

# CRAWLDoc: A Dataset for Robust Ranking of Bibliographic Documents

Fabian Karl<sup>1,\*</sup>, Ansgar Scherp<sup>1</sup>

<sup>1</sup>Universität Ulm, Germany

## Abstract

Publication databases rely on accurate metadata extraction from diverse web sources, yet variations in web layouts and data formats present challenges for metadata providers. This paper introduces CRAWLDoc, a new method for contextual ranking of linked web documents. Starting with a publication's URL, such as a digital object identifier, CRAWLDoc retrieves the landing page and all linked web resources, including PDFs, ORCID profiles, and supplementary materials. It embeds these resources, along with anchor texts and the URLs, into a unified representation. For evaluating CRAWLDoc, we have created a new, manually labeled dataset of 600 publications from six top publishers in computer science. Our method CRAWLDoc demonstrates a robust and layout-independent ranking of relevant documents across publishers and data formats. It lays the foundation for improved metadata extraction from web documents with various layouts and formats.

Our source code and dataset can be accessed at <https://github.com/FKarl/CRAWLDoc>.

## Keywords

Scholarly Dataset, Bibliographic Metadata, Information Retrieval, Language Model

## 1. Introduction

Databases such as Web of Science, Crossref, and DBLP are crucial academic resources of bibliographic information. Identifying high-quality metadata sources about new publications is essential for these services. While there are methods and tools for extracting bibliographic metadata [1, 2], these are typically restricted to a single document like a PDF. Currently, many potential web sources and content that may contain valuable metadata are underutilized. This is due to source heterogeneity of web layouts, document types, and formats, including full texts, publication PDFs, publisher landing pages, ORCIDs, and other web content.

We consider the example of DBLP, the de facto main metadata provider in computer science. The main strategy for integrating publisher-provided metadata is to implement publisher-specific wrappers, an approach that is time-consuming and requires maintenance whenever the publisher changes its website [3]. Thus, an automated service is needed to systematically search for bibliographic metadata sources across multiple web documents. Often, bibliographic information cannot be found on a single website, e. g., the publication's landing page, necessitating to harvest linked documents and identifying those relevant to the publication. Identifying relevant linked documents is challenging because two web documents with similar layouts and text can refer to different papers with paper-specific components like titles, authors, and affiliations. Another challenge is the heterogeneity of web data. Important documents can be in HTML or other formats like PDF. Another reason is that using wrappers or APIs relies on crawling publisher websites, which is expensive to maintain [3].

We propose a novel retrieval system CRAWLDoc (Contextual RAnking of Web-Linked Documents), see Figure 1, that can automatically identify relevant data sources from diverse web sources. Input is a Digital Object Identifier (DOI) of a publication, which is provided by publishers [4]. The web content linked from this seed URI is harvested and analyzed. We identify relevant linked content referring to the same paper as the DOI that may carry metadata. To this end, we embed the source document and linked

---

SCOLIA '25: First International Workshop on Scholarly Information Access (SCOLIA), April 10, 2025, Lucca, Italy

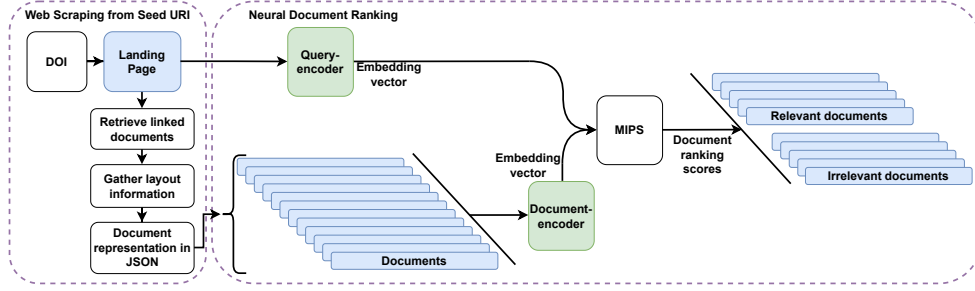
\*Corresponding author.

✉ [fabian.karl@uni-ulm.de](mailto:fabian.karl@uni-ulm.de) (F. Karl); [ansgar.scherp@uni-ulm.de](mailto:ansgar.scherp@uni-ulm.de) (A. Scherp)

ORCID [0009-0008-0079-5604](https://orcid.org/0009-0008-0079-5604) (F. Karl); [0000-0002-2653-9245](https://orcid.org/0000-0002-2653-9245) (A. Scherp)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



**Figure 1:** Illustration of document retrieval from heterogeneous web sources. The process starts with a DOI, which is resolved to access the landing page. Linked documents are ranked by treating the landing page as a query. A maximum inner product search (MIPS) ranks the top- $k$  documents based on embeddings from a small language model.

documents along with their associated anchor texts and URLs into a shared vector space and treat the publication’s landing page as the query. A ranking is computed by the similarity between the landing page embedding and the embeddings of linked documents, effectively identifying the most relevant sources for metadata extraction. By embedding the content, we effectively address the challenge that the web sources from which the data is extracted are highly diverse and vary in structure and format [3].

We evaluate CRAWLDoc on a new dataset derived from DBLP, which comprises 600 publications from the six largest computer science publishers. Our dataset is unique as it provides annotated relevancy labels for all outgoing links from publication landing pages, along with bibliographic metadata, including titles, years, authors’ names, and affiliations manually. Our experiments show that CRAWLDoc reliably identifies relevant web documents based on a single seed document.

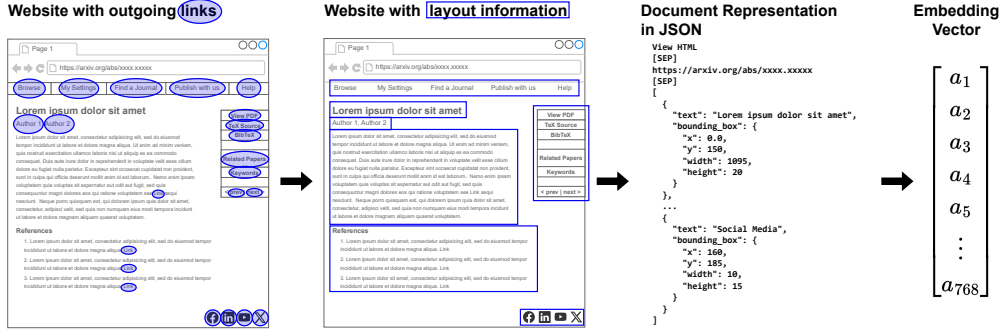
A leave-one-out experiment shows that our system is robust w. r. t. the retrieval from websites of various layouts from publishers that were excluded from the training dataset. In summary, our contributions are:

- A document-as-query approach CRAWLDoc to determine relevant documents that encodes web content of various formats, anchor text, and URIs in a single embedding space.
- Evaluating 600 publications from the six largest publishers in computer science.
- A robustness check by training on five publishers and testing on a held-out publisher.
- A new dataset of bibliographic metadata with author affiliations, along with relevancy information for linked web documents.

Below, we summarize related work. We introduce our CRAWLDoc metadata retrieval system in Section 3. The experimental apparatus is described in Section 4. The results are described in Section 5 and discussed in Section 6, before we conclude and outline future work.

## 2. Related Work

We discuss research in neural information retrieval and layout-aware language models. **Neural Information Retrieval** (NIR) is a prominent research area, utilizing neural networks to improve the retrieval process. The landscape of NIR research has been extensively surveyed [5, 6, 7], highlighting the use of learned representations of queries and documents, commonly referred to as embeddings. These embeddings capture semantic similarities that traditional information retrieval models might overlook [8, 9, 10]. The BERT model [11], although not specifically designed for information retrieval, has profoundly impacted NIR [12, 13, 14]. BERT-based models such as CEDR [15] have achieved impressive performance on various information retrieval benchmarks. The ColBERT model [16] introduced a late interaction paradigm, enabling efficient and effective passage retrieval. ColBERT’s ability to balance effectiveness and efficiency has made it a popular choice for large-scale retrieval tasks [17].



**Figure 2:** This figure illustrates our document representation methodology. The process begins with identifying all hyperlinks on the landing page, followed by integrating layout information via bounding boxes. The document is then converted into a uniform textual format and encoded into a vector representation.

**Layout-infused language models** consider both textual content and spatial layout. Layout-LMv3 [18] exemplifies this concept by pre-training multimodal transformers with a unified text and image masking objective. Another approach is DocLLM [19], which does not rely on expensive image encoders but relies solely on bounding box information from optical character recognition (OCR). LMDX [20] is a model-agnostic method to adapt arbitrary Large Language Models (LLMs) for document information extraction. It extracts text with OCR and enriches it with layout information. The model proposes an XML-style prompt for information extraction and trains a text-only LLM with text and bounding boxes. Layout-infused LLMs can face challenges with layout distribution shifts. Chen et al. [21] note that model performance can degrade by up to 20 points in macro F1 score under layout distribution shifts.

### 3. CRAWLDoc Metadata Retrieval

We introduce CRAWLDoc (Contextual Ranking of Web-Linked Documents), a novel system for identifying relevant bibliographic sources across web documents. Based on a seed URI, a DOI of a publication, CRAWLDoc scrapes linked resources. Subsequently, the retrieved web documents in the form of HTML or PDF are ranked using a Small Language Model (SLM) [22]. Our primary assumption is that all necessary information can be found within a one-hop crawl of the landing page associated with the DOI. This assumption is based on our observation that publishers present key bibliographic information on the landing page or pages directly linked to it e.g., the PDF of the publication.

**Web Scraping from Seed URI** The initial step of our system involves web scraping, starting with a DOI as the input and progressing to the scraping of the corresponding web page. After this starting point, all documents linked from the seed URI are retrieved, which may be formatted in HTML or PDF. Both PDF and HTML files undergo a series of steps to extract the relevant text and its associated bounding boxes to also capture layout information. For PDF documents, the text and its corresponding bounding box coordinates are directly extracted from the file using the PDFMiner Python library. In the case of HTML documents, the page is first rendered in a Firefox web browser (Version: 129.0.2) to accurately present the content’s formatting and layout, and then the text and bounding boxes are extracted. This information is then converted into a uniform textual JSON format and serves as the input for the neural document ranking. Figure 2 illustrates the different steps to create our document representation.

**Neural Document Ranking** In the second step, we employ a SLM to create embeddings of the documents along with their associated anchor texts and URLs. For each document, we construct a single input representation by concatenating the anchor text, URL, and document content using a special separator token ([SEP]). This representation is then embedded into a dense vector space. The document

---

originating from the DOI is embedded utilizing a query encoder, and all documents linked from the landing page are embedded with the document encoder. A Maximum Inner Product Search (MIPS) is performed with the embedding of the landing page and the embeddings of all scraped documents to create a Contextual Ranking of Web-Linked Documents (CRAWLDoc) based on the landing page.

We use the `jina-embeddings-v2` model [23] as neural retriever. It is based on a BERT [11] architecture and supports the symmetric bidirectional variant of ALiBi [24], allowing for a sequence length of up to 81,921 tokens. Due to memory restrictions, we limit our experiments to the first 2,048 tokens. The neural retriever is trained using contrastive learning with the InfoNCE loss function [25].

## 4. Experimental Apparatus

**Dataset** We take a subset of bibliographies from the six largest publishers in the DBLP Computer Science Bibliography dataset [26]. The publishers represent more than 80% of all publications listed in DBLP. This ensures the dataset contains a representative set of layouts encountered in bibliographic web content. We randomly select 100 publications for each publisher and split them into training, validation, and test sets in an 80/10/10 ratio with equal per-publisher distribution.

We obtained the metadata for each publication by manually retrieving the title, publication year, and authors’ names and affiliations. We retrieved the landing page of each publication and labeled every outgoing link on the landing page with a binary relevancy label. This label indicates whether the linked website or document is about the same publication. By manually creating this dataset, we ensure a high quality of the metadata and can accurately assess the document retrieval process in our proposed setup.

To prevent artificial inflation of our performance metrics, we identified and removed any instances in our test set where the landing page contained links to itself. The trivial nature of calculating document similarity to itself would otherwise result in an unrepresentative boost in ranking performance.

Our dataset consists of 600 publications with detailed metadata and 72,483 linked documents with binary relevancy labels. Per publication, we have an average of 3.63 (SD: 2.10) authors, with an average of 1.14 (SD: 0.41) affiliations per author. Furthermore, there is an average of 120.81 (SD: 76.52) linked websites per landing page and an average of only 5.45 (SD: 2.99) relevant websites per publication. To the best of our knowledge, we are the first to release a dataset that includes author affiliations as mentioned in the publications. Additionally, we are the first to provide relevancy labels for linked documents in the context of publication web data. For legal purposes, we are only able to publish the labels and not the actual websites. However, we do offer the source code for our procedure.

The DBLP dataset is released under CC0 1.0 Public Domain Dedication license. Our annotations have the same license.

**Procedure** Our experimental procedure for document ranking involves fine-tuning a neural document retriever using contrastive loss to improve document ranking. To ensure robust performance, we evaluate the ranking capabilities on both in-distribution and out-of-distribution data. We optimize the hyperparameters of the neural document retriever resulting in a learning rate of  $3e-05$ , 32 accumulation steps, and patience of 5, resulting in 16 epochs.

**Metrics** To evaluate the ranking of the web documents, we employ several metrics. The Mean Reciprocal Rank (MRR) evaluates the effectiveness of a retrieval system by considering the rank position of the first relevant result. The MRR focuses on the first relevant document in the ranked list, i. e., it favors a relevant document in the highest position. In contrast, Mean Average Precision (MAP) evaluates the precision of a retrieval system by averaging the precision scores at all ranks where relevant documents are found and then averaging these scores over all queries. Normalized Discounted Cumulative Gain (nDCG) [27] measures the usefulness of a document based on its position in the result list, assuming that highly relevant documents are more useful when appearing earlier. We further calculate the precision@k, recall@k, and F1@k which measure the proportion of relevant items in the top  $k$  results.

**Table 1**

Ranking performance metrics across publishers. Values are provided for each publisher and aggregated.

<b>Publisher:</b>	<b>IEEE</b>	<b>Springer</b>	<b>Elsevier</b>	<b>ACM</b>	<b>arXiv</b>	<b>MDPI</b>	<b>All</b>
<b>MRR</b>	1.000	0.800	1.000	1.000	1.000	1.000	0.967
<b>MAP</b>	1.000	0.998	0.970	0.999	1.000	0.954	0.987
<b>nDCG</b>	1.000	0.800	0.985	1.000	1.000	0.982	0.961

**Table 2**

Ranking performance evaluation of CRAWLDoc at different cut-off values  $k$ .

<b>k</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
<b>Recall@k</b>	0.344	0.551	0.692	0.792	0.870	0.900	0.932	0.941	0.943	0.951
<b>Precision@k</b>	0.972	0.892	0.822	0.754	0.698	0.617	0.557	0.500	0.455	0.416
<b>F1@k</b>	0.510	0.678	0.751	0.772	0.772	0.732	0.696	0.652	0.615	0.579

## 5. Results

The ranking metrics for identifying relevant linked documents are shown in Table 1. Overall, we achieve a very high average ranking performance with MRR of 0.967, MAP of 0.987, and nDCG of 0.961. The MRR, MAP, and nDCG values exhibit a consistently high level of performance for all six publishers, except for MRR and nDCG on the Springer dataset. The MRR for IEEE, Elsevier, ACM, arXiv, and MDPI all achieve the maximum score of 1.000, indicating that a relevant document is always placed at the top position. To understand the impact of layout information on ranking performance, we conducted an ablation study. The results without layout information showed slightly lower performance with a MRR of 0.950, MAP of 0.976, and nDCG of 0.952.

We have conducted a more detailed examination of the ranking performance with different cut-off values  $k$  presented in Table 2. The recall increases with increasing values of  $k$ , reaching 0.951 at  $k = 10$ . Precision declines from 0.972 for  $k = 1$  to 0.416 for  $k = 10$ . The F1@ $k$  score, which combines precision and recall, reaches its highest value of 0.772 for  $k = 4$  and  $k = 5$ .

We have evaluated robustness using a leave-one-out strategy, training on all but one publisher and testing on the left-out publisher. The results of the robustness analysis are shown in Table 3. We obtain a high performance across all publishers, with an average MRR of 0.959, MAP of 0.968, and nDCG of 0.961. This is less than one point for MRR and nDCG and less than two points for MAP compared to using the full training dataset shown in Table 1. The results were particularly strong for IEEE and arXiv, both achieving the maximum score of 1.000 for all three metrics. However, the performance was slightly lower for Springer, consistent with the result on the full training set.

## 6. Discussion

**Document Ranking** Our system shows impressive overall ranking performance, with documents ranked at the top being mostly relevant. This trend is seen when examining the evaluation of ranking performance at various cutoff values. We notice a sharp rise in recall@ $k$  for the first few documents but only minor enhancements after around five documents. The decline in precision@ $k$  as  $k$  values increase is a natural result considering that a publication has on average 5.45 relevant documents (see Section 4). This is also reflected in the F1@ $k$  score, which is peaking at  $k = 4$  and  $k = 5$ . Overall, the results show that CRAWLDoc maintains a good balance between precision and recall with a cut-off value of  $k = 5$ .

Upon investigating cases where the model ranked irrelevant documents higher than relevant ones, we could not identify a general error pattern. The errors we observed were predominantly paper-specific rather than systematic. For example, errors occurred when the model ranked links from the references



**Table 3**

Performance of the ranking task across all publishers in a leave-one-out test. The publisher the model is “tested on” is not part of the training data.

Tested on:	IEEE	Springer	Elsevier	ACM	arXiv	MDPI	Average
<b>MRR</b>	1.000	0.757	1.000	1.000	1.000	1.000	0.959
<b>MAP</b>	1.000	0.835	0.996	0.999	1.000	0.979	0.968
<b>nDCG</b>	1.000	0.772	0.999	1.000	1.000	0.992	0.961

section of a paper highly or when it assigned high ranks to different chapters of the same book. In particular, Springer publications presented more special cases than other publishers in our dataset.

**Robustness of Document Ranking** Our model demonstrates strong robustness across different publishers. While previous research, such as Chen et al. [21], has identified challenges for layout-infused LLMs when dealing with layout distribution shifts, our system shows consistent performance. This is evidenced by nearly equivalent performance between in-distribution and out-of-distribution data, suggesting effective generalization. Academic publishers often follow similar design patterns for their publication pages, reducing the effective layout distribution shift between sources. Our robustness evaluation considered six major publishers. The conventional nature of academic publication layouts suggests our model likely generalizes to a broader range of publishers.

**Generalization and Threat to Validity** The generalizability of our work refers to different publishers based on a leave-one-out test (Table 3) in which we test the system for publishers on which it has not been trained. Our robustness check demonstrates that a trained model can achieve comparably good results on out-of-distribution data within our tested scope. This finding suggests that our model has learned generalizable features of document relevance beyond the specific layouts and publishers in our training data. Our approach of transforming different document formats (HTML and PDF) into a uniform textual representation enhances its potential for generalization. This uniformity in representation suggests applicability to other web and document-related tasks.

While our study provides robust results, it is important to reflect on potential threats to validity. One such threat is the limited scope of our investigation, which focuses on only six publishers, primarily from the computer science field. However, the threat is reduced as these publishers represent more than 80% of computer science publications and provide various formats. Nevertheless, we acknowledge that true generalizability to the remaining 20% of publications, which may exhibit greater variability in their document layouts and metadata presentation, remains to be thoroughly tested in future work. An additional possible risk is the presence of recency bias in our dataset, given that most publications in DBLP are from more recent years. Nevertheless, we have found that older publications, including papers as far back as 1967, in our test set achieve similar performance to more recent ones, which eases this concern. This indicates that the performance of our model is not much influenced by the publication year.

Although there is no particular reason why other embedding models could not be used, our work does not focus on finding optimal embedding models for the retrieval tasks. We use Jina embeddings because they are widely used and have demonstrated strong results [23].

## 7. Conclusion and Future Work

Our Contextual RAnking of Web-Linked Documents (CRAWLDoc) retrieval system effectively identifies relevant bibliographic documents across diverse web sources. The key scientific findings include robustly identifying pertinent web documents and the system’s consistent performance across publishers with varying web layouts. The insights presented in this study can potentially advance the management and enrichment of comprehensive bibliographic databases.

---

Although our model’s performance is already very strong, rerankers [5] could improve document ranking accuracy. Future work could also explore alternative neural retriever setups like ColBERTv2 [17] and token-level representation of documents with MaxSim [16] instead of cosine similarity. In the next steps, we plan to run different metadata extractor components and setups on the CRAWLDoc-ranked list of web resources. Furthermore, we plan to evaluate CRAWLDoc in the context of the DBLP workflow.

## Acknowledgments

We thank Florian Reitz from DBLP for valuable feedback. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. This research is co-funded by the SmartER project (No. 515537520) of the DFG, German Research Foundation.

## Declaration on Generative AI

During the preparation of this work, the authors used Writefull and Grammarly to check spelling and grammar. They also used Writefull and Claude 3.5 Sonnet to improve the writing style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

## References

- [1] Grobid, <https://github.com/kermitt2/grobid>, 2008–2023.
- [2] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, *Int. J. Document Anal. Recognit.* 18 (2015) 317–335. URL: <https://doi.org/10.1007/s10032-015-0249-8>. doi:10.1007/s10032-015-0249-8.
- [3] R. Schenkel, Integrating and exploiting public metadata sources in a bibliographic information system, in: P. Mayr, I. Frommholz, G. Cabanac (Eds.), *Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018)*, Grenoble, France, March 26, 2018, volume 2080 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 16–21. URL: <https://ceur-ws.org/Vol-2080/paper2.pdf>.
- [4] M. Ley, DBLP - some lessons learned, *Proc. VLDB Endow.* 2 (2009) 1493–1500. URL: <http://www.vldb.org/pvldb/vol2/vldb09-98.pdf>. doi:10.14778/1687553.1687577.
- [5] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, J. Wen, Large language models for information retrieval: A survey, *CoRR abs/2308.07107* (2023). URL: <https://doi.org/10.48550/arXiv.2308.07107>. doi:10.48550/ARXIV.2308.07107. arXiv:2308.07107.
- [6] Y. Zhang, M. M. Rahman, A. Braylan, B. Dang, H. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. Lease, Neural information retrieval: A literature review, *CoRR abs/1611.06792* (2016). URL: <http://arxiv.org/abs/1611.06792>. arXiv:1611.06792.
- [7] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A deep look into neural ranking models for information retrieval, *Inf. Process. Manag.* 57 (2020) 102067. URL: <https://doi.org/10.1016/j.ipm.2019.102067>. doi:10.1016/J.IPM.2019.102067.
- [8] Y. Zhang, M. M. Rahman, A. Braylan, B. Dang, H. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. Lease, Neural information retrieval: A literature review, *CoRR abs/1611.06792* (2016). URL: <http://arxiv.org/abs/1611.06792>. arXiv:1611.06792.
- [9] B. Mitra, N. Craswell, An introduction to neural information retrieval, *Found. Trends Inf. Retr.* 13 (2018) 1–126. URL: <https://doi.org/10.1561/15000000061>. doi:10.1561/15000000061.

- 
- [10] Z. Abbasiantaeb, S. Momtazi, Text-based question answering from information retrieval and deep neural network perspectives: A survey, *WIREs Data Mining Knowl. Discov.* 11 (2021). URL: <https://doi.org/10.1002/widm.1412>. doi:10.1002/WIDM.1412.
- [11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: <https://doi.org/10.18653/v1/n19-1423>. doi:10.18653/v1/n19-1423.
- [12] X. Tian, J. Wang, Retrieval of scientific documents based on HFS and BERT, *IEEE Access* 9 (2021) 8708–8717. URL: <https://doi.org/10.1109/ACCESS.2021.3049391>. doi:10.1109/ACCESS.2021.3049391.
- [13] J. Wang, J. X. Huang, X. Tu, J. Wang, A. J. Huang, M. T. R. Laskar, A. Bhuiyan, Utilizing BERT for information retrieval: Survey, applications, resources, and challenges, *ACM Comput. Surv.* 56 (2024) 185:1–185:33. URL: <https://doi.org/10.1145/3648471>. doi:10.1145/3648471.
- [14] M. Li, É. Gaussier, Intra-document block pre-ranking for bert-based long document information retrieval - abstract, in: L. Tamine, E. Amigó, J. Mothe (Eds.), *Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022)*, Samatan, Gers, France, July 4-7, 2022, volume 3178 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: [https://ceur-ws.org/Vol-3178/CIRCLE\\_2022\\_paper\\_27.pdf](https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_27.pdf).
- [15] S. MacAvaney, A. Yates, A. Cohan, N. Goharian, CEDR: contextualized embeddings for document ranking, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, ACM, 2019, pp. 1101–1104. URL: <https://doi.org/10.1145/3331184.3331317>. doi:10.1145/3331184.3331317.
- [16] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over BERT, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, ACM, 2020, pp. 39–48. URL: <https://doi.org/10.1145/3397271.3401075>. doi:10.1145/3397271.3401075.
- [17] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, Colbertv2: Effective and efficient retrieval via lightweight late interaction, in: M. Carpuat, M. de Marneffe, I. V. M. Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Association for Computational Linguistics, 2022, pp. 3715–3734. URL: <https://doi.org/10.18653/v1/2022.naacl-main.272>. doi:10.18653/v1/2022.NAACL-MAIN.272.
- [18] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, Layoutlmv3: Pre-training for document AI with unified text and image masking, in: J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, L. Toni (Eds.), *MM '22: The 30th ACM International Conference on Multimedia*, Lisboa, Portugal, October 10 - 14, 2022, ACM, 2022, pp. 4083–4091. URL: <https://doi.org/10.1145/3503161.3548112>. doi:10.1145/3503161.3548112.
- [19] D. Wang, N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei, A. Nourbakhsh, X. Liu, Docllm: A layout-aware generative language model for multimodal document understanding, *CoRR abs/2401.00908* (2024). URL: <https://doi.org/10.48550/arXiv.2401.00908>. doi:10.48550/ARXIV.2401.00908. arXiv:2401.00908.
- [20] V. Perot, K. Kang, F. Luisier, G. Su, X. Sun, R. S. Boppana, Z. Wang, J. Mu, H. Zhang, N. Hua, LMDX: language model-based document information extraction and localization, *CoRR abs/2309.10952* (2023). URL: <https://doi.org/10.48550/arXiv.2309.10952>. doi:10.48550/ARXIV.2309.10952. arXiv:2309.10952.
- [21] C. Chen, Z. Shen, D. Klein, G. Stanovsky, D. Downey, K. Lo, Are layout-infused language models robust to layout distribution shifts? A case study with scientific documents, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics:*



- 
- ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 13345–13360. URL: <https://doi.org/10.18653/v1/2023.findings-acl.844>. doi:10.18653/v1/2023.FINDINGS-ACL.844.
- [22] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, M. Xu, Small language models: Survey, measurements, and insights, CoRR abs/2409.15790 (2024). URL: <https://doi.org/10.48550/arXiv.2409.15790>. doi:10.48550/ARXIV.2409.15790. arXiv:2409.15790.
- [23] M. Günther, J. Ong, I. Mohr, A. Abdessalem, T. Abel, M. K. Akram, S. Guzman, G. Mastrapas, S. Sturua, B. Wang, M. Werk, N. Wang, H. Xiao, Jina embeddings 2: 8192-token general-purpose text embeddings for long documents, CoRR abs/2310.19923 (2023). URL: <https://doi.org/10.48550/arXiv.2310.19923>. doi:10.48550/ARXIV.2310.19923. arXiv:2310.19923.
- [24] O. Press, N. A. Smith, M. Lewis, Train short, test long: Attention with linear biases enables input length extrapolation, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: <https://openreview.net/forum?id=R8sQPpGCv0>.
- [25] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748 (2018). URL: <http://arxiv.org/abs/1807.03748>. arXiv:1807.03748.
- [26] dblp Team, dblp computer science bibliography – Monthly Snapshot XML Release of April 2024, 2024. URL: <https://doi.org/10.4230/dblp.xml.2024-04-01>. doi:10.4230/dblp.xml.2024-04-01.
- [27] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: <http://doi.acm.org/10.1145/582415.582418>. doi:10.1145/582415.582418.