SIGIR 2023 Workshop on Retrieval Enhanced Machine Learning (REML @ SIGIR 2023)

Michael Bendersky
Google Research
United States
bemike@google.com

Danqi Chen
Princeton University
United States
danqic@cs.princeton.edu

Fernando Diaz
Google Research
Canada
diazf@acm.org

Hamed Zamani*
University of Massachusetts Amherst
United States
zamani@cs.umass.edu

ABSTRACT

Most machine learning models are designed to be self-contained and encode both “knowledge” and “reasoning” in their parameters. However, such models cannot perform effectively for tasks that require knowledge grounding and tasks that deal with non-stationary data, such as news and social media. Besides, these models often require huge number of parameters to encode all the required knowledge. These issues can be addressed via augmentation with a retrieval model. This category of machine learning models, which is called *Retrieval-enhanced machine learning* (REML), has recently attracted considerable attention in multiple research communities. For instance, REML models have been studied in the context of open-domain question answering, fact verification, and dialogue systems and also in the context of generalization through memorization in language models and memory networks. We believe that the information retrieval community can significantly contribute to this growing research area by designing, implementing, analyzing, and evaluating various aspects of retrieval models with applications to REML tasks. The goal of this *full-day hybrid workshop* is to bring together researchers from industry and academia to discuss various aspects of retrieval-enhanced machine learning, including effectiveness, efficiency, and robustness of these models in addition to their impact on real-world applications.

ACM Reference Format:
Michael Bendersky, Danqi Chen, Fernando Diaz, and Hamed Zamani. 2023. SIGIR 2023 Workshop on Retrieval Enhanced Machine Learning (REML @ SIGIR 2023). In *Proceedings of SIGIR '23.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

The vast majority of machine learning (ML) systems are designed to be self-contained, with both knowledge and reasoning encoded in model parameters. They suffer from a number of major shortcomings that can be (fully or partially) addressed, *if machine learning models have access to efficient and effective retrieval models*:

- **Knowledge grounding:** A number of important real-world problems, often called knowledge-intensive tasks, require access to external knowledge. They include (among others) open-domain question answering, task-oriented dialogues, and fact checking [16]. Therefore, ML systems that make predictions solely based on the data observed during training fall short when dealing with knowledge-intensive tasks. In addition, ML models related to non-stationary domains, such as news or social media, can significantly benefit from accessing fresh data [1, 13]. An information retrieval (IR) system can decouple reasoning from knowledge, allowing it to be maintained and updated independent of model parameters at a cadence aligned with the corpus.
- **Generalization:** Recent work has shown that many ML models can significantly benefit from retrieval augmentation. For instance, kNN-LM [11] linearly interpolates large language model (LM) predictions with the nearest neighbors of the given context input. This approach does not even require further training or fine-tuning. The authors showed substantial improvements in terms of perplexity in both in-distribution and out-of-distribution test sets, demonstrating the generalization of this approach. kNN-LM together with other approaches, such as BERT-kNN [10] and hybrid dialogue systems [12, 19], suggest that enhancing ML models using retrieval will have a large impact on their generalization.
- **Significant growth in model parameters:** Since all the required information for making predictions is often encoded in the ML models' parameters, increasing their capacity by increasing the number of parameters generally leads to higher accuracy [5]. For example, the number of parameters used in LMs has increased from 94 million in ELMo [15] to 1.6 trillion in Switch Transformers [4], an over 16× increase in just three years (2018 – 2021). Despite these successes, improving performance by increasing the number of model parameters can incur significant cost and limit access to a handful of organizations that have the resources to train them [2]. As such, this approach is neither

[*] Corresponding author.

SIGIR '23, , Taipei, Taiwan
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

scalable nor sustainable in the long run, and providing access to a scalable large collection (or memory) can potentially mitigate this issue.

- **Interpretability and explainability:** Because the knowledge in training data is encoded in learned model parameters, explanations of model predictions often appeal to abstract and difficult-to-interpret distributed representations. By grounding inference on retrieved information, predictions can more easily be traced to specific data, often stored in a human-readable format such as text.

The mentioned issues have been also recognized by multiple research groups for various learning problems [11, 14, 16–18]. Existing research in this area – often referred to as retrieval-augmented, or retrieval-enhanced ML (or REML, for short) – has thus far been primarily driven from a machine learning perspective: developing ML models that can more effectively leverage retrieval models for accurate prediction. In other words, most existing efforts take the retrieval part of REML for granted. For instance, they use term matching retrieval models, such as BM25, or off-the-shelf dense retrieval models for retrieving documents, and their main focus is data augmentation based on the retrieval results.

Recent research has demonstrated that errors in many REML models are mostly due to the failure of retrieval model as opposed to the augmented machine learning model, confirming the well-known "garbage in, garbage out" phenomenon [9]. Motivated by this observation and a recent perspective paper by Zamani et al. [20], we believe that the expertise of the IR research community is pivotal for further progress in REML models. Therefore, we propose to organize a workshop with a fresh perspective on retrieval-enhanced machine learning through an information retrieval lens.

2 THEME AND SCOPE

The workshop will focus on models, techniques, data collections, and evaluation methodologies for various retrieval-enhanced machine learning problems. These include but are not limited to:

- Effectiveness and/or efficiency of retrieval models for knowledge grounding, e.g., for open-domain question answering, dialogue systems, fact verification, and information extraction.
- Effectiveness and/or efficiency of retrieval models for generalization through memorization, e.g., nearest neighbor language models.
- Effectiveness and/or efficiency of retrieval models for memory networks.
- Effectiveness and/or efficiency of retrieval models for retrieval-augmented representation learning.
- Retrieval-enhanced optimization.
- Retrieval-enhanced domain adaptation.
- Retrieval-enhanced models for multi-media and multi-modal learning.
- Query generation for retrieval-enhanced models.
- Retrieval result utilization by machine learning models.
- Interactive retrieval-enhanced machine learning models.
- Retrieval-enhanced models for non-stationary data, such as news, social media, etc.

3 FORMAT AND PLANNED ACTIVITIES

We plan to organize a **full-day hybrid workshop**. The tentative schedule is presented in Table 1.

4 SPECIAL REQUIREMENTS

We require sufficient infrastructure for hybrid organization of the workshop, which includes remote broadcasting of on-site presentations as well as remote presentations from online participants. At least two of the organizers will organize the workshop in person. We also require poster stands during the second half of the workshop.

5 ORGANIZERS

The organization team consists of active IR and NLP researchers from both academia and industry with recent experience on REML research.

Michael Bendersky Michael Bendersky is a Principal Software Engineer / Engineering Director at Google Research. He is currently managing a team whose mission is improving algorithms, models, and metrics for information discovery and quality across Google products. His recent research interests include neural ranking and retrieval, unbiased learning-to-rank, ranking ensembles, query understanding, dynamic content understanding, and more. Michael is a Distinguished Member of the ACM. He holds a Ph.D. from the University of Massachusetts Amherst, and a B.Sc. and M.Sc. from the Technion, Israel Institute of Technology. Michael co-authored over 80 publications. He served on program and organizing committees for multiple academic conferences, and co-organized tutorials at SIGIR 2015, SIGIR 2019, ICTIR 2019, and WSDM 2022. He co-authored two books in the "Foundations and Trends in Information Retrieval" series: "Information Retrieval with Verbose Queries", and "Search and Discovery in Personal Email Collections".

Danqi Chen is an Assistant Professor of Computer Science at Princeton University. Her research focuses on training, adapting, and understanding large language models, and developing scalable and generalizable NLP systems for question answering, information extraction, and conversational agents. She is particularly interested in combining large language models with knowledge retrieval. Before joining Princeton, Danqi worked as a visiting scientist at Facebook AI Research. She received her Ph.D. from Stanford University (2018) and B.E. from Tsinghua University (2012), both in Computer Science. Danqi is a recipient of a Sloan Fellowship, a Samsung AI Researcher of the Year Award, outstanding paper awards from ACL 2016, EMNLP 2017, and ACL 2022, and multiple faculty awards. She served as the program chair of AKBC 2021 and (senior) area chair for many NLP conferences and co-organized workshops and tutorials at NAACL 2016, NeurIPS 2017, ACL 2018, EMNLP 2019, ACL 2020, and EMNLP 2021.

Fernando Diaz is a research scientist at Google Research Montréal and an incoming Associate Professor in the Language Technologies Institute (LTI) at Carnegie Mellon University. Fernando's research focuses on the design of information access systems, including search engines, music recommendation services and crisis response

Table 1: Tentative Schedule for the REML Workshop at SIGIR 2023.

Time	Agenda	Comment
9 - 9:15	Opening	
9:15 - 10	Keynote	Potential keynote speaker: William W. Cohen
10 - 10:30	Two invited talks	Invited talks from published REML articles at major conferences.
10:30 - 11	coffee break	
11 - 12:30	Five paper presentations	See Section 7 for more details on paper selection.
12:30 - 1:30	Lunch break	
1:30 - 2:15	Keynote	Keynote speaker: to be determined.
2:15 - 3	Discussion panel - Part 1	Topic: The role of IR community in REML research, including 15 minutes Q&A
3 - 3:30	Coffee break	
3:30 - 4:15	Discussion panel - Part 2	Topic: The future of REML research, including 15 minutes Q&A
4:15 - 5	Poster presentation or spotlight talks	See Section 7 for more details on paper selection.

platforms. He is particularly interested in understanding and addressing the societal implications of artificial intelligence more generally. Previously, Fernando was the assistant managing director of Microsoft Research Montréal, where he also led FATE Montréal, and a director of research at Spotify, where he helped establish its research organization on recommendation, search, and personalization. Fernando's work has received special recognition and awards at SIGIR, CIKM, CSCW, WSDM, ISCRAM, and ECIR. He is the recipient of the 2017 British Computer Society Karen Spärck Jones Award and holds a CIFAR AI Chair. Fernando has co-organized several NIST TREC tracks, WSDM (2013), Strategic Workshop on Information Retrieval (2018), FAT* (2019), SIGIR (2021), and the CIFAR Workshop on Artificial Intelligence and the Curation of Culture (2019). He received his BS in Computer Science from the University of Michigan Ann Arbor and his MS and PhD from the University of Massachusetts Amherst.

Hamed Zamani is an Assistant Professor at the University of Massachusetts Amherst, where he also serves as the Associate Director of the Center for Intelligent Information Retrieval (CIIR), one of the top academic research labs in Information Retrieval worldwide. Prior to UMass, he was a Researcher at Microsoft working on search and recommendation problems. His research focuses on designing and evaluating (interactive) information access systems, including search engines, recommender systems, and question answering. His work has led to over 80 refereed publications in the field, including some recent work on the topic of REML [3, 6–9, 20]. His research has received a few Best Paper and Honorable Mentions from SIGIR, CIKM, and ICTIR. He is a recipient of the NSF CAREER Award and Amazon's Alexa Prize grants. He is an Associate Editor of the ACM Transactions on Information Systems (TOIS), has organized multiple workshops at SIGIR, RecSys, WSDM, and WWW conferences, and served as a PC Chair at SIGIR 2022 (Short Papers).

6 POTENTIAL PROGRAM COMMITTEE

The program committee of the REML workshop consists of experts from various communities, including IR, NLP, and ML, working of different aspects of REML. The potential program committee include:

- Qingyao Ai, Tsinghua University
- Andrew Drozdov, UMass Amherst
- William Cohen, Google Research
- Nick Craswell, Microsoft
- Zhuyun Dai, Google Research
- Jeff Dalton, University of Glasgow
- Mostafa Dehghani, Google Brain
- Kelvin Guu, Google Research
- Helia Hashemi, UMass Amherst
- Claudia Hauff, Spotify
- Sebastian Hofstätter, Cohere AI
- Mohit Iyyer, UMass Amherst
- Urvashi Khandelwal, Google Research
- Julia Kiseleva, Microsoft Research
- Donald Metzler, Google Research
- Rodrigo Nogueira, UNICAMP Brazil and NeuralMind
- Aleksandra Piktus, Hugging Face
- Maithra Raghu, Samaya AI
- Jason Weston, Meta AI
- Chenyan Xiong, Microsoft Research
- Andrew Yates, University of Amsterdam
- Guido Zuccon, University of Queensland

7 SELECTION PROCESS

The accepted papers will be selected through a peer review process. Each paper will be evaluated by at least three PC members based on their originality, significance, technical soundness, presentation and clarity. We particularly encourage work-in-progress submissions and those with innovative ideas whose follow up work can potentially be submitted to the next major IR conferences.

The most attractive accepted papers will be given oral presentation. The remaining papers will be presented during a poster session or will be given short spotlight presentations.

The proceedings of the REML workshop will be **non-archival** and authors can resubmit their work to other peer-reviewed venues.

8 EXPECTED AUDIENCE AND ADVERTISEMENT

Specifically, given the rapid growth of interest in retrieval-enhanced models, we expect the audience to consist of both academic and industrial researchers involved in information retrieval and natural language processing research and engineering. We plan on advertising the workshop through existing social media channels (e.g., Twitter, Mastodon) as well as email lists (e.g., SIGIR-List, Corpora-List).

9 RELATED WORKSHOPS

There have been numerous workshops on open-domain question answering, dialogue systems, fact checking, and information extraction. Even though these workshops are indirectly related to REML, they have not *directly* focused on REML research. The most relevant workshops to REML are the recent ICML 2022 Workshop on **Knowledge Retrieval and Language Models (KRLM)**[1] and the recent ACL 2022 Workshop on **Semiparametric Methods in NLP: Decoupling Logic from Knowledge (Spa-NLP)**.[2] Unlike KRLM and Spa-NLP, the REML workshop will focus on retrieval-enhanced machine learning from *an information retrieval perspective.*

ACKNOWLEDGMENT

This research was supported in part by the Center for Intelligent Information Retrieval and in part by the Google Visiting Scholar program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Spurthi Amba Hombaiah, Tao Chen, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Dynamic Language Models for Continuously Evolving Content. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2514–2524. DOI:http://dx.doi.org/10.1145/3447548.3467162

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. DOI:http://dx.doi.org/10.1145/3442188.3445922

[3] Andrew Drozdov, Shufan Wang, Razieh Rahimi, Andrew McCallum, Hamed Zamani, and Mohit Iyyer. 2022. You can't pick your neighbors, or can you? When and How to Rely on Retrieval in the KNN-LM. In *Empirical Methods in Natural Language Processing*.

[4] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv:2101.03961* (2021).

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[6] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided Transformer: Leveraging Multiple External Sources for Representation Learning in Conversational Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1131–1140. https://doi.org/10.1145/3397271.3401061

[7] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2021. Learning Multiple Intent Representations for Search Queries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 669–679. DOI: http://dx.doi.org/10.1145/3459637.3482445

[8] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022. FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation. *CoRR* abs/2209.14290 (2022).

[9] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022. Multi-Task Retrieval-Augmented Text Generation with Relevance Sampling. In *Proceedings of the ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

[10] Nora Kassner and Hinrich Schütze. 2020. BERT-kNN: Adding a kNN Search Component to Pretrained Language Models for Better QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3424–3430. DOI:http://dx.doi.org/10.18653/v1/2020.findings-emnlp.307

[11] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HklBjCEKvH

[12] Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-Augmented Dialogue Generation. *CoRR* abs/2107.07566 (2021). arXiv:2107.07566 https://arxiv.org/abs/2107.07566

[13] Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.), Vol. 162. PMLR, 13604–13622. https://proceedings.mlr.press/v162/liska22a.html

[14] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. *CoRR* abs/1906.00067 (2019). arXiv:1906.00067 http://arxiv.org/abs/1906.00067

[15] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *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, 2227–2237. DOI:http://dx.doi.org/10.18653/v1/N18-1202

[16] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL '21)*. Association for Computational Linguistics, Online, 2523–2544. DOI:http://dx.doi.org/10.18653/v1/2021.naacl-main.200

[17] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. 2021. Passage Retrieval for Outside-Knowledge Visual Question Answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1753–1757. DOI:http://dx.doi.org/10.1145/3404835.3462987

[18] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *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, 809–819. DOI:http://dx.doi.org/10.18653/v1/N18-1074

[19] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1341–1350. DOI: http://dx.doi.org/10.1145/3357384.3357881

[20] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2875–2886. DOI:http://dx.doi.org/10.1145/3477495.3531722

[1] https://knowledge-retrieval-workshop.github.io/

[2] https://www.semiparametric.ml/