From Retrieval to Reranking: Evaluating LLM Strategies for Long-Form Cases

Anonymous ACL submission

Abstract

Case retrieval is a critical component of case-based reasoning in domains such as law and medicine, where decisions are informed by prior cases. The task is particularly challenging because both queries and candidate cases are often extremely long with relevant evidence sparsely distributed across lengthy texts. We systematically study two aspects of long case retrieval. First, we compare full-document embeddings with LLM-generated summaries, finding that summaries improve retrieval performance for weaker methods such as BM25 while full-context representations are more effective for embeddings produced by strong LLMs such as Qwen3. Second, we examine reranking strategies, contrasting retrieval heads that capture token-level evidence with LLM-based rerankers that perform higher-level reasoning. Experiments show complementary strengths: retrieval heads excel with strong query-document overlap, while LLM-based rerankers perform better when complex reasoning is needed. Our findings provide guidance for designing retrieval systems that balance context coverage, token-level similarity, and reasoning for long-form cases.[1]

1 Introduction

Human experts in high-stakes domains often rely on prior cases and accumulated experience to guide their reasoning and inform complex decisions. Case-based reasoning (CBR) formalizes this process, where decisions are grounded in past experiences and established precedents (Hatalis et al., 2025; Wilkerson and Leake, 2024). This capability is especially critical in high-stake domains (Wiratunga et al., 2024; Sivarajkumar et al., 2024; Jeong et al., 2024), in which each reasoning step in decision making must be fully attributable and transparent.

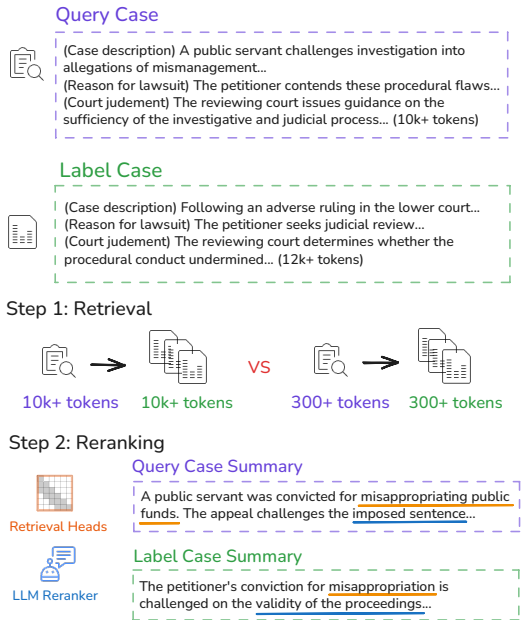

Figure 1: Our study on case retrieval task: Given a long-form query case and a corpus of long cases, retrievers first select candidate cases using either full-context representations or compressed LLM-generated summaries. Candidates are then reranked using retrieval heads or LLM rerankers.

A core component of CBR is case retrieval, which identifies relevant prior cases from large databases (De Mantaras et al., 2005; Yang, 2024). Recent advances increasingly leverage large language models (LLMs) for retrieval, owing to their strong long-context modeling capabilities and ability to follow complex instructions (Jiang et al., 2024; Weller et al., 2024; Zhang et al., 2025a). Despite this promise, case retrieval remains challenging: both queries and case documents are often extremely long, and LLMs are known to suffer from the lost-in-the-middle problem (Zhu et al., 2024; Liu et al., 2024). These challenges motivate our central research question: **How can different retrievers, including LLM-based retrievers, be effectively leveraged for case retrieval?**

[1] Code and results will be released after the review process.

Our study first examines how context representation affects retrieval performance. We compare two setting: *full-context* which preserve s all information but may dilute relevant signals, and *LLM-generated summaries* (Shukla et al., 2022), which concentrate important content but risk omitting details. Experiments on two legal and medical benchmarks reveal a clear trade-off between these representations. Summaries improve retrieval performance for weaker methods such as BM25, as they effectively filter out irrelevant information. In contrast, full-context representations are more effective for embeddings produced by strong LLMs such as Qwen3 (Yang et al., 2025), which are capable of handling long contexts.

After retrieval, a common practice is to rerank the candidate set using a stronger model. We study two complementary paradigms: retrieval heads (Zhang et al., 2025b; Wu et al., 2024), which score candidates via token-level relevance, and LLM-based rerankers (Rathee et al., 2025), which summarize and jointly reason over the entire candidate set. Retrieval heads perform best on tasks with strong lexical–semantic overlap, achieving up to 64.95 nDCG@10 on FIRE, whereas LLM-based rerankers are more effective when higher-level reasoning is required, improving nDCG@10 from 48.07 to 53.13 on CDS.

In summary, we present a systematic analysis of how context representations and reranking strategies influence case retrieval performance. Our findings provide practical guidance for designing effective retrieval pipelines that support reliable case retrieval in real-world applications.

2 Task Formulation

Given a corpus of long-form cases $\mathcal{D} = \{d_1, \ldots, d_N\}$ and a long-form query case q, the task of long case retrieval involves a two-stage process. First, a retrieval system selects a subset of potentially relevant cases from $\mathcal{D}$, and then a reranker orders these cases such that those most relevant to q appear at the top. Each query is associated with a small set of gold relevant cases $\mathcal{R}(q) \subset \mathcal{D}$. The ranked output $\pi(q)$ is evaluated using standard top-k metrics such as nDCG@k and Recall@k.

3 Method

This section describes the retrieval and reranking methods evaluated in our experiments, including LLM-based retrieval§3.1, retrieval heads as rerankers §3.2, and LLM reranking §3.3.

3.1 LLM-Based Retrieval

We model queries $q \in Q$ and documents $d \in D$ as token sequences from a vocabulary V. An encoder f maps each sequence to contextualized token embeddings, which are mean-pooled to obtain dense representations $\mathbf{e}_q$ and $\mathbf{e}_d$. Document relevance is measured by cosine similarity, and the top-k documents are retrieved as candidates.

To handle extremely long documents, we consider two retrieval settings. In **full-context retrieval**, each document is split into overlapping 10k-token segments, encoded independently, and aggregated into a document-level embedding. In **summary-based retrieval**, a generative model produces a concise (∼300-token) summary per document, which is encoded as a single vector. This design enables us to examine the trade-off between context coverage and information density.

3.2 Retrieval Heads as Rerankers

Retrieval heads rerank the top-50 candidates by aggregating attention from query to document tokens. To mitigate the effect of long documents and queries, we apply **length normalization** that averages only non-zero attention scores, and **query calibration** that retains query tokens with high variance in their attention distributions across document tokens:

$$\mathrm{Var}_h(t_q) = \frac{1}{|d_i|} \sum_{t_d \in d_i} \left(A^h_{t_q \to t_d} - \mu^h_{t_q} \right)^2, \quad (1)$$

$$\mu^h_{t_q} = \frac{1}{|d_i|} \sum_{t_d \in d_i} A^h_{t_q \to t_d}. \quad (2)$$

3.3 LLM Reranking

LLMs are used to rerank candidates either **pointwise**, scoring each query–document pair independently, or **listwise**, producing a reordered ranking of the entire candidate set. These models leverage reasoning over the query and document summaries, rather than relying solely on token-level similarity.

4 Experiments

This section first presents the dataset and experiment settings (§4.1 and §4.2). Then we provide detailed results (§4.3) and analysis (§4.4).

4.1 Dataset

We evaluate our methods on two case retrieval benchmarks. (1) **IFIR** (Song et al., 2025), which

Retriever	AILA		FIRE		COLIEE		CDS		PM	
	nDCG@10	Recall@10	nDCG@10	Recall@10	nDCG@10	Recall@10	nDCG@20	Recall@20	nDCG@20	Recall@20
BM25										
FULL CONTEXT	15.88	12.01	52.71	35.85	34.51	34.53	21.47	3.12	39.95	8.21
SUMMARY CONTEXT	16.12	12.26	55.11	42.59	38.49	40.51	21.30	3.68	42.63	8.96
NV-Embed-v2										
FULL CONTEXT	11.96	11.08	59.16	40.53	36.63	36.96	55.20	13.42	60.24	18.89
SUMMARY CONTEXT	16.11	16.10	54.94	39.22	36.44	36.94	54.00	13.13	61.03	24.90
Qwen3-Embedding-8B										
FULL CONTEXT	**19.52**	13.74	**61.75**	**43.69**	**41.37**	**46.42**	**56.62**	**16.10**	**68.61**	**32.96**
SUMMARY CONTEXT	18.91	**17.37**	59.82	42.25	40.10	41.03	53.83	12.65	62.34	27.41

Table 1: Retrieval performance across five benchmarks using different base retrievers under full context and summary context. Bold indicates the best performance among all models.

Retriever	AILA		FIRE		COLIEE		CDS		PM	
	nDCG@10	Recall@10	nDCG@10	Recall@10	nDCG@10	Recall@10	nDCG@20	Recall@20	nDCG@20	Recall@20
Retrieval Heads										
QWEN3-EMBEDDING-8B	**23.85**	**22.60**	63.84	49.02	42.18	45.73	48.07	12.87	58.40	27.44
QWEN3-4B	18.65	17.05	62.76	48.37	**43.53**	**48.97**	48.07	12.87	63.57	29.05
QWEN3-32B	21.87	21.66	**64.95**	**50.73**	42.74	47.00	46.88	13.47	**66.23**	**30.55**
LLM Pointwise Reranking										
QWEN3-4B-INSTRUCT-2507	19.01	22.35	53.94	40.40	19.63	13.30	56.12	15.58	53.38	23.22
QWEN3-32B	21.46	22.44	62.26	45.92	20.21	14.27	**61.92**	**15.84**	61.99	25.79
GPT-4.1	21.97	23.61	64.42	47.85	22.63	17.39	64.60	18.07	70.85	33.87
LLM Listwise Reranking										
QWEN3-4B-INSTRUCT-2507	19.79	12.94	50.42	36.26	20.62	16.98	50.30	12.68	64.00	25.05
QWEN3-32B	18.58	13.99	51.95	37.24	21.94	17.60	53.13	14.00	64.19	26.84
GPT-4.1	27.37	22.83	61.96	46.79	22.03	14.21	61.14	17.46	70.87	34.18

Table 2: Reranking performance across five benchmarks using different base models under summary context. Bold indicates the best performance among all non-GPT base models

comprises four long-form case retrieval datasets. AILA and FIRE are legal case retrieval datasets containing lawsuit cases with detailed case descriptions and corresponding final appeal outcomes. PM and CDS are medical case retrieval datasets that describe patients' conditions, disease progressions, and treatment records. (2) **COLIEE 2025** (Goebel et al., 2025), a legal case retrieval benchmark focusing on statutory and case-law reasoning. Across all datasets, both queries and documents can reach lengths of up to 160k tokens. For AILA, FIRE, and COLIEE, we report nDCG@10 and Recall@10 which smaller gold sets of approximately 3 relevant documents per query. For CDS and PM, we use nDCG@20 and Recall@20 because they have a larger gold set at 10.

4.2 Setups

Models. We use Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025) to generate summaries. We compare sparse and dense retrievers, including BM25 (Robertson and Walker, 1994), NV-Embed-v2 (Lee et al., 2024), and Qwen3-Embedding-8B (Zhang et al., 2025c). For retrieval heads, we evaluate Qwen3-Embedding-8B, Qwen3-4B-Instruct-2507, Qwen3-32B with 16 heads for 4B/8B and 64 heads for 32B models, covering 1—2% of total attention heads. For LLM-based reranking, we use Qwen3-4B-Instruct-2507, Qwen3-32B, and use ChatGPT-4.1 (OpenAI, 2025) as a reference for upper-bound performance.

4.3 Main Results

BM25 benefits most from summary context while Qwen3-Embedding-8B benefits most from full context. Table 1 shows that BM25 benefits most from summary context. For example, Recall@10 on FIRE is 42.59 over 35.85 with summaries. This suggests that compressing long documents helps the sparse retriever focus on key information. In contrast, Qwen3-Embedding-8B performs best when provided with the full context, achieving nDCG@20 of 68.61 on PM, compared to 62.34 with summaries. This advantage stems from the Qwen3 models' strong long-context capabilities, which enable it to identify salient information directly from extended inputs. As a result, retrieving from full-context documents avoids in-

formation loss introduced by summarization.

Retrieval heads and LLM reranking perform differently across datasets. Table 2 shows that retrieval heads and LLM-based rerankers exhibit distinct performance patterns across datasets. Retrieval heads perform strongly on FIRE and COLIEE, with Qwen3-32B achieving 64.95 nDCG@10 / 50.73 Recall@10 on FIRE. In contrast, LLM-based reranking yields clear gains on CDS, where Qwen3-32B listwise reranking improves retrieval heads on nDCG@10 from 48.07 to 53.13. We provide further analysis in the next section.

4.4 Analysis

Summaries help when they correctly identify key information but hurt when central points are missed. Query and document cases often contain multiple loosely related facts, which can obscure true relevance signals and hinder retrieval. When summaries successfully preserve the decisive reasoning from long contexts, they reduce noise and enable retrievers to focus on salient information. Conversely, if key points are omitted or distorted, summarization can mislead the retriever and degrade performance.

Summary effectiveness positively correlates with relevance density of documents. We calculate the relevance density between each query and its corresponding gold full-context document using Appendix Eq. 3. Appendix Table 4 shows that, across all datasets, queries whose retrieval performance improved after summarization tend to be associated with lower full-context relevance density than those that degraded. This indicates that long documents with sparse distributions of relevant information benefit more from summarization. The effect is particularly evident on PM and CDS, where large ΔDensity differences (+0.0659 and +0.0914) coincide with strong positive correlations (0.590 and 0.570) between relevance density change and retrieval improvement.

Retrieval heads are strongly aligned with semantic similarity. To analyze the behavior of different reranking paradigms, we measure rank correlations between BM25 and reranked outputs using Kendall's τ and Spearman's ρ. As shown in Appendix Table 5 and 6, retrieval heads consistently exhibit higher correlation with BM25 than LLM pointwise rerankers across all datasets. The positive $\Delta\tau$ and $\Delta\rho$ values, together with high win rates, indicate that retrieval heads largely preserve BM25-style semantic ranking patterns. In contrast, LLM rerankers frequently reorder candidates, reflecting their focus on reasoning-based relevance rather than lexical alignment.

Limitations of retrieval heads on instruction-following tasks. The strong alignment between retrieval heads and BM25 suggests that retrieval heads primarily function as token-level similarity matchers, rather than performing higher-level reasoning. While this behavior is effective when relevance is driven by lexical similarity, it becomes a limitation for tasks where relevance depends on understanding nuanced task intent. This effect is particularly pronounced on the CDS dataset, where rerankers must distinguish whether treatments or disease descriptions should be prioritized based on the query.

5 Related Work

Recent work on case retrieval highlights its importance and difficulty in domains like law and medicine (Hou et al., 2025; Tang et al., 2024). Early methods relying on simple text or embedding matching struggled to capture evidence scattered across long texts (Shao et al., 2020; Pradeep et al., 2023). Recent advances combine long-context LLMs with retrieval, using either hierarchical representations or LLM-generated summaries to handle lengthy inputs (Zhu et al., 2024; Xu et al., 2023). Retrieval heads, which identify attention patterns that highlight relevant tokens, have been shown to improve candidate selection in long documents (Wu et al., 2024; Zhang et al., 2025b). Complementarily, LLM-based rerankers perform pointwise or listwise reasoning over retrieved candidates, capturing higher-level relevance beyond token overlap (Mozafari and Jatowt, 2025; Luo et al., 2025).

6 Discussion

Our analysis highlights the complementary roles of context representation and reranking in long case retrieval. Building on these findings, future work may explore adaptive pipelines that dynamically select between full-context and summary representations, as well as hybrid reranking frameworks that combine evidence-based retrieval heads with reasoning-oriented LLM rerankers. Another promising direction is structure- or evidence-aware summarization that better preserves legally or clinically salient information in long cases.

Limitations

This work focuses on a limited set of LLM-based dense retrievers and rerankers, and therefore does not fully capture the diversity of possible model architectures or training paradigms for long case retrieval. In addition, existing benchmarks for long case retrieval are relatively small, reflecting the scarcity of publicly available legal and medical case datasets; this constraint may limit the statistical power of our analysis and the generalizability of the observed trends. Our study further relies on LLM-generated summaries as the primary mechanism for context compression, and alternative summarization or structured representation methods may lead to different trade-offs.

Ethics Statement

Our work focuses on improving retrieval of long-form legal and medical cases using large language model embeddings and reranking strategies. All datasets used are publicly available benchmarks (AILA, FIRE, CDS, PM, COLIEE 2025), and no private or sensitive patient or client data was collected or used. While our methods aim to enhance decision support in high-stakes domains, we emphasize that they are designed to assist human experts rather than replace professional judgment. Our experiments were conducted ethically, with proper citation of datasets and models, and we encourage transparency and caution in applying these techniques in real-world legal or medical settings.

References

Ramon Lopez De Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, MICHAEL T COX, Kenneth Forbus, et al. 2005. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, 20(3):215–240.

Randy Goebel, Yoshinobu Kano, Japan Calum Kawn, Mi-Young Kim, and Masaharu Yoshioka. 2025. International competition on legal information extraction and entailment (coliee 2025).

Kostas Hatalis, Despina Christou, and Vyshnavi Kondapalli. 2025. Review of case-based reasoning for llm agents: theoretical foundations, architectural components, and cognitive integration. *arXiv preprint arXiv:2504.06943*.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. Clerc: A dataset for us legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7898–7913.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Qin Luo, Erjia Chen, Zhao Shi, and Bang Wang. 2025. Anchor-based pairwise comparison via large language model for recommendation reranking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5001–5005.

Abdelrahman Abdallah Bhawna Piryani Jamshid Mozafari and Mohammed Ali Adam Jatowt. 2025. How good are llm-based rerankers? an empirical analysis of state-of-the-art reranking models.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. A new series of GPT models featuring major improvements on coding, instruction following, and long context—plus our first-ever nano model.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Guiding retrieval using llm-based listwise rankers. In *European Conference on Information Retrieval*, pages 230–246. Springer.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, volume 2020, pages 3501–3507.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical information retrieval: a literature review. *Journal of healthcare informatics research*, 8(2):313–352.

Tingyu Song, Guo Gan, Mingsheng Shang, and Yilun Zhao. 2025. Ifir: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. *arXiv preprint arXiv:2503.04644*.

Yanran Tang, Ruihong Qiu, Hongzhi Yin, Xue Li, and Zi Huang. 2024. Caselink: Inductive graph learning for legal case retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2199–2209.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024. Promptriever: Instruction-trained retrievers can be prompted like language models. *arXiv preprint arXiv:2409.11136*.

Kaitlynne Wilkerson and David Leake. 2024. On implementing case-based reasoning with large language models. In *International Conference on Case-Based Reasoning*, pages 404–417. Springer.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rui Yang. 2024. Casegpt: a case reasoning framework based on language models and retrieval-augmented generation. *arXiv preprint arXiv:2407.07913*.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025a. Diffusion vs. autoregressive language models: A text embedding perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025b. Query-focused retrieval heads improve long-context reasoning and re-ranking. *arXiv preprint arXiv:2506.09944*.

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

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Longembed: Extending embedding models for long context retrieval. *arXiv preprint arXiv:2404.12096*.

A Dataset Overview

Table 3 provides an overview of the datasets used in our experiments, summarizing their scale and length characteristics. We report the number of queries and documents, along with detailed statistics of query and corpus lengths, including mean and standard deviation as well as maximum token counts, for both full texts and their summaries. The table also presents the average number of relevant labels per query.

B Computational Resources and Model Parameters

All experiments were conducted using NVIDIA A100 GPUs. Specifically, Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025) was run on 2 A100 GPUs for summary generation of both queries and documents. Retrieval heads for Qwen3-32B and inference were executed on 2 A100 GPUs, while Qwen3-4B and Qwen3-8B models used a single A100 GPU. LLM-based reranking with ChatGPT-4.1 (OpenAI, 2025) was performed via API calls. For all inference tasks—including summary generation, retrieval head scoring, and LLM-based reranking, we set the temperature to 0.1 to ensure stable and consistent outputs.

C Relevance Density Analysis

To understand how context representation affects the distribution of relevant information, we define the *relevance density* of a document d with respect to a query q as

$$\text{Relevance Density}(q, d) = \frac{\cos(\mathbf{q}, \mathbf{d})}{\log(|d|)}, \quad (3)$$

where $\mathbf{q}$ and $\mathbf{d}$ are the embeddings of the query and document, and $|d|$ denotes the number of tokens in d. This formulation captures the average concentration of relevant information, normalized by document length, allowing comparison between full-context and summary representations.

We compute the full-document relevance density for each query across five benchmarks (AILA, FIRE, COLIEE, PM, and CDS) and classify queries as *Improved (Imp.)* or *Degraded (Deg.)* based on whether summaries enhanced retrieval. Table 4 reports the gold labels, percentages of improved/degraded queries, average full-document densities, ΔDensity (with standard deviation), their difference (Diff), and correlation (Corr) between ΔDensity and retrieval improvement. Positive differences and correlations indicate that summaries concentrate relevant content, thereby improving retrieval.

D Alignment of Retrieval-Head and Listwise LLM with BM25 Summaries

We evaluate how different reranking strategies align with BM25 in long case retrieval using summary context across five benchmarks: COLIEE, FIRE, PM, AILA, and CDS. For each query, we compute rank correlation metrics (Kendall's τ and Spearman's ρ) between BM25 and two rerankers: Retrieval-Head (RH) and LLM Pointwise (LP). Table 5 summarizes the overall alignment, showing the difference (Δ) between RH and LP, along with Win Rates (WR) indicating the fraction of queries where RH outperforms LP. Table 6 provides per-dataset metrics, standard deviations, and statistical significance (p-values). Positive Δ values and high WR demonstrate that RH consistently aligns more closely with BM25 than LP, with COLIEE and CDS showing the largest gains.

Dataset	Queries	Corpus	Query Tokens (mean±std)	Query Max	Query Summary (mean±std)	Query Summary Max	Labels /Query	Corpus Tokens (mean±std)	Corpus Max	Corpus Summary (mean±std)	Corpus Summary Max
AILA	40	2,914	654±258	1,363	274±101	485	3.0±2.2	4,554±4,207	50,702	397±104	883
FIRE	168	1,745	618±109	956	348±64	543	3.4±1.2	6,257±3,502	15,945	459±108	932
COLIEE	250	2,159	8,227±6,407	47,989	556±78	845	4.4±3.3	7,140±7,929	160,577	546±83	1,021
CDS	43	633,955	48±12	82	48±12	82	10.8±9.1	315±150	3,913	297±111	754
PM	59	241,006	29±5	44	29±5	44	20.6±12.9	436±365	3,474	398±297	702

Table 3: Statistics of datasets used in our experiments. We report the number of queries and documents, token lengths of queries and corpora (mean±standard deviation and maximum), summary lengths, and the average number of relevant labels per query.

Dataset	Gold	Imp. (%)	Deg. (%)	Full Density (Imp.)	Full Density (Deg.)	ΔDensity (Imp.)	ΔDensity (Deg.)	Diff	Corr
AILA	109	45 (41.3)	64 (58.7)	0.1031	0.1044	-0.0072 ± 0.0024	-0.0177 ± 0.0200	+0.0105	0.318
FIRE	664	254 (38.3)	410 (61.7)	0.0876	0.0923	$+0.0031 \pm 0.0108$	-0.0032 ± 0.0104	+0.0063	0.278
COLIEE	1054	468 (44.4)	586 (55.6)	0.1069	0.1077	-0.0043 ± 0.0130	-0.0110 ± 0.0115	+0.0067	0.266
PM	956	475 (49.7)	481 (50.3)	0.0628	0.1305	$+0.0601 \pm 0.0608$	-0.0059 ± 0.0196	+0.0659	0.590
CDS	720	309 (42.9)	411 (57.1)	0.1094	0.1400	$+0.0379 \pm 0.0618$	-0.0535 ± 0.0676	+0.0914	0.570

Table 4: Relevance density analysis comparing summary-based and full-context retrieval across five datasets. Columns report the number of gold labels, the percentage of queries that improved or degraded with summaries, full-document relevance density for improved and degraded queries, the corresponding density changes (ΔDensity), their difference, and the correlation between ΔDensity and retrieval improvement. Positive differences and correlations indicate that summarization concentrates relevant content and improves retrieval performance.

Dataset	**N**	τ(BM25→RH)	τ(BM25→LP)	$\boldsymbol{\Delta\tau}$	ρ(BM25→RH)	ρ(BM25→LP)	$\boldsymbol{\Delta\rho}$	WR(τ)	WR(ρ)
COLIEE	250	0.434	0.006	**+0.427**	0.588	0.009	**+0.579**	98.8%	98.8%
FIRE	168	0.340	0.232	**+0.107**	0.475	0.333	**+0.142**	81.5%	81.0%
PM	59	0.131	0.021	**+0.110**	0.189	0.031	**+0.158**	72.9%	71.2%
AILA	40	0.256	0.131	**+0.125**	0.365	0.190	**+0.175**	80.0%	75.0%
CDS	43	0.268	0.019	**+0.249**	0.383	0.026	**+0.357**	95.1%	95.1%

Table 5: Summary rank correlation analysis comparing Retrieval-Head (RH) and Listwise LLM (LP) against BM25 using summary context only. $\Delta\tau = \tau(\text{BM25}\rightarrow\text{RH}) - \tau(\text{BM25}\rightarrow\text{LP})$ and $\Delta\rho = \rho(\text{BM25}\rightarrow\text{RH}) - \rho(\text{BM25}\rightarrow\text{LP})$ measure how much more aligned RH is with BM25 compared to LP. Positive Δ values indicate RH is more aligned with BM25. WR = Win Rate (% of queries where RH > LP).

Dataset	**Metric**	**RH**	**LP**	**Difference**	**Std Dev**	**Win Rate**	**p-value**
COLIEE (n=250)	Kendall τ	0.4336	0.0061	**+0.4274**	0.1733	98.8%	0.000000
	Spearman ρ	0.5878	0.0085	**+0.5793**	0.2334	98.8%	0.000000
FIRE (n=168)	Kendall τ	0.3398	0.2324	**+0.1074**	0.1224	81.5%	0.000000
	Spearman ρ	0.4754	0.3329	**+0.1425**	0.1628	81.0%	0.000000
PM (n=59)	Kendall τ	0.1314	0.0209	**+0.1104**	0.1544	72.9%	0.000001
	Spearman ρ	0.1890	0.0311	**+0.1579**	0.2249	71.2%	0.000002
AILA (n=40)	Kendall τ	0.2561	0.1313	**+0.1249**	0.1288	80.0%	0.000000
	Spearman ρ	0.3651	0.1901	**+0.1751**	0.1828	75.0%	0.000001
CDS (n=43)	Kendall τ	0.2679	0.0190	**+0.2489**	0.1551	95.1%	0.000000
	Spearman ρ	0.3832	0.0259	**+0.3573**	0.2180	95.1%	0.000000

Table 6: Detailed per-dataset rank correlation metrics and statistical tests. RH = Retrieval-Head, LP = LLM Pointwise Reranker. Δ = Difference between RH and LP. p-value computed via t-test.