The Second Tutorial on Retrieval-Enhanced Machine Learning: Synthesis and Opportunities

Fernando Diaz
Carnegie Mellon University
Pittsburgh, PA, United States
diazf@acm.org

Andrew Drozdov
Databricks
New York, NY, United States
andrew.drozdov@databricks.com

To Eun Kim
Carnegie Mellon University
Pittsburgh, PA, United States
toeunk@cs.cmu.edu

Alireza Salemi
University of Massachusetts
Amherst, MA, United States
asalemi@cs.umass.edu

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

Abstract

Retrieval-Enhanced Machine Learning (REML) refers to the use of information retrieval (IR) methods to support reasoning and inference in machine learning tasks. Although relatively recent, these approaches can substantially improve model performance. This includes improved generalization, knowledge grounding, scalability, freshness, attribution, interpretability, and on-device learning. To date, despite being influenced by work in the information retrieval community, REML research has predominantly been presented in natural language processing (NLP) conferences. Our tutorial addresses this disconnect by introducing core REML concepts and synthesizing the literature from various domains in machine learning (ML), including, but not limited to, NLP. What is unique to our approach is the use of consistent notations to provide researchers with a unified and expandable framework. The tutorial will be presented in lecture format based on an existing manuscript, with supporting materials and a comprehensive reading list available at a website. Building on the momentum of our successful workshop at SIGIR 2023 and our tutorial at SIGIR-AP 2024, this year's tutorial features updated content with an emphasis on retrieval technologies used across the broader ML community. We also highlight their role in emerging, future-facing applications such as language agents and evolving scenarios where the extensive body of knowledge from IR can provide critical insights and capabilities.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Machine learning**.

Keywords

Information Retrieval, Machine Learning, Retrieval-Augmented Generation, Large Language Models

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3731695

ACM Reference Format:
Fernando Diaz, Andrew Drozdov, To Eun Kim, Alireza Salemi, and Hamed Zamani. 2025. The Second Tutorial on Retrieval-Enhanced Machine Learning: Synthesis and Opportunities. 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/10.1145/3726302.3731695

1 Motivation

Retrieval systems, originally designed for human use, are increasingly being integrated into machine learning (ML) models to extend their access to information beyond fixed model parameters [27]. A growing body of research shows that incorporating retrieval enhances model performance by improving generalization, knowledge grounding, scalability, freshness, attribution, interpretability, and on-device learning [56]. Moreover, these systems function as an external memory for ML models or 'agents', effectively extending their cognition to the external environment [28, 47].

In light of this progress, Zamani et al. [56] introduced Retrieval-Enhanced Machine Learning (REML), a research program dedicated to the development of information retrieval techniques for artificial intelligence systems. Since its publication, the machine learning and natural language processing communities have continued advancing REML design and applications [2, 4, 8, 10, 39].

Despite its growing impact, much of the current REML research has remained disconnected from the Information Retrieval (IR) community. As a result, insights from decades of IR research remain underutilized. For instance, many works default to simple retrieval techniques like BM25. At the same time, retrieval methods to date have–with some exception–focused on their use by people as end users, not models.

This tutorial explores the integration of retrieval systems into ML models—extending beyond the traditional scope of NLP. We examine both historical and contemporary uses of retrieval in ML and synthesize methods across domains using consistent mathematical notation, a feature that is currently lacking in the literature.

Rather than organizing the tutorial around applications, we structure it around the components of the REML framework: *Querying*, *Searching*, *Presentation*, *Consumption*, *Storing*, *Optimization*, and

Authors are listed in alphabetical order.

Evaluation. This structure facilitates a deeper understanding of each component's role and its interaction within the framework, while enabling generalization to new domains.

2 Objectives

The goal of this tutorial is to provide information retrieval researchers with a clear and formal overview of the diverse approaches in Retrieval-Enhanced Machine Learning (REML), enabling them to quickly engage in research in this area. The tutorial builds upon the manuscript [27], co-authored by all presenters. Specifically, the objectives are to: (i) survey and synthesize the range of REML approaches using common strategies, (ii) connect abstract themes in REML to existing information retrieval research, and (iii) outline a set of future research directions for both the information retrieval and machine learning communities. The tutorial will formally define key retrieval enhancement strategies using consistent notation, providing researchers with a structured and accessible entry point into the field.

3 Relevance and Impact to the Community

The information retrieval community has a long history of collaboration with the natural language processing field, particularly in supporting tasks such as question answering [50]. Research aimed at improving these tasks is regularly published at SIGIR and examined in forums like TREC.

As described by Zamani et al. [56], Retrieval-Enhanced Machine Learning (REML) can be viewed as both an evolution and generalization of this thread of IR research. It modernizes traditional domains, such as retrieval-based question answering, by embedding them within contemporary deep learning architectures. Furthermore, it expands beyond these domains by recognizing that retrieval need not be limited to a single stage in a reasoning pipeline (e.g., candidate generation), nor constrained to text corpora. Instead, retrieval can serve as a mechanism for accessing knowledge and memory involving abstract representations and concepts. This tutorial provides a theoretical unification of common themes across REML, and surfaces opportunities for future research in the area.

In fact, the IR community has already begun moving in this direction, as seen in initiatives such as TREC RAG 2024[1] and the upcoming TREC Million LLM 2025.[2] These efforts are focused on developing standardized and effective evaluation strategies for RAG systems [38, 40], as well as adapting long-standing IR knowledge—including distributed IR, federated search, and meta-search—to design search systems for machine users [21, 44].

The growing industrial interest in REML is also evident from a surge in real-world AI applications equipped with retrieval systems [5, 23, 30, 53]. Thus, a comprehensive tutorial on REML is timely and relevant not only to the academic community but also to broader applied and industrial audiences.

Building on our prior experience organizing the REML workshop at SIGIR 2023 [4] and the tutorial at SIGIR-AP 2024[3] [10], where we witnessed substantial interest and engagement, we will incorporate insights from those events. This updated tutorial reflects ongoing conversations and highlights the evolving role of IR in this new era.

[1]https://trec.nist.gov/data/rag2024.html
[2]https://trec-mllm.github.io
[3]https://retrieval-enhanced-ml.github.io/sigir-ap2024-tutorial/

Time (min)	Topic	In Manuscript
20	Introduction	Sec. 1-2
25	Querying	Sec. 3
10	Searching	Sec. 4
30	Presentation & Consumption	Sec. 5
10	Q & A	
25	Storing	Sec. 6
20	Optimization	Sec. 7
15	Evaluation	Sec. 8
15	Future Direction & Conclusion	Sec. 9-10
10	Q & A	
3 hours		

Table 1: Tutorial schedule and corresponding manuscript sections [27].

4 Detailed Schedule

Table 1 presents the overall schedule along with the corresponding sections from the manuscript [27]. The detailed content for each section, along with representative papers, is provided in the bullet points below.

1. **Introduction** (presenter: Diaz)
   - Prehistory and definition of REML.
   - Motivations: generalization, knowledge grounding, scalability, freshness, attribution, and on-device learning [16, 22, 29, 56].
   - Applications beyond NLP (e.g., computer vision [7, 32], robot navigation [54], reinforcement learning [12, 15], drug discovery [31], protein structure prediction [20])
2. **Querying** (presenter: Salemi)
   - *How query spaces are represented and constructed by predictive models.*
   - Input transformation (compression [36], expansion [52], conversion [46]), and decomposition [34, 59].
   - Modeling when and where to query (routing) [25, 37, 48].
   - Unified equation for Querying.
3. **Searching** (presenter: Salemi)
   - *How queries and stored items are combined to construct retrieval results.*
   - Sparse [13], dense [24], and reranking [35] models.
   - Generative retrievers [49, 57].
   - Unified equation for Searching.
4. **Presentation & Consumption** (presenter: Drozdov)
   - *How retrieval results are represented and consumed by the predictive models.*
   - Presentation (transformation [14], composition [45], truncation [17]), and
   - Consumption (different granularities [19], algorithms [3], efficiency [9], attribution [14]) strategies.
   - Unified equation for Presentation and Consumption.
5. **Storing** (presenter: Kim)
   - *How retrievable items are represented and indexed.*

- Storage construction and management in coupled [16, 33] and decoupled [1, 22] storage.
- How agents can index past experiences [33, 51].
- Unified equation for Storing.

6. **Optimization** (presenter: Zamani)
   - *How retrieval models use feedback provided by predictive models to update their parameters, and how predictive models are optimized for performance.*
   - Conditional optimization of retrieval [18, 55] or predictive models [6].
   - Joint optimization of retrieval and predictive models [17, 58].
   - Unified equation for Optimization.
7. **Evaluation** (presenter: Diaz)
   - *How REML components are benchmarked.*
   - Extrinsic and intrinsic evaluations, and LLM-based evaluations [11, 40, 41].
   - Unified equation for Evaluation.
8. **Future Direction & Conclusion** (presenter: Diaz)
   - Future directions of each component of REML [26, 42–44] and across domains and communities.

5 Supporting Materials

This tutorial builds on the material and structure of an existing manuscript co-authored by the organizers [27]. In line with the approach used in our SIGIR-AP 2024 tutorial [10], attendees will receive slide materials throughout the session, including an annotated bibliography to aid understanding and follow-up study.

We will also provide detailed documentation and a comprehensive reading list. An example of the format and content can be found on the website for our previous tutorial[4], which will be updated with newly published papers and resources in preparation for SIGIR 2025.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant numbers 2402873 and 2402874, and in part by the Office of Naval Research contract number N000142412612. 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] Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*. https://openreview.net/forum?id=ZJZmKGM6UB

[2] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. ACL 2023 Tutorial: Retrieval-based Language Models and Applications. *ACL 2023* (2023).

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[4] Michael Bendersky, Danqi Chen, Fernando Diaz, and Hamed Zamani. 2023. SIGIR 2023 Workshop on Retrieval Enhanced Machine Learning (REML @ SIGIR 2023). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, 3468–3471. https://doi.org/10.1145/3539618.3591925

[5] Harrison Chase. 2022. *LangChain*. https://github.com/langchain-ai/langchain

[6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1870–1879. https://doi.org/10.18653/v1/P17-1171

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=XSEBx0iSjFQ

[8] Rajarshi Das, Patrick Lewis, Sewon Min, June Thai, and Manzil Zaheer (Eds.). 2022. *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*. Association for Computational Linguistics. https://aclanthology.org/2022.spanlp-1.0

[9] Michiel De Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Joshua Ainslie, Sumit Sanghai, Fei Sha, and William W. Cohen. 2023. Pre-Computed Memory or on-the-Fly Encoding? A Hybrid Approach to Retrieval Augmentation Makes the Most of Your Compute. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 290, 14 pages.

[10] Fernando Diaz, Andrew Drozdov, To Eun Kim, Alireza Salemi, and Hamed Zamani. 2024. Retrieval-Enhanced Machine Learning: Synthesis and Opportunities. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo, Japan) *(SIGIR-AP 2024)*. Association for Computing Machinery, New York, NY, USA, 299–302. https://doi.org/10.1145/3673791.3698439

[11] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 150–158. https://aclanthology.org/2024.eacl-demo.16

[12] Fernando Fernández and Manuela Veloso. 2006. Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. 720–727.

[13] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, 2288–2292. https://doi.org/10.1145/3404835.3463098

[14] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[15] Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adrià Puigdomènech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, Michal Valko, Simon Osindero, Timothy Lillicrap, Nicolas Heess, and Charles Blundell. 2022. Retrieval-Augmented Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 7740–7765.

[16] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 368, 10 pages.

[17] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1437–1447.

[18] Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *International Conference on Learning Representations*. https://openreview.net/forum?id=NTEz-6wysdb

[19] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7969–7992. https://doi.org/10.18653/v1/2023.emnlp-main.495

[20] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature* 596, 7873 (2021), 583–589.

[21] Evangelos Kanoulas, Panagiotis Eustratiadis, Yongkang Li, Yougang Lyu, Vaishali Pal, Gabrielle Poerwawinata, Jingfen Qiao, and Zihan Wang. 2025. Agent-centric Information Access. arXiv:2502.19298 [cs.IR] https://arxiv.org/abs/2502.19298

[22] 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

[23] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam,

[4] https://retrieval-enhanced-ml.github.io/sigir-ap2024-tutorial/

Heather Miller, et al. 2024. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

[24] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, 39–48. https://doi.org/10.1145/3397271.3401075

[25] Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, Xi Wang, and Guido Zuccon. 2023. Selecting which Dense Retriever to use for Zero-Shot Search. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Beijing, China) *(SIGIR-AP '23)*. Association for Computing Machinery, New York, NY, USA, 223–233. https://doi.org/10.1145/3624918.3625330

[26] To Eun Kim and Fernando Diaz. 2024. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. arXiv:2409.11598 [cs.IR] https://arxiv.org/abs/2409.11598

[27] To Eun Kim, Alireza Salemi, Andrew Drozdov, Fernando Diaz, and Hamed Zamani. 2024. Retrieval-Enhanced Machine Learning: Synthesis and Opportunities. arXiv:2407.12982 [cs.LG] https://arxiv.org/abs/2407.12982

[28] Julian Kiverstein. 2018. Extended cognition. *The Oxford handbook of 4E cognition* 1 (2018), 3–15.

[29] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[30] Jerry Liu. 2022. *LlamaIndex*. https://doi.org/10.5281/zenodo.1234

[31] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2024. Conversational drug editing using retrieval and domain feedback. In *The twelfth international conference on learning representations*.

[32] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. 2022. Retrieval Augmented Classification for Long-Tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6959–6969.

[33] Bodhisattwa Prasad Majumder, Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. 2024. CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization. https://openreview.net/forum?id=d5DGVHMdsC

[34] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 6097–6109. https://doi.org/10.18653/v1/P19-1613

[35] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Found. Trends Inf. Retr.* 13, 1 (dec 2018), 1–126. https://doi.org/10.1561/1500000061

[36] Ryan Musa, Xiaoyan Wang, Achille Fokoue, Nicholas Mattei, Maria Chang, Pavan Kapanipathi, Bassem Makni, Kartik Talamadupula, and Michael Witbrock. 2019. Answering Science Exam Questions Using Query Reformulation with Background Knowledge. In *Automated Knowledge Base Construction (AKBC)*. https://openreview.net/forum?id=HJxYZ-5paX

[37] Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2023. Knowledge-in-Context: Towards Knowledgeable Semi-Parametric Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=a2jNdqE2102

[38] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework. *arXiv preprint arXiv:2411.09607* (2024).

[39] Maithra Raghu, Urvashi Khandelwal, Chiyuan Zhang, Matei Zaharia, and Alexander Rush (Eds.). 2022. *Workshop on Knowledge Retrieval and Language Models*. ICML.

[40] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=J9oefdGUuM

[41] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *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, 338–354.

[42] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, 752–762. https://doi.org/10.1145/3626772.3657783

[43] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, 2395–2400. https://doi.org/10.1145/3626772.3657957

[44] Alireza Salemi and Hamed Zamani. 2024. Towards a Search Engine for Machines: Unified Ranking for Multiple Retrieval-Augmented Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, 741–751. https://doi.org/10.1145/3626772.3657733

[45] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=GN921JHCRw

[46] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=Yacmpz84TH

[47] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research* (2023).

[48] Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. 2024. Adapting to Non-Stationary Environments: Multi-Armed Bandit Enhanced Retrieval-Augmented Generation on Knowledge Graphs. *arXiv preprint arXiv:2412.07618* (2024).

[49] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 21831–21843. https://proceedings.neurips.cc/paper_files/paper/2022/file/892840a6123b5ec99ebaab8be1530fba-Paper-Conference.pdf

[50] Ellen M. Voorhees. 2003. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text REtrieval Conference*.

[51] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=ehfRiF0R3a

[52] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 9414–9423. https://doi.org/10.18653/v1/2023.emnlp-main.585

[53] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. 2024. Openhands: An open platform for ai software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*.

[54] Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. 2024. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313* (2024).

[55] Sohee Yang and Minjoon Seo. 2020. Is Retriever Merely an Approximator of Reader? *CoRR* abs/2010.10999 (2020). arXiv:2010.10999 https://arxiv.org/abs/2010.10999

[56] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[57] Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and Effective Generative Information Retrieval. In *Proceedings of the 2024 ACM Web Conference (WWW '24)*. Association for Computing Machinery.

[58] Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2022. Retgen: A joint framework for retrieval and grounded text generation modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11739–11747.

[59] Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2223–2235. https://doi.org/10.18653/v1/2022.emnlp-main.142