Representation Matters: A Budget-Controlled Evaluation of Scientific Document Embeddings

Anonymous ACL submission

Abstract

We introduce ArXivDocQA, an open-domain retrieval benchmark constructed directly from the raw LaTeX sources of scientific papers. By operating on LaTeX, ArXivDocQA retains fine-grained structural information—including figures, tables, equations, and section boundaries—while enabling controlled construction of decontextualized queries grounded in specific parts of a document. This design enables evaluation under realistic scientific search settings with explicit control over the evidence source. We systematically compare text-only, image-based, and multimodal retrieval representations under varying storage budgets, and show that document-as-image representations—on which many state-of-the-art document retrieval models are trained—are not universally optimal, and their performance depends strongly on where the relevant evidence resides in the document (text, tables, or figures).

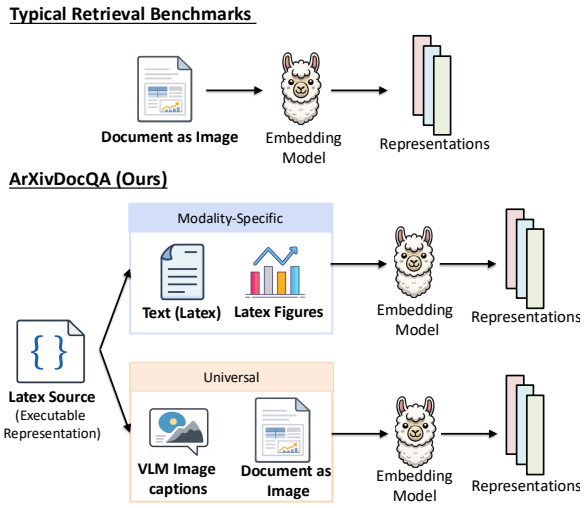


Figure 1: Typical retrieval benchmarks predominantly treat PDF pages as images. ArXivDocQA is purpose-built for scientific document retrieval and operates directly on LaTeX sources, enabling the exploration of both modality-specific and unified document representations.

1 Introduction

Scientific document retrieval requires locating evidence that may appear in prose, equations, tables, or figures. Yet most existing retrieval systems represent documents either as plain text or as page-level images, without accounting for whether the relevant evidence is textual, tabular, or visual. This raises a fundamental question: *which document representations are most appropriate for scientific retrieval, and under what conditions?*

Recent work has increasingly favored document-as-image representations, as exemplified by models such as ColPali (Faysse et al., 2024), VisRAG (Yu et al., 2024), and the Jina document retrieval models (Günther et al., 2025), motivated by the success of vision–language models trained on rendered pages. As a consequence, many existing benchmarks represent scientific documents primarily as page-level images (Ma et al., 2024; Macé et al., 2025a; Wasserman et al., 2025; Chia et al., 2025). While this choice preserves visual layout, it obscures information that is explicit in the document source, including figure and table boundaries, equation environments, section hierarchy, and machine-readable links between textual references and visual or tabular evidence—information that is present in LaTeX sources but flattened or lost in rendered page images. Scientific papers differ from natural documents in that they are authored in LaTeX, a structured format that explicitly encodes document organization and content types. When papers are represented only as images, this structure is no longer directly accessible: boundaries between content types must be inferred visually, cross-references are no longer explicit, and distinctions between prose, equations, and figures are not machine-readable. As a result, it remains unclear whether page-level image representations are well suited for all retrieval scenarios, particularly when

the relevant evidence is expressed primarily in text, mathematical discussion, or structured tables rather than in visually salient figures.

Moreover, recent evidence shows that vision–language model performance on scientific documents can vary substantially with document layout and rendering templates (Cheng et al., 2025). This sensitivity suggests that representations derived directly from the document source—such as LaTeX markup that explicitly encodes text content, figures, tables, equations, and structural boundaries—may offer more stable and controllable inputs for retrieval than rendered page images alone.

More broadly, there is an inherent trade-off between textual and visual document representations. Text-based representations align well with language model pretraining and preserve explicit semantic content such as terminology, equations, and logical structure, while visual representations capture layout and multimodal cues but require models to infer content boundaries and relationships implicitly from appearance (Faysse et al., 2024; Wei et al., 2025; Lyu et al., 2025). Much of the recent shift toward visual representations has been driven by document collections where accurate text extraction or structural markup is unavailable (Macé et al., 2025a; Ma et al., 2024). In contrast, scientific papers on arXiv provide clean LaTeX sources that preserve both content and structure, offering a unique testbed for systematically comparing textual and visual document representations under controlled conditions.

A second challenge concerns storage. Different representations lead to substantially different index sizes. Text-based retrieval systems trade off storage via chunk size and overlap, while vision-based systems depend on image resolution and the number of visual tokens produced by the encoder. Architectural choices further amplify these effects: late-interaction models (Khattab and Zaharia, 2020; Faysse et al., 2024) store multiple vectors per document unit, often increasing index size by orders of magnitude compared to single-vector embeddings. Despite this, most benchmarks report performance at a single operating point, obscuring the trade-offs between retrieval accuracy and storage cost that are critical for real-world deployment.

In this work, we introduce **ArXivDocQA**, a benchmark designed to systematically analyze scientific document retrieval across document representations and storage budgets. Using raw LaTeX sources, we generate targeted, evidence-grounded queries and construct distinct document representations for retrieval, including text-only, figure-only, caption-based, page-level image, and interleaved text–image representations. Queries in ArXivDocQA are explicitly grounded in text, tables, or figures, enabling retrieval behavior to be analyzed as a function of evidence type.

Furthermore, we treat index size as a controllable variable rather than a fixed constant. Index size directly affects inference latency and memory footprint, making it a primary concern in practical retrieval systems. By varying text chunking strategies and visual token budgets, we measure retrieval performance as a function of storage cost and characterize how accuracy changes as representations are compressed or expanded. Our results show that increasing index size does not necessarily improve retrieval performance. Instead, performance is primarily determined by whether the document representation preserves the type of evidence required by the query.

Contributions. We introduce **ArXivDocQA**, a scientific document retrieval benchmark built from raw LaTeX sources that enables controlled comparison of textual, visual, and hybrid document representations under realistic storage constraints. Using this benchmark, we show that document-as-image representations—despite their prominence in prior work—are not universally optimal, and that retrieval performance depends strongly on the type of evidence targeted by the query (text, tables, or figures). We further uncover that indexing documents using captions generated by vision–language models achieves the strongest overall retrieval performance, despite requiring substantially smaller indices than many image-based representations.

2 Related Work

In this section, we situate ArXivDocQA within the landscape of recent multimodal document retrieval and understanding benchmarks, highlighting how it addresses key limitations in existing datasets. Table 1 provides a high-level summary of the key differences.

ArXivDocQA is formulated as an *open-domain scientific document retrieval* task, in which models must identify the single document containing the relevant evidence from a large corpus. This setting differs from benchmarks such as ArXivQA (Li et al., 2024), which focus on document-specific questions over isolated figures. In ArXivDocQA,

	#Pages	#Documents	#Queries	LaTeX Exists	Open-Domain Queries	Scientific Documents
M-LongDoc	–	180	10,070	✗	✗	✓
MMLongBench-Doc	6,400	135	1,091	✗	✗	✓
ViDoRe V2	3,266	66	3,000	✗	✓	✗
REAL-MM-RAG	8,000	163	–	✗	✓	✗
Ours	156,524	9,234	679	✓	✓	✓

Table 1: Comparison of document-level benchmarks with query counts.

queries are context-independent and grounded in text, tables, or figures, requiring retrieval rather than within-document question answering.

Several benchmarks emphasize reasoning over long contexts rather than retrieval. MMLongBench-Doc (Ma et al., 2024) and M-LongDoc (Chia et al., 2025) evaluate multi-page understanding within fixed documents, where the relevant context is provided explicitly. ArXivDocQA instead frames scientific understanding as a retrieval problem over a large document collection, requiring models to distinguish between many closely related papers.

Other benchmarks focus more directly on retrieval over structured or multimodal documents. `ViDoRe V1` (Faysse et al., 2024) evaluates retrieval over curated candidate sets derived from VQA datasets, where the task is to identify relevant pages rather than search an unrestricted corpus. `ViDoRe V2` (Macé et al., 2025b) extends this setting to open-domain retrieval but concentrates on visually rich documents from specialized domains such as ESG and medical reports. In contrast, ArXivDocQA targets large-scale *scientific* document retrieval and uniquely exposes raw LaTeX sources, enabling controlled analysis of retrieval grounded in text, tables, and figures. As shown in Table 1, ArXivDocQA is substantially larger in both document and page count and is the only benchmark to provide both open-domain queries and source-level LaTeX access.

Recent work has also examined how representation choices affect retrieval performance. MRMR (Zhang et al., 2025a) studies reasoning-intensive retrieval over interleaved documents and finds that caption-based text retrieval can outperform native multimodal embeddings. REAL-MM-RAG (Wasserman et al., 2025) focuses on table-dense financial and technical reports and reports stronger performance from page-as-image representations, while ViDoSeek (Wang et al., 2025) observes similar trends on presentation slides. Consistently, ViDoRe V1 also reports advantages for page-level image retrieval.

ArXivDocQA enables a systematic comparison of retrieval strategies under these conditions. In our experiments, we study multiple embedding strategies and find that leveraging LaTeX source representations yields stronger retrieval performance than representing scientific documents solely as rendered page images.

3 Dataset

We introduce *ArXivDocQA*, an open-source scientific document retrieval benchmark constructed from raw LaTeX sources. The LaTeX source is used directly to define document representations and to generate queries, without relying on OCR-extracted PDF text. This allows document content and structure to be preserved consistently across all representations derived from the same underlying source.

The benchmark contains approximately 9,000 full-length scientific papers, spanning substantially more pages than prior multimodal retrieval benchmarks (Table 1). This scale supports evaluation over large document collections in which many papers share similar topics, notation, and experimental structure.

Queries are generated and filtered through a multi-stage process involving decontextualization and verification, producing a balanced benchmark for systematic evaluation of unimodal and multimodal retrieval methods.

3.1 Problem Formulation

We study open-domain scientific document retrieval. Let $D = \{d_1, \ldots, d_N\}$ denote a corpus of scientific documents. Each document d_i is represented as a set of embedding units

$$E_i = \{e_{i1}, \ldots, e_{iM_i}\},$$

where each embedding unit corresponds to a document component such as a text chunk, figure, or page, depending on the chosen representation and model.

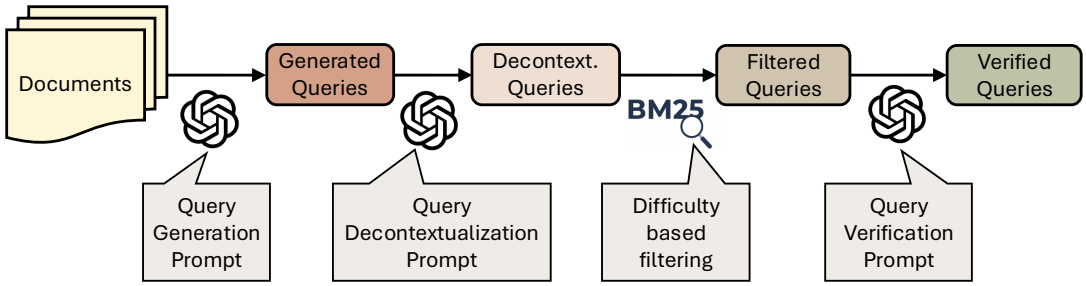


Figure 2: Figure illustrating the dataset construction pipeline.

Let q denote a context-independent natural language query targeting a specific piece of scientific evidence (text, table, or figure).

Given a query q, a retrieval system computes a similarity score between q and each embedding unit $e_{ij} \in E_i$ using a scoring function $s(q, e_{ij})$. The document-level score for d_i is then defined as

$$S(q, d_i) = \max_{e_{ij} \in E_i} s(q, e_{ij}),$$

and documents are ranked according to $S(q, d_i)$.

3.2 Query Generation and Filtering

We construct open-domain retrieval queries from three distinct evidence types present in scientific documents: *text*, *figures*, and *tables*. Each evidence type defines a separate pool of candidate queries, and the same multi-step generation and filtering pipeline is applied independently to each pool.

Evidence Modalities.

- **Text-based queries** are generated from raw LaTeX sources by randomly sampling text chunks from the `.tex` file.
- **Figure-based queries** are generated from figures extracted from the LaTeX source, using the rendered figure images as input.
- **Table-based queries** are generated from tables identified in the LaTeX source by detecting `\begin{table}` environments.

For each query type, we apply the same four-step pipeline: synthetic query generation, decontextualization, difficulty-based filtering, and final verification (see Figure 2). Prompt templates used in each stage are provided in Appendix B.[1]

[1] In preliminary experiments, directly prompting models to produce context-independent queries led to low-quality outputs.

We note that queries in ArXivDocQA are intentionally grounded in localized evidence (a specific text span, table, or figure) rather than requiring multi-hop reasoning across multiple document components. This design choice reflects our focus on document retrieval rather than document-level reasoning: the primary challenge is identifying the correct document from a large corpus, not aggregating evidence within a document.

3.2.1 Generation

For each evidence source (text, figure, or table), we prompt `gpt-5-mini` to generate a single query targeting the underlying scientific content. The prompt enforces that the query (i) requires expert-level reasoning (e.g., about implications, trends, limitations, or constraints), (ii) avoids direct restatement and minimizes lexical overlap through abstraction and paraphrasing, (iii) is answerable from the document without relying on keyword or phrase matching or referencing document-specific elements (e.g., sections, figures, or experiment names), and (iv) consists of exactly one realistic, concise sentence. Full prompt templates are provided in Appendix B.

3.2.2 Decontextualization.

To make these queries compatible with open-domain retrieval, we rewrite each synthetic query into a context-independent form that removes explicit references to figures, tables, or document-local structure. In practice, this rewriting step frequently produces queries that are superficially decontextualized but remain underspecified, as they fail to introduce enough scientific context to stand alone.

For example:

Original: *If the variable on the x-axis represents time, what can be inferred about the rate of change of the parameter over time?*
Rewritten: *If the independent variable represents*

time, what can be inferred about the rate of change of the parameter over time?

Although `gpt-5-mini` is supposed to return `null` when it cannot generate a valid rewrite, it frequently outputs very slight paraphrases like the one above, which are still unclear without the original context. This makes a final verification step necessary.

3.2.3 Difficulty-Based Filtering.

We remove trivially easy queries using an automated difficulty filter that leverages retrieval behavior. For each query, we run BM25 over a chunked document corpus and remove the query if its gold document ranks within the top five results, as this suggests the answer can be found through shallow lexical overlap. This step eliminates roughly 40% of all queries.

3.2.4 Final Verification.

All remaining queries undergo a final verification step using `gpt-5-mini` to ensure that they constitute valid open-domain retrieval queries. This verification checks that queries are interpretable without document context. After verification, the query counts are reduced from 620 to 443 (text), 930 to 178 (figure), and 673 to 58 (table).

3.2.5 Human Evaluation

Annotators are presented with 40 random queries *without* access to the source document and asked to evaluate each query along two dimensions: **Context Independence**, assessing whether the query can be understood and answered without document-local references; and **Well-Formedness**, assessing whether the query is a coherent, meaningful, and realistic scientific question. Two PhD students independently verified that all evaluated queries satisfied both criteria across text-, table-, and figure-based subsets. A list of queries and their corresponding ground-truth documents is provided in Appendix C.

4 Experiments

4.1 Experimental Setup

We use the proposed **ArXivDocQA** dataset to assess how different document representations support open-domain retrieval from scientific papers. To obtain a detailed view of the design space, we conduct a systematic evaluation of both modality-specific representations (text-only and figure-only) and universal representations that encode multiple modalities, under varying storage budgets.

Rather than reporting performance at a single operating point, we treat *index size* as the primary independent variable and analyze the resulting accuracy–storage tradeoff. This allows us to compare representations on equal footing and to characterize how retrieval performance evolves as the available storage budget increases.

Retrieval performance is measured using normalized discounted cumulative gain at rank 5 (nDCG@5), which evaluates whether the relevant document is ranked near the top of the retrieval list.

Representations and Models. We consider five document representations: (i) **Text (LaTeX)**, which indexes raw LaTeX source text. For this representation, we first *flatten* the source to a single file to account for projects split across multiple `.tex` inputs (e.g., via `\input` or `\include`). We then apply lightweight normalization to reduce non-semantic markup: we remove comments and strip common formatting commands (e.g., `\cite`, `\ref`, `\label`, `\footnote`, and styling macros such as `\emph`, `\textbf`), while preserving scientific content such as plain text, math, and section structure. (ii) **LaTeX Figures**, which indexes rendered figures extracted from the LaTeX sources. Document text is ignored in this setting. We collect all figure assets referenced in the source (including `.png`, `.jpg`, `.pdf`, `.eps`, and related formats) and convert them into a unified format prior to embedding. (iii) **VLM Captions**, which indexes textual descriptions of figures generated by a vision–language model alongside the document text, by appending the captions to the end of the document. (iv) **Document-as-Image**, which indexes full document pages rendered as images; and (v) **Interleaved Text + Images**, which jointly indexes text and figures while preserving their original order in the document. We parse the LaTeX source to extract textual spans and figure references, resolve each reference to its rendered image, and construct an interleaved sequence that reflects the document's narrative flow. Retrieval units are formed by segmenting the text into chunks and associating each chunk with the surrounding figures. This process produces multimodal units that pair text with zero or one images.

Across these representations, we evaluate several embedding models, depending on modality compatibility: **Qwen3-Embedding-8B** (Zhang et al., 2025b), **ColQwen2 v1** (Faysse et al., 2024), **Open-**

Representation	Model	# units	Index Size (GB)	Text	Table	Figure	Avg.
Text (LaTeX)	Qwen	334,991	2.56	**0.71**	0.46	0.50	0.64
	ColQwen	40,067	40.54	0.59	0.41	0.55	0.56
	OpenCLIP	334,991	1.28	0.05	0.10	0.12	0.07
	Ops-MM-Embedding	334,991	2.24	0.52	0.38	0.46	0.49
LaTeX Figures	ColQwen	138,847	22.09	0.06	0.12	**0.73**	0.22
	OpenCLIP	138,847	0.49	0.01	0.03	0.29	0.08
	Ops-MM-Embedding	138,847	0.86	0.15	0.22	0.65	0.27
VLM Captions	Qwen	202,936	1.54	0.70	**0.47**	0.59	**0.65**
	ColQwen	202,932	52.10	0.51	0.40	0.71	0.55
	OpenCLIP	202,932	1.65	0.07	0.10	0.15	0.09
	Ops-MM-Embedding	202,932	2.88	0.52	0.34	0.48	0.49
Document-as-Image	ColQwen	156,512	70.56	0.47	0.34	0.65	0.50
	OpenCLIP	156,512	0.60	0.01	0.03	0.06	0.02
	Ops-MM-Embedding	156,512	1.04	0.46	0.39	0.50	0.46
Interleaved Text + Images	ColQwen	1,202,897	49.59	0.37	0.32	0.60	0.42

Table 2: Retrieval performance (NDCG@5) using the best configuration for each *representation–model* pair. Weighted Avg. is computed as a query-count weighted mean over text, table, and figure queries ($n_{\text{text}} = 443$, $n_{\text{table}} = 58$, $n_{\text{figure}} = 147$). Hyperparameter sweeps are reported separately.

CLIP ViT-G/14 (Ilharco et al., 2021), and **Ops-MM-Embedding v1 (7B)** (Lin et al., 2025). For each representation–model pair, we report results using the best-performing configuration; full hyperparameter sweeps are reported separately.

ColQwen is a late-interaction model that encodes queries and documents into sets of token-level embeddings and computes document relevance via max-similarity aggregation across embedding units, resulting in substantially larger indices than single-vector models. In contrast, Qwen, OpenCLIP, and Ops-MM-Embedding produce a single embedding per input unit and rely on standard vector similarity for retrieval.

ColQwen was originally introduced for page-level visual document retrieval and is not trained on text documents. Nevertheless, we apply ColQwen to text chunks by treating each chunk as an embedding unit and using the model's language encoder. This allows us to evaluate a late-interaction retrieval model on purely textual representations and to compare its behavior directly with single-vector text embedding models under identical document inputs.

Index Size and Number of Units. Index size is defined as the total storage required for all embedding vectors, excluding model parameters. The *number of units* corresponds to the number of indexed elements (e.g., text chunks, figures, pages, or interleaved segments), which varies by representation and directly determines index size.

For **text-based representations**, index size is controlled by varying the *chunk size* used to segment documents prior to embedding. Smaller chunks increase the number of units and the total index size, while larger chunks reduce storage at the cost of coarser representations. This mechanism is used for all text indexings. Note that varying chunk size does not substantially change the index size for ColQwen, since text is encoded at the token level and stored as a set of embeddings regardless of chunk boundaries.

For vision-based representations, index size is controlled via the `max_pixels` parameter, which caps the total number of input pixels processed per image. Images are resized to approximately preserve aspect ratio while satisfying this budget. The number of visual tokens scales with the effective image resolution and can be approximated as

$$T_{\text{vis}} \propto \frac{H'W'}{P^2}, \quad \text{with } H'W' \leq \texttt{max\_pixels},$$

where $H' \times W' \leq \texttt{max\_pixels}$ is the resized image resolution and P is the vision encoder's patch size. Reducing `max_pixels` therefore decreases the number of visual tokens and the resulting index size, trading visual detail for storage efficiency.

To ensure fair comparison across representations, we evaluate each method across a range of storage budgets rather than at a single operating point. Results are reported by selecting the optimal hyperparameters.

4.2 Main Results

Table 2 reports retrieval performance measured by NDCG@5, using the best configuration for each representation–model pair. Results are reported separately for text, table, and figure queries.

Retrieval accuracy is highest when the document representation matches the evidence modality targeted by the query. Table 2 shows a clear correspondence between query modality and document representation. Textual representations perform best on text and table queries, while figure-based representations dominate on figure queries. In contrast, representations that mismatch the required evidence modality (e.g., vision-only representations for text or table queries) consistently underperform. Overall, retrieval accuracy is largely determined by whether the indexed representation preserves the modality of evidence required by the query.

While modality-specific representations achieve the highest accuracy when the query evidence type is known, they do not generalize across all query types. We therefore analyze representations designed to support multiple modalities within a single index. Across models, **document-as-image representations are not the strongest universal choice, despite the prevalence of page-level image training in recent retrieval systems (Faysse et al., 2024; Xu et al., 2025).** When comparing representations under the same embedding model, document-as-image consistently underperforms alternatives that retain explicit textual structure.

Within universal representations, **VLM caption-based indexing yields the best overall tradeoff between cross-modal coverage and retrieval accuracy.** Across all model–representation pairs, the highest weighted average performance is achieved by Qwen with Text (LaTeX) indexing. Notably, although ColQwen is trained as a vision–language model using rendered document images and captions, its architecture includes a shared language backbone that is also trained on text-only data. **Despite not being explicitly optimized for text-only document retrieval, ColQwen achieves its strongest average performance when indexing textual representations and VLM-generated captions, rather than page-level images.** This result indicates that strong multimodal encoders can effectively exploit structured textual representations at indexing time, even when their training emphasizes visual document inputs.

4.3 Storage and Retrieval Performance

Figure 3 shows how retrieval performance varies with storage budget across evidence types. Although increasing storage often improves accuracy, the effect is not uniform and depends on the type of evidence being retrieved.

In some settings, performance improves rapidly with modest increases in storage and then saturates, while in others gains are more gradual or inconsistent. Importantly, larger indices do not reliably yield better retrieval performance.

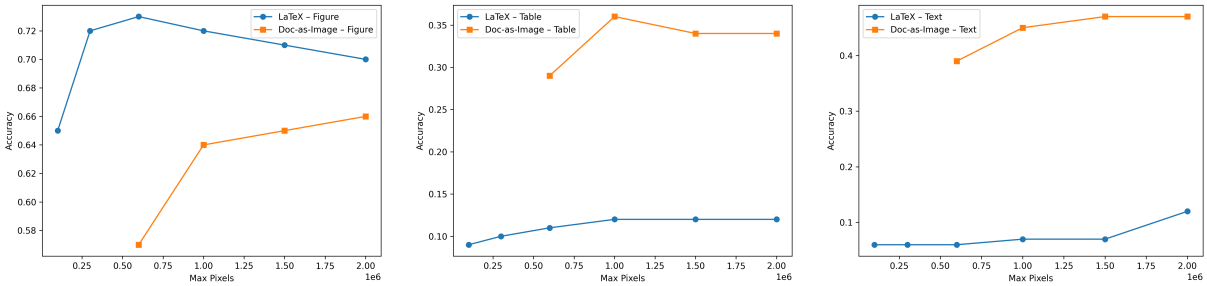
For example, Table 2 shows that ColQwen-based document-as-image representations require tens of gigabytes of storage to approach the performance of figure-centric representations, while ColQwen with LaTeX figures achieves strong accuracy at significantly smaller storage budgets. Similarly, text-based indexing with Qwen attains competitive performance at sub-gigabyte scales, despite using orders of magnitude fewer stored tokens than many vision-based configurations.

Overall, these results indicate that raw storage footprint is a weak predictor of retrieval quality. Instead, performance is driven primarily by representation choice and how effectively it preserves the target evidence. Nevertheless, storage-related parameters—such as the number of stored tokens—remain important hyperparameters. In practice, these are governed by design choices including text chunk size and visual token budgets (e.g., `max_pixels`).

4.4 Analysis

4.4.1 OCR Text vs. LaTeX Sources

Table 3 compares retrieval using text extracted from rendered PDFs with retrieval over raw LaTeX sources. Using LaTeX text generally leads to stronger retrieval performance at similar storage cost, with the effect more pronounced for single-vector text embedding models. Overall, these results suggest that cleaner source text provides more effective representations for scientific document retrieval than PDF-extracted text.



(a) Figure-based retrieval (b) Table-based retrieval (c) Text-based retrieval

Figure 3: Retrieval accuracy as a function of storage budget, controlled via the visual token budget (`max_pixels`), for figure-, table-, and text-based queries.

Model		Units	GB	Text	Table	Fig.
Qwen (512)	PDF	370k	2.83	0.66	0.44	0.41
	LaTeX	335k	2.56	**0.71**	**0.46**	**0.50**
Qwen (1024)	PDF	175k	1.33	0.64	0.44	0.39
	LaTeX	158k	1.20	**0.70**	**0.46**	**0.42**
ColQwen (4096)	PDF	44k	39.95	0.58	**0.42**	0.54
	LaTeX	40k	40.54	**0.59**	0.41	**0.55**

Table 3: Comparison of text extracted from rendered PDFs (via `pypdf`) and raw LaTeX sources.

4.4.2 Qualitative Failure Analysis

Document-as-image representations systematically fail when the relevant evidence is semantic rather than visual. We analyze cases where document-as-image retrieval fails while text-based retrieval succeeds, isolating failure modes attributable to representation choice rather than model capacity or retrieval scale. Across many such queries, the key limitation is that the evidence required for retrieval is distributed across textual and mathematical discussion—such as assumptions, limitations, or explanatory arguments—rather than concentrated in a visually distinctive figure or page. Consequently, page-level image representations tend to retrieve visually similar but semantically unrelated documents, such as pages dense with equations or plots. For example, for the query (evidence in text) "How does increasing the horizontal repeat length relative to the operating wavelength affect the difficulty of approximating contributions from distant repeats and the computational work required per layer?", the relevant evidence is conveyed through extended mathematical formulation and accompanying textual explanation rather than a single visually distinctive element. As shown in the appendix D, this relationship must be inferred from the structure of the equations and their surrounding discussion, making it accessible to text-based retrieval but difficult to identify from page-level visual similarity alone.

5 Conclusion and Future Work

Conclusion. We introduce **ArXivDocQA**, a benchmark for open-domain scientific document retrieval built from raw LaTeX sources that enables systematic evaluation of document representations under realistic storage constraints. ArXivDocQA provides a foundation for studying how representation choices and storage budgets shape retrieval behavior in scientific search.

Future Work. Several directions remain open. A natural extension is to train retrieval models directly on interleaved scientific documents that combine text, figures, tables, and equations, rather than treating modalities independently. In addition, late-interaction retrieval models have not been explicitly trained for document-level scientific text retrieval; adapting and training such models on full scientific corpora may yield further gains. Finally, ArXivDocQA can be extended to support more complex retrieval settings, such as queries requiring aggregation across multiple pieces of evidence or document components.

6 Limitations

ArXivDocQA evaluates document-level retrieval at scale, which introduces the possibility of false negatives: multiple documents in the corpus may plausibly address a query, even though only a single document is labeled as relevant.

In addition, queries in ArXivDocQA are generated automatically and can be overspecified. In particular, some GPT-generated queries are longer or more detailed than typical user-issued search queries.

7 Ethics Statement

We do not believe there are significant ethical issues associated with this research.

References

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, and 1 others. 2025. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2025. M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9244–9261, Suzhou, China. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550, Suzhuo, China. Association for Computational Linguistics.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. *Preprint*, arXiv:2411.02571.

Zhiheng Lyu, Xueguang Ma, and Wenhu Chen. 2025. Pixelworld: How far are we from perceiving everything as pixels? *Preprint*, arXiv:2501.19339.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: benchmarking long-context document understanding with visualizations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Quentin Macé, António Loison, and Manuel Faysse. 2025a. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*.

Quentin Macé, António Loison, and Manuel Faysse. 2025b. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Mengyao Xu, Gabriel Moreira, Ronay Ak, Radek Osmulski, Yauhen Babakhin, Zhiding Yu, Benedikt Schifferer, and Even Oldridge. 2025. Llama nemoretriever colembed: Top-performing text-image retrieval model. *Preprint*, arXiv:2507.05513.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025a. Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. *Preprint*, arXiv:2510.09510.

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*.

A Hyperparameters

Table 4 reports retrieval accuracy (nDCG@5) under varying storage budgets, grouped by document representation. Storage is controlled through representation-specific hyperparameters, including text chunk size, visual token budgets (`max_pixels`), and their combinations for interleaved representations.

Representation	Model	# units	Index Size (GB)	Text	Table	Figure
Text	*Qwen*					
	chunk=512	334,991	2.56	0.71	0.46	0.50
	chunk=1024	157,692	1.20	0.70	0.46	0.42
	chunk=4096	40,067	0.34	0.63	0.42	0.30
	ColQwen					
	chunk=1024	157,692	39.85	0.54	0.43	0.56
	chunk=4096	40,067	40.54	0.59	0.41	0.55
LaTeX Figures	*ColQwen*					
	max_pixels $= 100K$	138,847	6.75	0.06	0.09	0.65
	max_pixels $= 300K$	138,847	10.74	0.06	0.10	0.72
	max_pixels $= 600K$	138,847	22.09	0.06	0.11	0.73
	max_pixels $= 1M$	138,847	35.43	0.07	0.12	0.72
	max_pixels $= 2M$	138,847	65.99	0.12	0.12	0.70
VLM captions	*Qwen*					
	VLM captions + text	202,936	1.54	0.66	0.43	0.53
	ColQwen					
	VLM captions + text	202,932	52.10	0.51	0.40	0.71
Document-as-Image	*ColQwen*					
	max_pixels $= 600K$	156,512	28.37	0.39	0.29	0.57
	max_pixels $= 1M$	156,512	47.23	0.45	0.36	0.64
	max_pixels $= 1.5M$	156,512	70.56	0.47	0.34	0.65
	max_pixels $= 2M$	156,512	94.79	0.47	0.34	0.66
Interleaved Text + Images	*ColQwen*					
	max_pixels $= 1M$ & chunk=1024	1,202,897	49.59	0.37	0.32	0.60

Table 4: Retrieval accuracy (NDCG@5) under varying storage budgets, grouped by document representation.

B Query Generation

The query generation pipeline consists of four stages, three of which involve prompting `gpt-5-mini`. In this section, we report the system prompts used at each stage.

B.1 Synthetic Query Generation Prompt (Text)

PROMPT

You are given extracted text from a scientific research paper. Your task is to generate a single, high-quality synthetic query that would meaningfully test a document retrieval system.

Instructions:

1. The query must require expert-level reasoning over implications, trends, limitations, or constraints discussed in the document, and must not be a direct restatement of any sentence from the input.
2. The query must minimize lexical overlap with the input text by avoiding distinctive phrases or terminology, relying instead on abstraction and paraphrasing rather than keyword matching.
3. The query must be answerable from the document but not trivially retrievable via keyword or phrase matching, and must not reference sections, figures, experiments, or document-specific wording.
4. The query must ask exactly one focused question, without combining multiple sub-questions or enumerating parameters or conditions.
5. The query must be realistic and concise, phrased as a single sentence that a knowledgeable researcher would plausibly ask, without verbose framing or artificial difficulty.
6. If no query satisfying these criteria can be generated, return null.

Required Output Format:

```
{
  "query": "<generated question or null>"
}
```

Here is the document content:
{paper_text}

Figure 4: Prompt used for generating synthetic, open-domain retrieval queries from scientific text.

B.2 Query Decontextualization Prompt

PROMPT

You are a scientific question rewriter. You are given an original question that references a specific portion of a research paper. Your task is to rewrite it into a **context-independent, open-domain scientific query** that targets the same underlying concept, without relying on document-local or visual references.

Do **not** refer to any figure, plot, panel, image, document, or use deictic expressions such as this, that, above, or below.

Requirements:

1. Preserve the core scientific intent, variables, and conditions present in the original question.
2. Replace visual or deictic phrasing with concept-level wording (e.g., remove references such as "based on the graph" and ask directly about the relationship or effect).
3. If symbols (e.g., f_{spec}) appear without definition, retain them exactly as written and do not invent meanings. A minimal parenthetical alias may be included only if it appears in the input.
4. Remove all references to figures, plots, tables, panels, or document-local indices.
5. Ensure the rewritten query can be answered by a knowledgeable reader without access to the original document or image.
6. Retain units, ranges, and experimental or observational conditions if present.
7. Avoid unresolved pronouns or placeholders (e.g., "the parameter", "the system") unless the domain makes them unambiguous.
8. If the original question contains multiple sub-questions, keep only one and discard the rest.
9. The final query must be a single, concise sentence with no superfluous framing or background.

Required Output Format:

```
{
  "query": "<single rewritten question or null>",
  "reasoning": "<one-sentence rationale>"
}
```

If a valid context-independent query cannot be produced, set "query" to null and briefly explain why in "reasoning".

Figure 5: Prompt used for decontextualizing document-dependent scientific questions into open-domain queries.

B.3 Query Verification Prompt

PROMPT

You are a validator that checks whether a decontextualized question is well-formed for open-domain retrieval. Judge only from the provided JSON fields. Do **not** assume access to the original figure, table, or paper.

What "valid decontextualized question" means:
A question is **valid** if and only if all of the following criteria are satisfied:

1. **Context-independent**: The question contains no references to local context such as "this figure," "the table above," "these results," or any indexical phrasing that requires the original document or image.
2. **Answerable in principle**: A knowledgeable person or external source could answer the question without access to the original paper or figure. The domain and variables must be sufficiently specified. Crucially, the question must not rely on parameters, symbols, or notations that are defined arbitrarily or only within the source paper (e.g., a tuning parameter with no standard meaning in the field).
3. **Intent preserved**: The question targets the same underlying information need as the original question, but generalized beyond the local figure or document context.
4. **Clarity and unambiguous entities**: Any entities, variables, or notations must be interpretable by an expert in the relevant field without requiring the specific paper. Unresolved pronouns or placeholders (e.g., "the parameter," "the system") are not allowed unless they are standard and unambiguous within the domain.

Guiding Principle for Ambiguity:
Requiring background domain knowledge is acceptable and expected for real search queries. However, ambiguity arising from terms that are defined only within the source document or that depend on the original figure context is not acceptable.

Common failure modes (label them if present):

- underspecified_parameter (especially if defined arbitrarily in the source paper)
- still_context_bound
- domain_missing_or_vague
- ambiguity_pronouns_placeholders
- unanswerable_generic

Required Output Format (JSON only):

```
{
  "is_valid": boolean,
  "score": integer,
  "decision_rationale": string,
  "confidence": integer
}
```

Scoring rubric:

- **0** (invalid): fails context-independence or answerability
- **1** (weak): partially decontextualized but still problematic
- **2** (good): valid with minor tightening possible
- **3** (excellent): clearly valid, crisp, and widely answerable

Figure 6: Prompt used to verify whether a generated question is a valid, context-independent query suitable for open-domain retrieval.

C Example Queries

arXiv ID: 0810.1349
Which experimental parameter ranges (such as filament grafting density, filament length and stiffness, interparticle separation, and ionic strength) would make crowding of surface-anchored rodlike filaments, rather than ion-mediated osmotic effects, the dominant origin of observed interparticle repulsion?

arXiv ID: 1108.3966
What experimental observations indicate that energy relaxation of individual circuit elements is the dominant factor limiting the fidelity of a three-party controlled logical operation?

arXiv ID: 1801.02602
What constraints on near-term demonstrations claiming to outperform conventional simulators follow if modest-scale coherent information processors with local imperfections produce measurement outcomes that are well described by simple low-complexity analytic models and show vanishing alignment with ideal noiseless behavior?

arXiv ID: 1811.08489
What conditions on data collection, annotation practices, or available observable features are required for equalizing label–attribute co-occurrence to reliably prevent models from exaggerating associations between a sensitive attribute and predicted targets?

arXiv ID: 1210.1079
What primary physical mechanism determines the maximum distance over which a high-density coherent quasiparticle pulse can be sustained by periodic traversal through a localized optically pumped gain region?

arXiv ID: 1711.03114
How reliably can first-year, single-season, intermittently sampled monitoring of a flux-limited quasar sample be used to revise the empirical scaling between emission-region size and source luminosity for luminous, high-redshift quasars with long intrinsic echo times?

arXiv ID: 1604.06227
How does the presence of rigid, crystalline-like cluster cores that travel with persistent momentum and produce highly ramified, low-mass aggregates affect the applicability of diffusion- or hydrodynamics-based late-stage coarsening predictions for two-dimensional fluid-to-solid quenches?

Figure 7: Representative examples of decontextualized, evidence-grounded queries in ArXivDocQA. Each query targets a specific piece of scientific evidence.

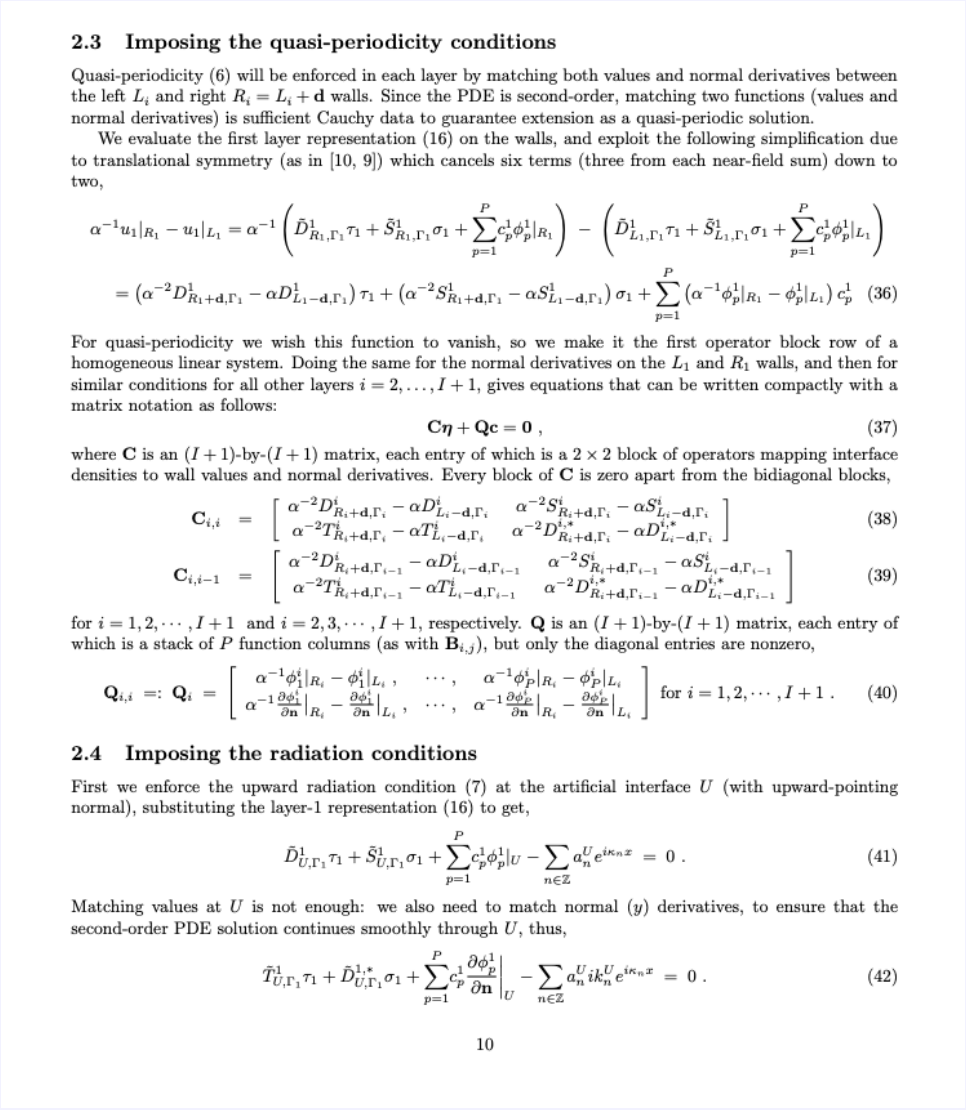
D Qualitative Failure Analysis Examples

In several cases, the evidence is expressed through extended textual and mathematical discussion rather than through a single visually distinctive figure or page. In these cases, page-level image representations lack stable semantic anchors and tend to retrieve pages that are visually similar but semantically unrelated (e.g., equation-dense pages from different papers). Text-based representations succeed by directly capturing the discourse-level scientific content required by the query.

Illustrative example. Figure 8 shows a representative example. The relevant evidence is not localized to a single figure; instead, it must be inferred from the mathematical formulation and accompanying discussion of the quasi-periodic operator construction and near–far field decomposition. As a result, document-as-image retrieval returns visually similar but incorrect pages from other papers, while text-based retrieval correctly identifies the gold document.

Query: How does increasing the horizontal repeat length relative to the operating wavelength affect the difficulty of approximating contributions from distant repeats and the computational work required per layer?

Gold document: arXiv:1410.5003



2.3 Imposing the quasi-periodicity conditions

Quasi-periodicity (6) will be enforced in each layer by matching both values and normal derivatives between the left L_i and right $R_i = L_i + \mathbf{d}$ walls. Since the PDE is second-order, matching two functions (values and normal derivatives) is sufficient Cauchy data to guarantee extension as a quasi-periodic solution.

We evaluate the first layer representation (16) on the walls, and exploit the following simplification due to translational symmetry (as in [10, 9]) which cancels six terms (three from each near-field sum) down to two,

$$\alpha^{-1}u_1|_{R_1} - u_1|_{L_1} = \alpha^{-1}\left(\tilde{D}^1_{R_1,\Gamma_1}\tau_1 + \tilde{S}^1_{R_1,\Gamma_1}\sigma_1 + \sum_{p=1}^{P} c_p^1\phi_p^1|_{R_1}\right) - \left(\tilde{D}^1_{L_1,\Gamma_1}\tau_1 + \tilde{S}^1_{L_1,\Gamma_1}\sigma_1 + \sum_{p=1}^{P} c_p^1\phi_p^1|_{L_1}\right)$$
$$= \left(\alpha^{-2}D^1_{R_1+\mathbf{d},\Gamma_1} - \alpha D^1_{L_1-\mathbf{d},\Gamma_1}\right)\tau_1 + \left(\alpha^{-2}S^1_{R_1+\mathbf{d},\Gamma_1} - \alpha S^1_{L_1-\mathbf{d},\Gamma_1}\right)\sigma_1 + \sum_{p=1}^{P}\left(\alpha^{-1}\phi_p^1|_{R_1} - \phi_p^1|_{L_1}\right)c_p^1 \quad (36)$$

For quasi-periodicity we wish this function to vanish, so we make it the first operator block row of a homogeneous linear system. Doing the same for the normal derivatives on the L_1 and R_1 walls, and then for similar conditions for all other layers $i = 2, \ldots, I+1$, gives equations that can be written compactly with a matrix notation as follows:

$$\mathbf{C}\boldsymbol{\eta} + \mathbf{Q}\mathbf{c} = \mathbf{0}\ , \quad (37)$$

where $\mathbf{C}$ is an $(I+1)$-by-$(I+1)$ matrix, each entry of which is a 2×2 block of operators mapping interface densities to wall values and normal derivatives. Every block of $\mathbf{C}$ is zero apart from the bidiagonal blocks,

$$\mathbf{C}_{i,i} = \begin{bmatrix} \alpha^{-2}D^i_{R_i+\mathbf{d},\Gamma_i} - \alpha D^i_{L_i-\mathbf{d},\Gamma_i} & \alpha^{-2}S^i_{R_i+\mathbf{d},\Gamma_i} - \alpha S^i_{L_i-\mathbf{d},\Gamma_i} \\ \alpha^{-2}T^i_{R_i+\mathbf{d},\Gamma_i} - \alpha T^i_{L_i-\mathbf{d},\Gamma_i} & \alpha^{-2}D^{i,*}_{R_i+\mathbf{d},\Gamma_i} - \alpha D^{i,*}_{L_i-\mathbf{d},\Gamma_i} \end{bmatrix} \quad (38)$$

$$\mathbf{C}_{i,i-1} = \begin{bmatrix} \alpha^{-2}D^i_{R_i+\mathbf{d},\Gamma_{i-1}} - \alpha D^i_{L_i-\mathbf{d},\Gamma_{i-1}} & \alpha^{-2}S^i_{R_i+\mathbf{d},\Gamma_{i-1}} - \alpha S^i_{L_i-\mathbf{d},\Gamma_{i-1}} \\ \alpha^{-2}T^i_{R_i+\mathbf{d},\Gamma_{i-1}} - \alpha T^i_{L_i-\mathbf{d},\Gamma_{i-1}} & \alpha^{-2}D^{i,*}_{R_i+\mathbf{d},\Gamma_{i-1}} - \alpha D^{i,*}_{L_i-\mathbf{d},\Gamma_{i-1}} \end{bmatrix} \quad (39)$$

for $i = 1, 2, \cdots, I+1$ and $i = 2, 3, \cdots, I+1$, respectively. $\mathbf{Q}$ is an $(I+1)$-by-$(I+1)$ matrix, each entry of which is a stack of P function columns (as with $\mathbf{B}_{i,j}$), but only the diagonal entries are nonzero,

$$\mathbf{Q}_{i,i} =: \mathbf{Q}_i = \begin{bmatrix} \alpha^{-1}\phi_1^i|_{R_i} - \phi_1^i|_{L_i}\ , & \cdots, & \alpha^{-1}\phi_P^i|_{R_i} - \phi_P^i|_{L_i} \\ \alpha^{-1}\frac{\partial\phi_1^i}{\partial\mathbf{n}}|_{R_i} - \frac{\partial\phi_1^i}{\partial\mathbf{n}}|_{L_i}\ , & \cdots, & \alpha^{-1}\frac{\partial\phi_P^i}{\partial\mathbf{n}}|_{R_i} - \frac{\partial\phi_P^i}{\partial\mathbf{n}}|_{L_i} \end{bmatrix} \text{ for } i = 1, 2, \cdots, I+1\ . \quad (40)$$

2.4 Imposing the radiation conditions

First we enforce the upward radiation condition (7) at the artificial interface U (with upward-pointing normal), substituting the layer-1 representation (16) to get,

$$\tilde{D}^1_{U,\Gamma_1}\tau_1 + \tilde{S}^1_{U,\Gamma_1}\sigma_1 + \sum_{p=1}^{P} c_p^1\phi_p^1|_U - \sum_{n\in\mathbb{Z}} a_n^U e^{i\kappa_n x} = 0\ . \quad (41)$$

Matching values at U is not enough: we also need to match normal (y) derivatives, to ensure that the second-order PDE solution continues smoothly through U, thus,

$$\tilde{T}^1_{U,\Gamma_1}\tau_1 + \tilde{D}^{1,*}_{U,\Gamma_1}\sigma_1 + \sum_{p=1}^{P} c_p^1\left.\frac{\partial\phi_p^1}{\partial\mathbf{n}}\right|_U - \sum_{n\in\mathbb{Z}} a_n^U ik_n^U e^{i\kappa_n x} = 0\ . \quad (42)$$

10

Figure 8: Qualitative failure example where document-as-image retrieval fails while text-based retrieval succeeds. The relevant evidence resides in mathematical discussion rather than in page-level features.