MAGDA: A Clustering-Based Retrieval-Augmented Generation System for Gender Gap Analysis in Digital Libraries

Marta Santacroce[1,*], Riccardo Benassi[1], Michele Luca Contalbo[1], Matteo Paganelli[1], Sara Pederzoli[1], Maurizio Vincini[1] and Francesco Guerra[1]

[1]*University of Modena and Reggio Emilia, Modena, Italy*

Abstract

This work introduces MAGDA (**M**ulti-document **A**ggregation via **G**lobal **D**ocument-level clustering **A**rchitecture), a domain-specific RAG system that uses a clustering-based chunking and retrieval strategy to capture semantic relations across documents in the Gender*More corpus. By aggregating inter-document information, MAGDA improves the grounding and relevance of generated answers. We evaluate the system on a curated set of real queries and validated responses, showing that domain-adapted RAG pipelines combined with cross-document processing significantly enhance information access and explainability in specialized scientific digital libraries.

Keywords

Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Digital Libraries, Gender-Gap Analysis, Domain-Specific Retrieval, Clustering-Based Chunking, Inter-Document Reasoning, Explainable AI.

1. Introduction

The increasing availability of scientific documents in digital libraries has amplified the need for intelligent systems capable of navigating, interpreting, and extracting knowledge from complex, heterogeneous sources. Although recent progress in Large Language Models (LLMs) has markedly improved automatic text understanding, these models are still limited by their reliance on static training corpora and by the opacity of their internal representations. When confronted with structured scientific material, such as institutional reports [1], statistical summaries [2], and tables [3], LLMs often fail to provide grounded, traceable answers, exposing the need for methods that combine generative capabilities with explicit document access [4]. Retrieval-Augmented Generation (RAG) [5] has emerged as an effective paradigm for addressing these shortcomings within digital library ecosystems. By integrating generative models with retrieval from curated corpora, RAG systems allow responses to be aligned with selected sources, enhancing reliability, reducing hallucinations, and supporting transparent information access. Such properties are particularly relevant in academic scenarios, where reproducibility, provenance, and metadata quality constitute core requirements.

However, conventional RAG pipelines exhibit limited effectiveness in scientific scenarios where the documents present highly specific terminology [6], or require information across multiple cited documents [7]. In particular, their retrieval component is highly sensitive to chunk granularity: large chunks preserve broader context but produce noisier and less discriminative embeddings, while smaller chunks yield cleaner semantic representations yet frequently miss key information that may help disambiguate domain-specific terminology [8, 9]. As a consequence, off-the-shelf RAG systems often

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.

marta.santacroce@unimore.it (M. Santacroce); riccardo.benassi@unimore.it (R. Benassi); micheleluca.contalbo@unimore.it (M. L. Contalbo); matteo.paganelli@unimore.it (M. Paganelli); sara.pederzoli@unimore.it (S. Pederzoli); maurizio.vincini@unimore.it (M. Vincini); francesco.guerra@unimore.it (F. Guerra)

0009-0005-2351-7154 (M. Santacroce); 0009-0007-4819-259X (R. Benassi); 0009-0008-7526-8534 (M. L. Contalbo); 0000-0001-8119-895X (M. Paganelli); 0009-0000-7659-662X (S. Pederzoli); 0000-0001-9262-2939 (M. Vincini); 0000-0001-6864-568X (F. Guerra)

retrieve passages that are either overly generic or only partially relevant. Moreover, since documents in scientific collections frequently cite or depend on one another, relying exclusively on intra-document chunking strategies prevents the model from capturing cross-document semantic relations.

To address these limitations, we introduce MAGDA[1], a domain-specific RAG system featuring a inter-document clustering-based chunking and retrieval strategy. The method first segments documents into semantically dense units, then applies clustering across the entire corpus to group chunks that belong to the same conceptual topic, regardless of their original document. Each chunk is subsequently enriched with its cluster-level embedding, combining local specificity with global topic semantics. This representation enables retrieval to account for inter-document relations that are typically invisible to standard pipelines, increasing the discriminative power of segment representations. Building on this enriched semantic structure, MAGDA performs retrieval through a hybrid scoring mechanism. A lexical scorer (BM25-style) and a dense encoder-based similarity measure are jointly used to rank candidates, producing a top-k list that reflects both syntactic match and semantic affinity. These candidates are then refined by a cross-encoder re-ranking step, which evaluates query-chunk pairs in context, yielding a final set of passages used for generation.

To evaluate MAGDA in realistic conditions, we constructed a question-answering dataset of 82 manually designed queries derived from gender-gap reports and institutional analyses. These documents are curated within the Gender*More[2] initiative at the University of Modena and Reggio Emilia, and feature domain-specific terminology related to gender-gap analysis and frequent cross-references across the reports. The queries require a total of 35 documents, but the whole retrieval corpus is made of 831 documents. The results indicate that the dataset is challenging, as the best reported Recall@1 achieves only 68.29%. Moreover, MAGDA shows clear improvements over other baselines using different chunking strategies.

2. Related Work

In this section, we review the two lines of research most closely related to MAGDA. The first concerns *context structuring and chunking for RAG*, which seeks to improve retrieval by refining how text is segmented and semantically represented. The second focuses on *RAG systems for scientific literature*, where most advances center on citation recommendation and introduce domain-aware retrieval mechanisms and corpus-level semantics to identify relevant scholarly evidence.

2.1. Context structuring and chunking for RAG

Recent advances in RAG have highlighted the central role of text segmentation in structuring, retrieving, and synthesizing knowledge. One line of research focuses on improving segment granularity and semantic coherence. In dense retrieval settings, fine-grained segmentation has been shown to improve performance by enabling more precise alignment between queries and relevant content [8], while LLM-guided approaches such as LumberChunker [10] and MoC [11] leverage discourse and logical structure to produce semantically meaningful units rather than relying on fixed window sizes. Another direction integrates local and global information through hierarchical or graph-based chunking. Models such as RAPTOR [12], TreeRAG [13], and GraphRAG [14] construct multi-level or graph-structured views of documents, combining sentence-level evidence with higher-order summaries or community-level relations, thereby enabling coarse-to-fine retrieval and context aggregation. In parallel, some methods operate directly in latent space, modifying or bypassing explicit text segmentation. Late Chunking [15] and Landmark Embedding [16] rely on representative embeddings, suggesting that segmentation decisions can be deferred to the embedding stage. More recent approaches adapt segment granularity dynamically at query time. Systems such as MoG [17], DCS [18], and MixGR [19] dynamically select

[1] https://github.com/marta01santacroce/GENDERMORE.git
[2] https://gendermore.unimore.it/home (last accessed 09-12-2025)

variable-length units or fuse multi-granular evidence (sentences, propositions, subqueries) to better align retrieved content with the informational scope of the query.

In this landscape, MAGDA builds on these insights by organizing fine-grained segments through inter-document semantic clustering in embedding space. This approach enriches local evidence with global thematic structure, enabling precise retrieval and broader scientific sensemaking without relying on hierarchical indices or manually constructed knowledge graphs.

2.2. RAG for scientific literature

RAG for scientific literature mainly focuses on citation recommendation, where models retrieve papers relevant to a specific citation context. Early deep learning systems combined textual and scholarly signals: BERT-GCN [20] fuses contextual embeddings with citation graph structure, while Medic et al. [21] augment local citation contexts with metadata, showing that relevance depends on both semantic similarity and scholarly influence. A complementary line of research learns citation-informed dense representations. Models such as SciNCL [22] and SPECTER [23] apply contrastive learning over citation neighborhoods, shifting retrieval from surface similarity toward research affinity. These objectives act as domain adaptation layers that specialize otherwise general encoders for scholarly corpora. More recent approaches adopt hierarchical and two-stage retrieval. SymTax [24] performs fast semantic prefetching followed by taxonomic embedding–based reranking, whereas HAtten [25] uses hierarchical text encoding for coarse retrieval and a SciBERT [26] re-ranker for fine-grained matching. LLM-driven systems have also emerged, exemplified by CiteAgent [27], an autonomous GPT-4o–based agent capable of reading, searching, and recommending papers.

These modeling advances are supported by a growing ecosystem of benchmarks. Citation recommendation datasets cover both global queries, where a paper serves as the query and its citations as targets [28, 23] and local inline citations, using citation mentions within text as queries [21, 20, 27, 29].

3. Approach

MAGDA is a domain-specific RAG platform designed to provide evidence-grounded answers over a curated digital library on gender-gap analysis. The system integrates two complementary pipelines (see Figure 1): an offline **ingestion and indexing** stage, which transforms heterogeneous documents into retrieval-ready representations, and an online **retrieval–reranking–generation** stage, which handles user queries. During ingestion, documents are converted to plain text while preserving structural cues, segmented into semantically coherent chunks, and organized into clusters that capture both local content and cross-document thematic relations. This dual-level representation allows the system to exploit fine- and coarse-grained semantic signals when searching for relevant information. At query time, user inputs are mapped into the same embedding space and matched against the enriched document representations using a hybrid semantic–lexical retrieval strategy. Candidate passages are then refined through cross-encoder re-ranking to highlight the most pertinent evidence, which is finally provided as context to an LLM. In this way, the model generates answers that are strictly grounded in the retrieved material, cite sources explicitly, and offer concise explanations of each passage's relevance, ensuring both transparency and factual consistency.

3.1. Ingestion and Indexing Pipeline

In this first step, the documents from the Gender*More initiative, comprising bibliometric analyses, institutional evaluations, and research studies on gender disparities, are first extracted from PDF format into normalized plain text while retaining structural cues such as page boundaries, section headings, and linearized tables. The resulting text is segmented using a **structure-aware recursive splitter**[3] that preserves paragraph- and section-level coherence under constraints on length and overlap, ensuring both contextual integrity and compatibility with LLM context windows.

[3] Recursive Splitter (last accessed 09-12-2025)

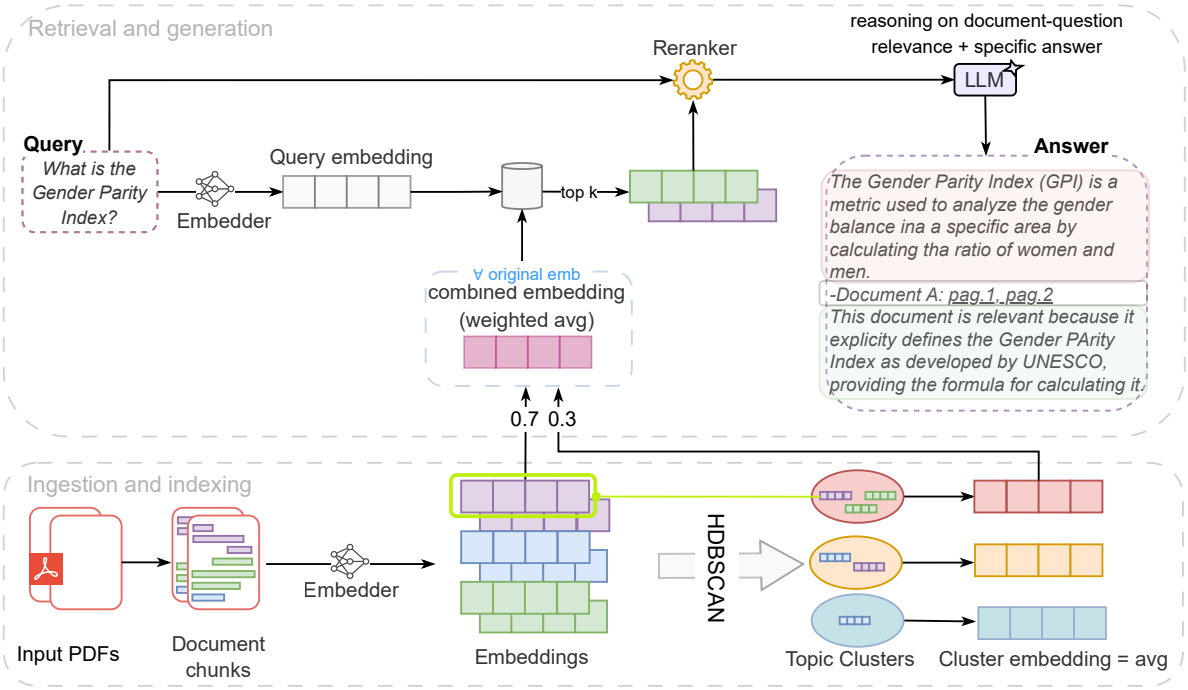

Figure 1: MAGDA Functional Architecture.

To overcome the limitations of purely local representations, MAGDA augments this segmentation with an inter-document clustering stage that captures higher-level thematic organization. Dense embeddings are computed for each chunk and subsequently grouped using the **HDBSCAN algorithm** [30] (with a minimum cluster size of five chunks), which identifies semantically coherent clusters extending across documents and produces centroid embeddings reflecting shared topic structure. Each chunk is enriched with its cluster-level representation, yielding a dual-context vector that combines fine-grained semantics with broader thematic cues. These two signals are integrated through a weighted combination of the chunk embedding and its cluster centroid, assigning weight 0.7 to the chunk embedding and 0.3 to the cluster embedding. All embeddings are computed using the `multilingual-e5-large-instruct` dense retriever [31].

The resulting enriched vectors, together with metadata such as document identifier, title, page number, section label, and cluster id, are stored in a PostgreSQL backend extended with PgVector to support efficient similarity search and metadata-aware retrieval.

3.2. Retrieval, Reranking, and Generation Pipeline

The retrieval and generation workflow in MAGDA follows a structured three-stage process. A user query is first encoded using the same `multilingual-e5-large-instruct` embedding model employed during ingestion, and matched against the vector representations stored in the vector database. Retrieval adopts a **hybrid strategy** that combines dense semantic similarity with a lexical component based on BM25-style scoring computed over the chunk text, enabling the system to capture both conceptual matches and exact term matches relevant for technical expressions, acronyms, and numeric values. The top-k candidate chunks returned by the hybrid retriever are then re-ranked with the **cross-encoder** model `ms-marco-MiniLM-L-6-v2`, which jointly encodes each query–chunk pair to produce a refined relevance score. This re-ranking step improves precision in the highest-ranked positions, a critical property when only a small evidence set can be inserted into the LLM context window.

The highest-ranked passages are injected into a prompt provided to **gpt-4o-mini**, which is instructed to generate answers grounded exclusively in the retrieved context, explicitly reference source documents, and avoid speculation when evidence is insufficient (see Figure 2). In parallel, the system produces concise explanations of each chunk's relevance, enhancing transparency and user trust (see Figure 3).

Answer Prompt

You are an AI assistant specialized in answering questions based on the past history and the provided context. Follow these guidelines:
1. Use only the given context and previous conversation to generate your answer;
2. Provide a clear and relevant response. Avoid unnecessary details;
3. Do NOT mention the context, sources, or any document references in your response.

If you cannot answer reply with this message: "I'm sorry but I could not find any relevant documents to answer your query. Try rephrasing the query or ask me something more specific. If you want a more in-depth explanation of the previous answer, take up the question you asked and ask me explicitly."

Figure 2: LLM response prompt.

Relevance Prompt

You are a helpful assistant that explains why each document was considered relevant. For each document, answer exactly in the following format:

Document: file_name.pdf;
Explanation: explanation text here.

Repeat for each document.

Figure 3: Prompt to explain the relevance of the document with reference to the query.

4. Experimental Evaluation

In this section, we present an experimental evaluation aimed at validating whether MAGDA effectively retrieves information relevant to user queries. In particular, we evaluate MAGDA through two experiments on a curated benchmark from the Gender*More corpus: one comparing standard chunking baselines, and one measuring the gains introduced by our clustering-based strategy. We also present a practical use case illustrating how MAGDA enables interpretable exploration of gender-gap indicators.

4.1. The Gender*More dataset

To evaluate MAGDA in a real operational scenario, we collaborated with domain experts to construct a domain-specific question answering dataset from the Gender*More corpus, specifically designed to assess retrieval-augmented generation performance. The dataset contains questions reflecting genuine information needs, including trends in the gender composition of academic staff (e.g., *"Does the academic system discriminate against women in the fields of mathematics and engineering?"*), differences in publication patterns (e.g., *"Is there a relationship between the gender of the last author and the order of the co-first authors?"*), and summaries of key findings from bibliometric analyses (e.g., *"What are the main findings of the bibliometric analysis on gender and sustainability research?"*). Each question is paired with its expected answer, the relevant document, and the exact passages annotated with page and section references, providing a precise ground-truth for evaluating both retrieval and generation components of RAG systems. The benchmark comprises 82 questions requiring a total of 35 documents, while the whole retrieval corpus is made of 831 documents.

4.2. Experimental Setup

We evaluate MAGDA in terms of both retrieval and generation quality. Retrieval performance is measured using Recall@k, which counts the proportion of queries for which the relevant document

Table 1
Retrieval and generation results across all evaluated chunking strategies on the top@20 retrieved documents. "*Recursive Splitter w/clust* –" denotes a weighted combination of the original chunk embedding and its corresponding cluster embedding, where weights and control the contribution of local (chunk-level) and global (cluster-level) semantic representations, respectively.

Chunking Strategy	*Generation*					*Retrieval*				
	$BERT_{f1}$	**BLEU**	**$R1_{f1}$**	**$R2_{f2}$**	**RL_{f1}**	**MRR**	**R@1**	**R@2**	**R@5**	**R@10**
Character Splitter	83.95	8.28	39.61	18.84	29.18	69.37	59.76	64.63	69.51	69.51
Semantic Splitter	83.93	8.36	38.84	17.44	27.83	69.39	59.76	65.85	69.51	73.17
Recursive Splitter	**84.56**	10.12	39.26	17.35	29.13	73.95	63.41	73.17	84.15	86.59
Recursive Splitter w/clust 0.9–0.1	84.16	**12.34**	**41.87**	**19.82**	**31.41**	74.62	64.63	75.60	85.36	89.02
Recursive Splitter w/clust 0.8–0.2	84.02	11.91	40.62	18.27	31.14	75.24	66.67	76.65	85.19	87.65
Recursive Splitter w/clust 0.7–0.3	83.84	12.33	40.33	18.18	30.42	**76.77**	67.07	**76.83**	**86.59**	**90.24**
Recursive Splitter w/clust 0.6–0.4	84.10	12.13	41.13	18.17	30.73	76.17	**68.29**	76.82	85.18	89.09
Recursive Splitter w/clust 0.5–0.5	83.73	11.91	40.26	18.05	30.02	76.64	**68.29**	76.81	85.36	87.80

appears among the top-k results, and Mean Reciprocal Rank (MRR), which reflects the ranking position of the correct document. Generation quality is assessed on the top-k retrieved chunks using BLEU, ROUGE, and BERTScore, capturing both n-gram overlap and contextual semantic similarity with the ground-truth answers.

Experiment 1: Baseline comparison. To contextualize MAGDA's performance, we compare it against baseline systems that share the same retriever, re-ranker, and generation components but differ in chunking strategy. In particular, we consider: (i) *Character Splitter*, which produces fixed-length chunks of 512 characters; (ii) *Semantic Splitter*, which merges consecutive sentences when their embeddings, computed with `multilingual-e5-large-instruct`, are sufficiently similar; and (iii) *Recursive Splitter*, which attempts to preserve larger textual units (e.g., paragraphs) and falls back recursively to finer-grained splitting when a 512-character limit is exceeded. This setup isolates the effect of segmentation granularity on downstream retrieval and generation quality.

Experiment 2: Clustering-enhanced chunking. In a second experiment, we evaluate the impact of our clustering-based strategy on retrieval and generation quality. All components of the RAG pipeline are kept fixed, while the chunk representation is augmented with a cluster-level embedding derived from inter-document semantic grouping. We test multiple weighting schemes between chunk-level and cluster-level embeddings (e.g., 0.9–0.1, 0.8–0.2, 0.7–0.3, 0.6–0.4, 0.5–0.5 allowing us to systematically analyze how different balances between fine- and coarse-grained semantic signals affect ranking precision and answer grounding. This experiment isolates the contribution of the clustering mechanism and quantifies the gains attributable to our proposed representation.

4.3. Results

The results reported at the top of Table 1 reveal a clear trend: traditional chunking strategies, like character, semantic and recursive splitting, provide reasonable performance but fall short compared to recursive splitting with inter-document clustering. Among the standard methods, the Recursive Splitter performs best, achieving an MRR of 73.95 and a Recall@10 of 0.86.59, together with the highest BERTScore F1 (84.56) and precision (83.83). Still, introducing inter-document representations leads to a substantial improvement. Recursive Splitter with inter-document clustering achieves the strongest retrieval performance overall, with an MRR of 76.77 and a Recall@10 of 90.24, the highest recorded score. Generation metrics also benefit from this strategy: BLEU increases to 12.33, ROUGE-1 to 40.33, ROUGE-2 to 18.18 and ROUGE-L to 30.42, all outperforming the traditional splitters. Overall, the evaluation demonstrates that, by adding inter-document representations for each chunk, the system consistently outperforms traditional chunking strategies, both in retrieval and generation.

The bottom part of Table 1 shows that even a minimal introduction of cluster-level information

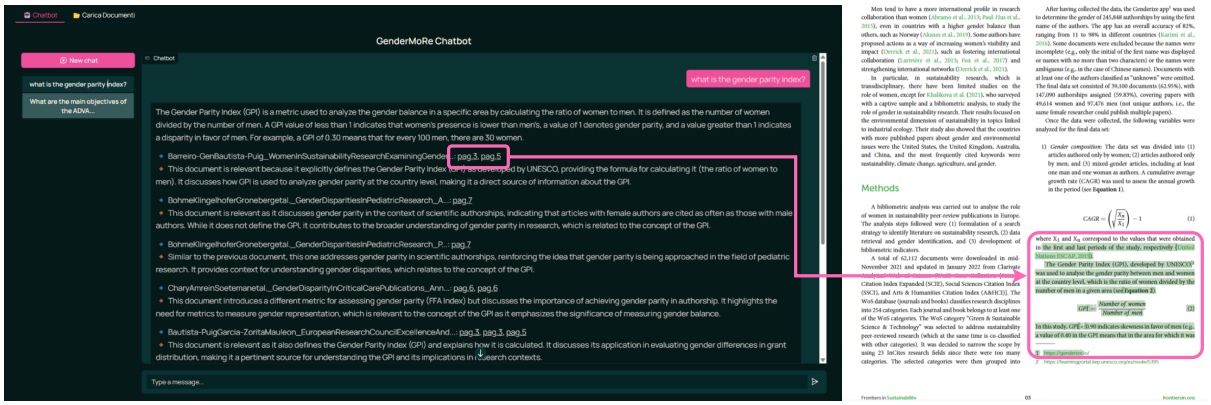

Figure 4: MAGDA in use. The system enables visualization of the referenced documents and highlights the sections of each document relevant to the query.

(0.9–0.1) improves MAGDA effectiveness in both retrieval and generation. The weighting scheme adopted in MAGDA (0.7–0.3) yields the highest retrieval performance, and affects generation quality only marginally.

4.4. Use case

Figure 4 shows a typical interaction with MAGDA during the exploration of gender-gap indicators. A user submits an analytical query through the interface, such as *"What is the Gender Parity Index?"*, and the system returns a grounded natural-language answer together with the most relevant document excerpts. Each retrieved passage is linked to its original source and exact location within the document.

The interface allows the user to (i) inspect the specific text segments used to produce the answer, (ii) visualize the sections within the original PDF, and (iii) read a concise explanation of why each document was selected as relevant. This interaction supports concrete analytical operations such as the verification of indicator definitions adopted in institutional reports, the rapid cross-checking of analytical claims across multiple sources, and the comparison of how the same metric is described in different studies. Furthermore, MAGDA can quickly obtain an overview of the existing literature for gender-gap analysis across different subtopics, such as trends in gender composition of academic staff and differences in publication patterns, allowing the user to move from targeted queries to a broader, cross-document understanding of the domain without manual document exploration.

5. Conclusions

We presented MAGDA, a domain-specific RAG system for evidence-based question answering over institutional documents for gender-gap analysis. MAGDA introduces a novel inter-document clustering-driven chunk enrichment strategy that integrates local and global semantic information into the retrieval process. The experimental results on the Gender*More dataset show that this approach consistently improves both retrieval accuracy and answer quality over standard chunking strategies.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: paraphrase and reword, improve writing style, grammar and spelling check. After using ChatGPT, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Y. Bai, S. Tu, J. Zhang, H. Peng, X. Wang, X. Lv, S. Cao, J. Xu, L. Hou, Y. Dong, J. Tang, J. Li, LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 3639–3664. URL: https://aclanthology.org/2025.acl-long.183/. doi:10.18653/v1/2025.acl-long.183.

[2] X. Li, Z. Li, C. Shi, Y. Xu, Q. Du, M. Tan, J. Huang, AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 773–783. URL: https://aclanthology.org/2024.lrec-main.69/.

[3] M. L. Contalbo, S. Pederzoli, F. D. Buono, V. Valeria, F. Guerra, M. Paganelli, GRI-QA: a comprehensive benchmark for table question answering over environmental data, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 15764–15779. URL: https://aclanthology.org/2025.findings-acl.814/. doi:10.18653/v1/2025.findings-acl.814.

[4] S. Jeong, J. Baek, S. Cho, S. J. Hwang, J. Park, Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7036–7050. URL: https://aclanthology.org/2024.naacl-long.389/. doi:10.18653/v1/2024.naacl-long.389.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[6] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, ACM Trans. Inf. Syst. 44 (2025). URL: https://doi.org/10.1145/3748304. doi:10.1145/3748304.

[7] J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, A. D. White, Paperqa: Retrieval-augmented generative agent for scientific research, 2023. URL: https://arxiv.org/abs/2312.07559. arXiv:2312.07559.

[8] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, D. Yu, Dense X retrieval: What retrieval granularity should we use?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15159–15177. URL: https://aclanthology.org/2024.emnlp-main.845/. doi:10.18653/v1/2024.emnlp-main.845.

[9] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, X. Huang, Searching for best practices in retrieval-augmented generation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 17716–17736. URL: https://aclanthology.org/2024.emnlp-main.981/. doi:10.18653/v1/2024.emnlp-main.981.

[10] A. V. Duarte, J. D. Marques, M. Graça, M. Freire, L. Li, A. L. Oliveira, LumberChunker: Long-form narrative document segmentation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 6473–6486. URL: https://aclanthology.org/2024.findings-emnlp.377/. doi:10.18653/v1/2024.findings-emnlp.377.

[11] J. Zhao, Z. Ji, Z. Fan, H. Wang, S. Niu, B. Tang, F. Xiong, Z. Li, MoC: Mixtures of text chunking learners for retrieval-augmented generation system, in: W. Che, J. Nabende, E. Shutova, M. T.

Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 5172–5189. URL: https://aclanthology.org/2025.acl-long.258/. doi:10.18653/v1/2025.acl-long.258.

[12] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, C. D. Manning, RAPTOR: recursive abstractive processing for tree-organized retrieval, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024. URL: https://openreview.net/forum?id=GN921JHCRw.

[13] W. Tao, X. Xing, Y. Chen, L. Huang, X. Xu, TreeRAG: Unleashing the power of hierarchical storage for enhanced knowledge retrieval in long documents, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 356–371. URL: https://aclanthology.org/2025.findings-acl.20/. doi:10.18653/v1/2025.findings-acl.20.

[14] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, From local to global: A graph RAG approach to query-focused summarization, CoRR abs/2404.16130 (2024). URL: https://doi.org/10.48550/arXiv.2404.16130. doi:10.48550/ARXIV.2404.16130. arXiv:2404.16130.

[15] M. Günther, I. Mohr, B. Wang, H. Xiao, Late chunking: Contextual chunk embeddings using long-context embedding models, CoRR abs/2409.04701 (2024). URL: https://doi.org/10.48550/arXiv.2409.04701. doi:10.48550/ARXIV.2409.04701. arXiv:2409.04701.

[16] K. Luo, Z. Liu, S. Xiao, T. Zhou, Y. Chen, J. Zhao, K. Liu, Landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 3268–3281. URL: https://aclanthology.org/2024.acl-long.180/. doi:10.18653/v1/2024.acl-long.180.

[17] Z. Zhong, H. Liu, X. Cui, X. Zhang, Z. Qin, Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025, Association for Computational Linguistics, 2025, pp. 5756–5774. URL: https://aclanthology.org/2025.coling-main.384/.

[18] B. Sheng, J. Yao, M. Zhang, G. He, Dynamic chunking and selection for reading comprehension of ultra-long context in large language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 31857–31876. URL: https://aclanthology.org/2025.acl-long.1538/. doi:10.18653/v1/2025.acl-long.1538.

[19] F. Cai, X. Zhao, T. Chen, S. Chen, H. Zhang, I. Gurevych, H. Koeppl, MixGR: Enhancing retriever generalization for scientific domain through complementary granularity, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 10369–10391. URL: https://aclanthology.org/2024.emnlp-main.579/. doi:10.18653/v1/2024.emnlp-main.579.

[20] C. Jeong, S. Jang, E. L. Park, S. Choi, A context-aware citation recommendation model with BERT and graph convolutional networks, Scientometrics 124 (2020) 1907–1922. URL: https://doi.org/10.1007/s11192-020-03561-y. doi:10.1007/S11192-020-03561-Y.

[21] Z. Medić, J. Snajder, Improved local citation recommendation based on context enhanced with global information, in: M. K. Chandrasekaran, A. de Waard, G. Feigenblat, D. Freitag, T. Ghosal, E. Hovy, P. Knoth, D. Konopnicki, P. Mayr, R. M. Patton, M. Shmueli-Scheuer (Eds.), Proceedings of the First Workshop on Scholarly Document Processing, Association for Computational Linguistics, Online, 2020, pp. 97–103. URL: https://aclanthology.org/2020.sdp-1.11/. doi:10.18653/v1/2020.sdp-1.11.

[22] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, G. Rehm, Neighborhood contrastive learning for scientific document representations with citation embeddings, in: Y. Goldberg, Z. Kozareva,

Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 11670–11688. URL: https://aclanthology.org/2022.emnlp-main.802/. doi:10.18653/v1/2022.emnlp-main.802.

[23] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, SPECTER: document-level representation learning using citation-informed transformers, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 2270–2282. URL: https://doi.org/10.18653/v1/2020.acl-main.207. doi:10.18653/V1/2020.ACL-MAIN.207.

[24] K. Goyal, M. Goel, V. Goyal, M. K. Mohania, Symtax: Symbiotic relationship and taxonomy fusion for effective citation recommendation, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 8997–9008. URL: https://doi.org/10.18653/v1/2024.findings-acl.533. doi:10.18653/V1/2024.FINDINGS-ACL.533.

[25] N. Gu, Y. Gao, R. H. R. Hahnloser, Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I, volume 13185 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 274–288. URL: https://doi.org/10.1007/978-3-030-99736-6_19. doi:10.1007/978-3-030-99736-6_19.

[26] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: https://aclanthology.org/D19-1371/. doi:10.18653/v1/D19-1371.

[27] O. Press, A. Hochlehnert, A. Prabhu, V. Udandarao, O. Press, M. Bethge, Citeme: Can language models accurately cite scientific claims?, in: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/0ef47f7b768e1a012e3d995ac8d8fac7-Abstract-Datasets_and_Benchmarks_Track.html.

[28] C. Bhagavatula, S. Feldman, R. Power, W. Ammar, Content-based citation recommendation, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 238–251. URL: https://aclanthology.org/N18-1022/. doi:10.18653/v1/N18-1022.

[29] A. Ajith, M. Xia, A. Chevalier, T. Goyal, D. Chen, T. Gao, LitSearch: A retrieval benchmark for scientific literature search, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15068–15083. URL: https://aclanthology.org/2024.emnlp-main.840/. doi:10.18653/v1/2024.emnlp-main.840.

[30] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 160–172.

[31] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).