ReNeuIR at SIGIR 2025: The Fourth Workshop on Reaching Efficiency in Neural Information Retrieval

Sebastian Bruch
Northeastern University
Boston, MA, United States
s.bruch@northeastern.edu

Maik Fröbe
Friedrich-Schiller-Universität Jena
Jena, Germany
maik.froebe@uni-jena.de

Tim Hagen
University of Kassel and hessian.AI
Kassel, Germany
tim.hagen@uni-kassel.de

Franco Maria Nardini
ISTI-CNR
Pisa, Italy
francomaria.nardini@isti.cnr.it

Martin Potthast
University of Kassel,
hessian.AI, and ScaDS.AI
Kassel, Germany
martin.potthast@uni-kassel.de

Abstract

Measuring effectiveness and efficiency in information retrieval has a strong empirical background. While modern retrieval systems substantially improve effectiveness, the community has not yet agreed on how to measure efficiency, making it difficult to contrast effectiveness and efficiency fairly. Efficiency-oriented system comparisons are difficult due to factors such as hardware configurations, software versioning, and experimental settings. Efficiency affects users, researchers, and the environment and can be measured in many dimensions beyond time and space, such as resource consumption, water usage, and sample efficiency. Analyzing the efficiency of algorithms and their trade-off with effectiveness requires revisiting and establishing new standards and principles, from defining relevant concepts to designing new measures and guidelines to assess the findings' significance. ReNeuIR's fourth iteration aims to bring the community together to debate these questions and collaboratively test and improve benchmarking frameworks for efficiency based on discussions and collaborations of its previous iterations, including a shared task focused on efficiency and reproducibility.

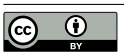
CCS Concepts

• **Information systems → Search engine architectures and scalability**.

Keywords

Efficiency, neural IR, sustainable IR, retrieval, ranking, algorithms

ACM Reference Format:
Sebastian Bruch, Maik Fröbe, Tim Hagen, Franco Maria Nardini, and Martin Potthast. 2025. ReNeuIR at SIGIR 2025: The Fourth Workshop on Reaching Efficiency in Neural Information Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3726302.3730358



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730358

1 Motivation and Theme

Modern information systems use complex algorithms to find relevant information. Recommendation systems help to discover movies, and search systems help to find answers to questions. *Retrieving and ranking* relevant items for explicit or implicit information needs from a collection is the common task of those applications.

Retrieval approaches evolved from Boolean retrieval over statistical models and feature-based Learning to Rank systems [34] to modern deep learning methods such as dense retrieval [27, 61], document expansion [51], large language models, etc. These developments mark the beginning of the new Neural Information Retrieval (NIR) era that enabled emerging applications like retrieval-augmented generation (RAG) [30]. Frequently, even larger deep learning models re-rank the candidates found by some smaller first-stage retriever, e.g., with expensive large language models significantly advancing the state-of-the-art in ranking [31, 49, 50, 52].

This shift from efficient statistical heuristics to more effective, but more computationally expensive, deep learning models became widespread in many areas of IR research. Although this advancement can substantially improve the effectiveness of search systems, it substantially increased the number of learnable parameters, requiring larger datasets and computational resources for training and inference, increasing the costs and the environmental impact. Consequently, the community questioned whether optimizing for effectiveness alone is the way to go and how to trade effectiveness for efficiency and examine the impact of such trade-offs [55, 57, 63].

Balancing efficiency and effectiveness for learning to rank systems motivated the research on *learning to efficiently rank* [10], leading to several innovations. For example, multi-stage rankers separate a light-weight retriever from more complex re-ranker(s) [1, 4, 16, 17, 33, 41, 58]. From probabilistic data structures [2, 3], to cost-aware training and *post hoc* pruning of decision forests [5, 18, 23, 37, 39, 48], to early-exit strategies and fast inference algorithms [6, 7, 13, 14, 29, 36, 38], the IR community considered the practicality and scalability of complex ranking algorithms, arriving at solutions that provide competitive trade-offs for huge datasets. As complex neural network-based models come to dominate ranking research, there is renewed interest in this research, with many recent approaches beeing inspired by past ideas [24, 28, 32, 35, 42–46, 50, 51, 56, 59, 60, 62], along with a series of novel approaches [22, 25, 26, 47, 53].

Despite these efforts, efficiency has typically been measured in terms of space or time, primarily for online inference. However, Scells et al. [55] showed that complex neural models are energy hungry. Moreover, energy consumption of models during training and inference are often ignored or under-reported, perhaps as an indirect result of effectiveness-driven competitions in IR [54].

Following this evolution of ideas, we propose the ReNeuIR workshop to discuss efficient and effective models in IR and their trade-offs to allow the community to discuss and debate these themes in the context of modern search systems. Following the successful previous editions of ReNeuIR, we will have (1) a scientific track for workshop papers and (2) a hands-on efficiency-oriented shared task to foster the development of new evaluation measures that paint a holistic picture of IR models by considering both efficiency and effectiveness. The shared task will run for the second time, so that previously submitted approaches serve as strong baselines. SIGIR is an appropriate venue for the proposed workshop as IR researchers—who increasingly use and develop neural network-based models—would help identify specific questions and challenges in this space, allowing us to define future directions collectively. We hope our forum will foster collaboration across interested groups.

2 Topics

To promote the themes discussed in the preceding section and enable a critical analysis and debate of each point, we solicit contributions on the following topics, including but not limited to: 1) Novel NIR models that reach competitive quality but are designed to provide fast training or fast inference; 2) Efficient NIR models for decentralized IR tasks such as conversational search; 3) Strategies to speed up training or inference of NIR models; 4) Sample-efficient training of NIR models; 5) Efficiency-driven distillation, pruning, quantization, retraining, and transfer learning; 6) Empirical investigation of the complexity of existing NIR models through an analysis of quality, interpretability, robustness, and environmental impact; and, 7) Evaluation protocols for efficiency in NIR.

3 Efficiency-Oriented IR Shared Task

The shared task aims to collect and measure NIR systems to encourage the development of new IR measures that incorporate efficiency and effectiveness. The methodology of the shared task was developed in the first two iterations of the ReNeuIR workshop [8, 9, 11] by comparing a range of different possibilities [12] and did run successfully in the previous 2024 edition of ReNeuIR [21].

Submitted retrieval pipelines need to (1) index, (2) retrieve, (3) and re-rank different workloads from ir_datasets [40] while their resource consumption on the same server will be monitored. As in 2024 [21], we will make all run files publicly available with their resource consumption. To enable a comparable and reproducible evaluation of systems, we will use TIRA/TIREx [19, 20].[1]

Within TIRA, we run the retrieval pipelines on different datasets derived from the MS MARCO passage dataset [15] while varying the number of queries and passages to cover different workloads. The software is executed in the TIRA sandbox, ensuring reproducibility (all dependencies must be installed in the uploaded Docker image) and keeping the query and document distribution of the test data secret. This also reveals cases where the systems efficiency highly depends on the observed distribution. Reusing MS MARCO lowers the barrier to entry, as many retrieval systems already exist for this benchmark, allowing participants to focus on efficiency.

To simplify participation, we have published public baselines with submission instructions.[2] We prepare the software required so that participants can compare the efficiency measurements we perform in TIRA with measurements on their own hardware. We encourage participants to run their Docker images on their systems as well so that multiple, diverse hardware setups can be used to monitor the resource consumption of systems. The use of Docker ensures reproducibility by enabling identical retrieval systems to be deployed in diverse environments. Our shared task primarily targets researchers with efficient, optimized retrieval systems, as well as those using standard implementations as audience.

4 Workshop Format

We plan ReNeuIR as a full-day workshop in-person-only, with the five organizers ensuring their presence onsite. We plan to have two keynote talks. The first is in the morning after the workshop's opening, and the second is at the beginning of the afternoon session. We plan to have from 5 to 10 shared task participants. The second session of the morning will be devoted to the technical presentation of the shared task, while a dedicated session after the first coffee break will be devoted to allowing shared task participants to present their activities and engage in the discussion among the workshop attendees. Additionally, we anticipate between 5 and 15 regular paper submissions (Section 2), aiming to accept ca. 5 to be presented in one session after the second keynote speaker in the afternoon (15 min per paper). We may also encourage "encore" presentations from the main conference or other late-breaking work. After the afternoon coffee break, we plan to have a 1.5-hours slot devoted to identifying the lessons learned and brainstorming on improving the shared task's design. The panel aims to engage the workshop attendees and allow them to contribute to the discussion. Table 1 summarizes our tentative schedule. Based on the first three editions of ReNeuIR at SIGIR 2022—2024, we expect 40 to 60 participants.

Table 1: Tentative schedule of the workshop.

09:00 - 09:15 Welcome and Introduction
09:15 - 10:00 Keynote Talk I
10:00 - 10:30 Recap of the Shared Task Initiative
10:30 - 11:00 Coffee Break
11:00 - 12:30 Presentations from Shared Task Participants
12:30 - 13:30 Lunch Break
13:30 - 14:15 Keynote Talk II
14:15 - 15:30 Paper Session
15:30 - 16:00 Coffee Break
16:00 - 17:30 Panel Discussion and Concluding Remarks

5 Selection Process

Submitted papers will be selected following an anonymous review from three PC members; invited talks and presentations will be facilitated directly by the organizing committee.

[1] Overview of the infrastructure: https://webis.de/facilities.html

[2] https://github.com/reneuir/reneuir-code

6 Expected Target Audience

For our scientific call (similar to the previous editions of ReNeuIR), we target a diverse audience of researchers, practitioners, and industry professionals focused on advancing efficiency to enable a holistic efficiency/effectiveness evaluation of modern retrieval systems. For the shared task, we target researchers and students working on efficient, optimized systems and standard implementations.

7 Organization

Sebastian Bruch, co-organizer of the ReNeuIR workshops at SIGIR in 2022 and 2023, completed his Ph.D. in Computer Science at the University of Maryland, College Park in 2013. His research since has centered around probabilistic data structures and approximate algorithms for retrieval, efficient and effective algorithms for learning-to-rank, and stochastic ranking functions. He is the author of "Foundations of Vector Retrieval" and the co-author of "Efficient and Effective Tree-based and Neural Learning to Rank." Sebastian's works have appeared in leading IR and systems journals including ACM TOIS, TKDE, and FnTIR. Sebastian is currently a Senior Research Scientist at Northeastern University in Boston, Massachusetts, United States.

Maik Fröbe is a PhD student at the Webis group with research interests in information retrieval. He has co-organized the Touché shared task since 2020 and the SCAI shared task since 2021 and was the main organizer of task 5 at SemEval-2023. He is an active developer of TIRA [19]/TIREx [20], which improved the reproducibility of a number of shared tasks and has an archive of more than 500 research prototypes and is increasingly used in teaching initiatives.

Tim Hagen is a PhD student in information retrieval at the University of Kassel. He is an active developer of TIRA/TIREx, focussing on how to simplify the automatic collection of central experimental metadata on hardware configurations, code versions, and resource consumption, aiming to allow decentralized run submissions that still contain all central metadata for meaningful comparisons.

Franco Maria Nardini is a Senior Researcher with ISTI-CNR in Pisa, Italy. His research interests focus on Web Information Retrieval, Machine Learning, and Data Mining. He authored over 100 papers in peer-reviewed international journals, conferences, and other venues. He has been General Co-Chair of ECIR 2025, Program Committee Co-Chair of SPIRE 2023, Tutorial Co-Chair of ACM WSDM 2021, Demo Papers Co-Chair of ECIR 2021, and Organizer of ReNeuIR at SIGIR (2022, 2023, 2024). He is a co-recipient of the ACM SIGIR 2024 Best Paper Runner-Up Award, the ECIR 2022 Industry Impact Award, the ACM SIGIR 2015 Best Paper Award, and the ECIR 2014 Best Demo Paper Award. He is a member of the editorial board of ACM TOIS and a PC member of SIGIR, ECIR, SIGKDD, CIKM, WSDM, IJCAI, and ECML-PKDD.

Martin Potthast is professor at the University of Kassel, Germany. In his research on information retrieval and natural language processing, he has put a special focus on the reproducibility of experimental evaluations. With TIRA, he has introduced and presented the first working prototype of a cloud-based evaluation framework that implements the evaluation-as-a-service paradigm, enabling the submission of working software instead of merely software runs while minimzing organizer overhead. Martin is co-initiator of the PAN network for digital text forensics, hosted at the CLEF conference, where he has organized annual shared tasks since 2009. Since 2012, TIRA was the exclusive submission system. Besides PAN, Martin has co-initiated several shared tasks at various editions of WSDM, MediaEval, SemEval, CoNLL, and INLG, and most recently the annual Touché-Lab at CLEF on argument retrieval.

8 Program Committee

The tentative program committee of ReNeuIR 2024 is (in alphabetical order): Sohia Althammer (Cohere), Zhuyun Dai (Google), Luke Gallagher (RMIT University), Carlos Lassance (Cohere), Xueguang Ma (University of Waterloo), Sean MacAvaney (University of Glasgow), Matthias Petri (Amazon Alexa), Harrisen Scells (Leipzig University), Salvatore Trani (ISTI-CNR), Andrew Yates (University of Amsterdam), Shengyao Zhuang (University of Queensland).

9 Publicity Plan and Audience Reach

The organizers will circulate the Call for Papers (CfP) on relevant mailing lists and to research institutions participating in related research projects. Similar to its last iteration, the workshop will have a website and be present on social media[3] to announce the CfP and important dates. We publish news and updates before, during, and after the workshop to stimulate interest and extend our reach.

10 Related workshops

ReNeuIR started in 2022 at SIGIR in Madrid [8, 9] with two keynotes, three paper sessions, and a lively discussion to identify research gaps and future work. We discussed as a community that efficiency is not simply latency; that a holistic, concrete definition of efficiency is needed for researchers and reviewers; and found research gaps for efficiency-centered evaluations, datasets, platforms, and tools.

The second edition of ReNeuIR took place at SIGIR 2023 in Taipei [11]. Highlights were a keynote talk, a session of paper presentations, and a joint poster session with the GenIR and REML workshops co-hosted at SIGIR. The last session of the workshop has been devoted to discuss an efficiency-oriented shared task to develop fair *efficiency-first* execution and measurement frameworks.

The third edition of ReNeuIR was held at SIGIR 2024 in Washington [21] with two keynotes, three paper sessions (with invitations from the main conference), one session to discuss the results of the shared task with 60 submitted systems, and a joint poster session with the GenIR, IR-RAG, and LLM4Eval workshops.

We believe that the first three editions of ReNeuIR helped identify and unify a community of researchers who are active in the space of efficiency in IR, raising awareness of ongoing work and existing gaps. Moreover, with the three editions of the workshop we contribute to build a shared task initiative with a shared methodology and a software infrastructure for collaborative execution of prototypes. The fourth edition of ReNeuIR serves as a community-building exercise and a forum to keep the community abreast of the progress made over the elapsed year. Moreover, the fourth ReNeuIR workshop aims to push forward and consolidate the efficiency-oriented shared task that allows the development of new evaluation measures towards a more complete evaluation of neural models in IR that considers both efficiency and effectiveness. ReNeuIR 2025 will report on the results of the shared task, allowing discussions in the community to identify new challenges in this research area.

[3]Website: http://ReNeuIR.org; Social media like: https://twitter.com/ReNeuIRWorkshop

References

[1] N. Asadi. *Multi-Stage Search Architectures for Streaming Documents*. University of Maryland, 2013. Ph.D. Dissertation.

[2] N. Asadi and J. Lin. Fast candidate generation for two-phase document ranking: Postings list intersection with Bloom filters. In *Proc. CIKM*, pages 2419–2422, 2012.

[3] N. Asadi and J. Lin. Fast candidate generation for real-time tweet search with Bloom filter chains. *ACM Trans. Inf. Syst.*, 31(3):13.1–13.36, 2013.

[4] N. Asadi and J. Lin. Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures. In *Proc. SIGIR*, pages 997–1000, 2013.

[5] N. Asadi and J. Lin. Training efficient tree-based models for document ranking. In *Proc. ECIR*, pages 146–157, 2013.

[6] N. Asadi, J. Lin, and A. P. de Vries. Runtime optimizations for tree-based machine learning models. *IEEE Trans. Know. Data Eng.*, 26(9):2281–2292, 2014.

[7] L. Beretta, F. M. Nardini, R. Trani, and R. Venturini. An optimal algorithm for finding champions in tournament graphs. *IEEE Trans. Knowl. Data Eng*, 35(10): 10197–10209, 2023.

[8] S. Bruch, C. Lucchese, and F. M. Nardini. Report on the 1st workshop on reaching efficiency in neural information retrieval (ReNeuIR 2022) at SIGIR 2022. *SIGIR Forum*, 56(2), 2022.

[9] S. Bruch, C. Lucchese, and F. M. Nardini. ReNeuIR: Reaching efficiency in neural information retrieval. In *Proc. SIGIR*, pages 3462–3465, 2022.

[10] S. Bruch, C. Lucchese, and F. M. Nardini. Efficient and effective tree-based and neural learning to rank. *Found. Trnd. Inf. Retr.*, 17(1):1–123, 2023.

[11] S. Bruch, J. Mackenzie, M. Maistro, and F. M. Nardini. Reneuir at SIGIR 2023: The second workshop on reaching efficiency in neural information retrieval. In *Proc. SIGIR*, pages 3456–3459, 2023.

[12] S. Bruch, J. Mackenzie, M. Maistro, and F. M. Nardini. A proposed efficiency benchmark for modern information retrieval systems. 2023. URL https://reneuir.org/assets/pdfs/ReNeuIR_2023_benchmark_proposal.pdf.

[13] F. Busolin, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and S. Trani. Early exit strategies for learning-to-rank cascades. *IEEE Access*, 11:126691–126704, 2023.

[14] B. B. Cambazoglu, H. Zaragoza, O. Chapelle, J. Chen, C. Liao, Z. Zheng, and J. Degenhardt. Early exit optimizations for additive machine learned ranking systems. In *Proc. WSDM*, pages 411–420, 2010.

[15] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and J. Lin. MS MARCO: Benchmarking ranking models in the large-data regime. In *Proc. SIGIR*, pages 1566–1576, 2021.

[16] J. S. Culpepper, C. L. Clarke, and J. Lin. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proc. ADCS*, pages 17–24, 2016.

[17] V. Dang, M. Bendersky, and W. B. Croft. Two-stage learning to rank for information retrieval. In *Proc. ECIR*, pages 423–434. 2013.

[18] D. Dato, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Fast ranking with additive ensembles of oblivious and non-oblivious regression trees. *ACM Trans. Inf. Syst.*, 35(2):15.1–15.31, 2016.

[19] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, and M. Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Proc. ECIR*, pages 236–241.

[20] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. The Information Retrieval Experiment Platform. In *Proc. SIGIR*, pages 2826–2836, July 2023.

[21] M. Fröbe, J. Mackenzie, B. Mitra, F. M. Nardini, and M. Potthast. Reneuir at SIGIR 2024: The third workshop on reaching efficiency in neural information retrieval. In *Proc. of SIGIR*, pages 3051–3054. ACM, 2024.

[22] L. Gao, Z. Dai, and J. Callan. Understanding BERT rankers under distillation. In *Proc. ICTIR*, pages 149–152, 2020.

[23] A. Gigli, C. Lucchese, F. M. Nardini, and R. Perego. Fast feature selection for learning to rank. In *Proc. ICTIR*, page 167–170, 2016.

[24] M. Gordon, K. Duh, and N. Andrews. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proc. Workshop on Representation Learning for NLP*, pages 143–155, July 2020.

[25] S. Hofstätter, H. Zamani, B. Mitra, N. Craswell, and A. Hanbury. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proc. SIGIR*, pages 2021–2024, 2020.

[26] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. TinyBERT: Distilling BERT for natural language understanding. In *Proc. EMNLP Findings*.

[27] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proc. EMNLP*, 2020.

[28] C. Lassance and S. Clinchant. An efficiency study for SPLADE models. In *Proc. SIGIR*, pages 2220–2226, 2022.

[29] F. Lettich, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Parallel traversal of large ensembles of decision trees. *IEEE Trans. on Par. Dist. Sys.*, 30(9):2075–2089, 2019.

[30] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS*, 2020.

[31] J. Lin, R. Nogueira, and A. Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers, 2021.

[32] Z. Lin, J. Liu, Z. Yang, N. Hua, and D. Roth. Pruning redundant mappings in transformer models via spectral-normalized identity prior. In *Proc. EMNLP Findings*, 2020.

[33] S. Liu, F. Xiao, W. Ou, and L. Si. Cascade ranking for operational e-commerce search. In *Proc. SIGKDD*, pages 1557–1565, 2017.

[34] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trnd. Inf. Retr*, 3(3): 225–331, 2009.

[35] Z. Liu, F. Li, G. Li, and J. Cheng. EBERT: Efficient BERT inference with dynamic structured pruning. In *Proc. ACL-IJCNLP Findings*, pages 4814–4823, 2021.

[36] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Quickscorer: A fast algorithm to rank documents with additive ensembles of regression trees. In *Proc. SIGIR*, pages 73–82, 2015.

[37] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, and S. Trani. Post-learning optimization of tree ensembles for efficient ranking. In *Proce. SIGIR*, pages 949–952, 2016.

[38] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Exploiting CPU SIMD extensions to speed-up document scoring with tree ensembles. In *Proc. SIGIR*, pages 833–836, 2016.

[39] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, and S. Trani. X-DART: blending dropout and pruning for efficient learning to rank. In *Proc. SIGIR*, pages 1077–1080, 2017.

[40] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. Simplified data wrangling with ir_datasets. In *Proc. SIGIR*, pages 2429–2436, 2021.

[41] J. Mackenzie, J. S. Culpepper, R. Blanco, M. Crane, C. L. Clarke, and J. Lin. Query driven algorithm selection in early stage retrieval. In *Proc. WSDM*, pages 396–404, 2018.

[42] J. Mackenzie, A. Mallia, A. Moffat, and M. Petri. Accelerating learned sparse indexes via term impact decomposition. In *Proc. EMNLP Findings*, 2022.

[43] J. Mackenzie, A. Trotman, and J. Lin. Efficient document-at-a-time and score-at-a-time query evaluation for learned sparse representations. *ACM Trans. Inf. Syst.*, 41(4), 2023.

[44] A. Mallia, J. Mackenzie, T. Suel, and N. Tonellotto. Faster learned sparse retrieval with guided traversal. In *Proc. SIGIR*, pages 1901–1905, 2022.

[45] Y. Matsubara, T. Vu, and A. Moschitti. Reranking for efficient transformer-based answer selection. In *Proc. SIGIR*, pages 1577–1580, 2020.

[46] J. S. McCarley, R. Chakravarti, and A. Sil. Structured pruning of a BERT-based question answering model. *arXiv:1910.06360*, 2021.

[47] B. Mitra, S. Hofstätter, H. Zamani, and N. Craswell. Improving transformer-kernel ranking model using conformer and query term independence. In *Proc. SIGIR*, pages 1697–1702, 2021.

[48] F. M. Nardini, C. Rulli, S. Trani, and R. Venturini. Distilled neural networks for efficient learning to rank. *IEEE Trans. Knowl. Data Eng.*, 35(5):4695–4712, 2023.

[49] R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv:1901.04085*, 2020.

[50] R. Nogueira, W. Yang, K. Cho, and J. Lin. Multi-stage document ranking with BERT. *arXiv:1910.14424*, 2019.

[51] R. Nogueira, W. Yang, J. Lin, and K. Cho. Document expansion by query prediction. *arXiv:1904.08375*, 2019.

[52] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document ranking with a pretrained sequence-to-sequence model. In *Proc. EMNLP Findings*, pages 708–718, 2020.

[53] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*, 2020.

[54] K. Santhanam, J. Saad-Falcon, M. Franz, O. Khattab, A. Sil, R. Florian, M. A. Sultan, S. Roukos, M. Zaharia, and C. Potts. Moving beyond downstream task accuracy for information retrieval benchmarking. *arXiv:2212.01340*, 2022.

[55] H. Scells, S. Zhuang, and G. Zuccon. Reduce, Reuse, Recycle: Green information retrieval research. In *Proc. SIGIR*, pages 2825–2837, 2022.

[56] L. Soldaini and A. Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *Proc. ACL*, pages 5697–5708, 2020.

[57] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In *Proc. ACL*, pages 3645–3650, 2019.

[58] L. Wang, J. Lin, and D. Metzler. A cascade ranking model for efficient ranked retrieval. In *Proc. SIGIR*, pages 105–114, 2011.

[59] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proc. ACL*, 2020.

[60] J. Xin, R. Tang, Y. Yu, and J. Lin. BERxiT: Early exiting for BERT with better fine-tuning and extension to regression. In *Proc. EACL*, pages 91–104, 2021.

[61] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proc. ICLR*, 2021.

[62] S. Zhuang and G. Zuccon. Fast passage re-ranking with contextualized exact term matching and efficient passage expansion. In *Proc. Workshop on ReNeuIR at SIGIR*, 2022.

[63] G. Zuccon, H. Scells, and S. Zhuang. Beyond CO2 emissions: The overlooked impact of water consumption of information retrieval models. In *Proc. ICTIR*, pages 283–289, 2023.