Tutorial Proposal: Retrieval-based Language Models and Applications

Akari Asai[†] **Sewon Min**[†] **Zexuan Zhong**[‡] **Danqi Chen**[‡]
[†] University of Washington [‡]Princeton University
{akari,sewon}@cs.washington.edu
{zzhong,danqic}@cs.princeton.edu

1 Description

Language models (LMs) such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) have shown impressive abilities in a range of natural language processing (NLP) tasks. However, relying solely on their parameters to encode a wealth of world knowledge requires a prohibitively large number of parameters and hence massive compute, and they often struggle to learn long-rail knowledge (Roberts et al., 2020; Kandpal et al., 2022; Mallen et al., 2022). Moreover, these parametric LMs are fundamentally incapable of adapting over time (De Cao et al., 2021; Lazaridou et al., 2021; Kasai et al., 2022), often hallucinate (Shuster et al., 2021), and may leak private data from the training corpus (Carlini et al., 2021). To overcome these limitations, there has been growing interest in retrieval-based LMs (Guu et al., 2020; Khandelwal et al., 2020; Borgeaud et al., 2022; Zhong et al., 2022; Izacard et al., 2022b; Min et al., 2022), which incorporate a non-parametric datastore (e.g., text chunks from an external corpus) with their parametric counterparts. Retrieval-based LMs can outperform LMs without retrieval by a large margin with much fewer parameters (Mallen et al., 2022), can update their knowledge by replacing their retrieval corpora (Izacard et al., 2022b), and provide citations for users to easily verify and evaluate the predictions (Menick et al., 2022; Bohnet et al., 2022).

Previously, retrieval and LMs have been studied mostly separately, and only recently researchers have integrated them and built systems in which retrieval and LMs interact more organically, and a number of retrieval-based LMs have been proposed due to growing interest. They differ in their neural architectures (e.g., the granularity of retrieval units, how to integrate retrieved information), learning algorithms, and different uses in downstream applications. In this tutorial, we aim to provide a comprehensive and coherent overview of recent advances in retrieval-based LMs. We will start by first providing preliminaries covering the foundations of LM (e.g., masked LMs, autoregressive LMs) and retrieval systems (e.g., nearest-neighbor search methods widely used in neural retrieval systems; Karpukhin et al. 2020). We will then focus on recent progress in *architectures*, *learning approaches*, and *applications* of retrieval-based LMs.

A taxonomy of architectures We introduce a taxonomy of architectures of retrieval-based LMs based on a variety of dimensions. Retrieval-based LMs can be categorized by the granularity of retrieved units stored in the datastore: either 1) a chunk of text (Borgeaud et al., 2022; Izacard et al., 2022b), or 2) a token (Khandelwal et al., 2020; Zhong et al., 2022; Min et al., 2022), or 3) an entity mention (Févry et al., 2020; de Jong et al., 2022). We also plan to cover techniques for refining data stores and improving similarity search (He et al., 2021; Alon et al., 2022). At the same time, retrieval-base LMs can be categorized based on how the retrieved information is integrated with the parametric encoder: 1) whether retrieved components are concatenated with the original input text (Lewis et al., 2020; Guu et al., 2020; Izacard et al., 2022b), 2) whether the retrieved components are latent and integrated into the intermediate layers of Transformers (de Jong et al., 2022; Févry et al., 2020; Borgeaud et al., 2022), or 3) distribution of tokens from the retrieved components and the LMs are interpolated (Khandelwal et al., 2020; Zhong et al., 2022; Yogatama et al., 2021).

Scalable learning algorithms Then, we discuss the *training approaches* of retrieval-based LMs. Since a retrieval datastore is typically very large, how to train retrieval-based LMs effectively and efficiently remains challenging. We first discuss pipelined approaches that train retrieval components and LMs separately, either through large-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 6: Tutorial Abstracts, pages 41–46

scale pre-training (Izacard et al., 2022a) or multi-task instruction tuning (Asai et al., 2022). Several other works train retrieval-based LMs with a fixed retrieval module (Borgeaud et al., 2022; Yogatama et al., 2021). We then discuss joint training under reasonable resource requirements: either through in-batch approximations to a full datastore, or updating the datastore with updated parameters asynchronously. The former uses fractions of the full corpus that are carefully designed during joint training (Zhong et al., 2022; de Jong et al., 2022; Min et al., 2022). The latter, on the other hand, aims to use full corpus during training with asynchronous index update for every certain time steps (Izacard et al., 2022b; Guu et al., 2020).

Adaption to downstream tasks After discussing the basic building blocks of retrieval-based LMs, we show how retrieval-based LMs are adapted to downstream applications. We first briefly summarize the two approaches to adapt a model to a new task: zero-shot or few-shot prompting without any parameter updates (Shi et al., 2022; Wang et al., 2022), and fine-tuning on target task data (Lewis et al., 2020). We then discuss methods designed to build more powerful retrieval-based LMs for certain downstream tasks, such as dialogue (Shuster et al., 2021), semantic parsing (Pasupat et al., 2021), and machine translation (Khandelwal et al., 2021; Zheng et al., 2021).

Up to this point, our tutorial has mainly focused on retrieving and integrating English plain text. At this end, we will cover recent extensions of retrieval-based LMs beyond English text, including multilingual (Asai et al., 2021), multimodal (Chen et al., 2022; Yasunaga et al., 2022) and code (Parvez et al., 2021) retrieval. These works often extend dense retrieval models to enable retrieval between heterogeneous input spaces (e.g., cross-lingual, cross-modal) and have shown that referring retrieved knowledge leads to knowledge-intensive generation.

Finally, we will use an exercise to showcase the effectiveness of retrieval-based LMs. We conclude our tutorial by discussing several important questions and future directions, including (1) how we can further improve the scalability of retrieval-based LMs without sacrificing performance, (2) when retrieval-based LMs are particularly useful in the era of rapidly evolving LMs, and (3) what is necessary to enable applications of retrieval-based LMs for more diverse domains.

2 Tutorial Outline

1. Introduction (15 minutes)

- An overview of the tutorial
- Why retrieval-based LMs?

2. Preliminaries (15 minutes)

- Language models: Auto-regressive LMs vs. masked LMs
- Dense retrieval methods
- Approximate nearest neighbor search

3. Retrieval-based LMs: A taxonomy of architectures (40 minutes)

- Granularity of datastore: tokens, entity mentions, and chunks of text
- How retrieved information is integrated: incorporation in the input layer, intermediate layers, and the output layer

4. Retrieval-based LMs: Scalable learning algorithms (40 minutes)

- Pipelined training
- Training with In-batch approximations
- Joint training of retrieval and LMs with asynchronous updates of corpus

5. Retrieval-based LMs: Downstream adaptations (40 minutes)

- Adaptation methods: zero-shot/few-shot prompting and fine-tuning on downstream tasks
- Downstream applications and task-specific modifications (e.g., dialogue, semantic parsing)

6. Extensions beyond English text (10 minutes)

- Multilingual retrieval-based LMs
- Multimodal retrieval-based LMs
- Code generation

7. Demostration: An exercise to show retrieval-augmented LMs (10 minutes)

8. Conclusions and future directions (10 minutes)

3 Tutorial Information

Type of the tutorial Cutting-edge.

Length This is a 3-hour tutorial.

Target audience The tutorial will be accessible to anyone who has a basic knowledge of machine learning and natural language processing. We think the topic will be of interest to both NLP researchers/students in academia and NLP practitioners in the industry.

Breadth We estimate that 20% of the work covered in this tutorial will be by the presenters and the remaining 80% by others. The papers we will cover are from both academia and industry.

Diversity considerations. The speakers are from two academic institutions with an affiliation with an industry research group, including both a professor and Ph.D. students. Three out of four speakers are female. The methods covered by our tutorials can scale up to various languages or domains, and we also briefly cover several papers focusing on multilingual and expert-domain extensions of the core frameworks. We will reach out to academic communities such as WiNLP[1] and Masakhane[2] to encourage them to attend our tutorial for participation of diverse audiences. Since retrieval-based LMs are alternatives to LMs with a significantly large number of parameters, we expect this tutorial to be especially useful to researchers with modest resources who do no have access to very large models.

An estimate of the audience size Given that language models are now used in a range of NLP tasks and retrieval-based approaches have been applied to diverse domains, we estimate that the number of audiences will be around 150+.

Venues. We prefer ACL due to the growing interest in the area and the travel constraints of some of the speakers. EMNLP is our second preferred choice, and we currently do not consider EACL.

Technical equipment. We would like to have Internet access to show online demos.

Open access We plan to make all teaching material available online and agree to allow the publication of slides and video recordings in the ACL anthology.

[1] http://www.winlp.org/

[2] https://www.masakhane.io/

Ethical considerations Retrieval-based LMs are often more powerful and parameter-efficient than LMs, and do not require full re-training to update world knowledge, which makes it more energy-efficient and can reduce carbon footprints. Prior work also shows that referring to external world knowledge can reduce harmful biases and hallucinations, although retrieval-based LMs can still be plausible sounding but incorrect or non-sensical outputs. We note that, as retrieval-based LMs may retrieve raw data from a corpus, which can leak privacy-sensitive information, especially when they are built on top of a private corpus. We acknowledge this to caution those who manage to apply retrieval-based LMs to privacy-sensitive domains.

Pedagogical material We plan to do some short hands-on exercises to let the audience try different retrieval-based LMs with few-shot prompting using Colab.

Past tutorials.

- ACL 2020 tutorial on Open-domain QA (Chen and Yih, 2020): This tutorial provides comprehensive reviews of open-domain question answering, some of which consist of a retriever and a generative model, while we focus on the recent progress of architectures and learning algorithms of retrieval-based LMs for diverse NLP tasks, not limiting its focus to open-domain QA. Most of the papers will be discussed in this tutorial have been published since the Open-domain QA tutorial three years ago. Moreover, one of the instructors, Danqi was an instructor of this ACL 2020 tutorial.

- SIGIR 2022 tutorial on Recent Advances in Retrieval-Augmented Text Generation (Cai et al., 2022): This tutorial focuses mainly on recent retrieval-augmented text generation approaches with a focus on two applications: dialogue and machine translation. Our tutorial puts more emphasis on the architecture and learning methods of retrieval-based LMs that can be applicable to diverse NLP tasks.

4 Presenters

Akari Asai Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Prof. Hannaneh Hajishirzi. Her research lies

in natural language processing and machine learning. Her recent research focuses on question answering, retrieval-based LMs, multilingual NLP, and entity-aware representations. She received the IBM Fellowship in 2022. She is a lead organizer of the Workshop on Multilingual Information Access (NAACL 2022) and serves as an area chair in question answering at EACL 2023.

Sewon Min Sewon Min is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, and a visiting researcher at Meta AI. Her research spans question answering, representation and retrieval of factoid knowledge, and language modeling. She was a co-instructor and a co-organizer of multiple tutorials and workshops at ACL, NAACL-HLT, EMNLP, NeurIPS and AKBC, including a tutorial on Few-Shot NLP with Pretrained Language Models (ACL 2022), a tutorial on NLP for Long Sequences (NAACL-HLT 2021), and the Workshop on Semiparametric Methods in NLP (ACL 2022).

Zexuan Zhong Zexuan Zhong is a Ph.D. student in the Department of Computer Science at Princeton University, advised by Prof. Danqi Chen. His research interests lie in natural language processing and machine learning. His recent research focuses on retrieval-based LMs, generalization of retrieval models, and efficient models in NLP. He received a J.P. Morgan PhD Fellowship in 2022.

Danqi Chen Danqi Chen is an Assistant Professor of Computer Science at Princeton University and co-leads the Princeton NLP Group. Her recent research focuses on training, adapting, and understanding large LMs, and developing scalable and generalizable NLP systems for question answering, information extraction, and conversational agents. 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 industry faculty awards. Danqi served as the program chair for AKBC 2021 and (senior) area chairs for many *ACL conferences. She taught a tutorial on "Open-domain Question Answering" at ACL 2020.

5 Reading List

- Unsupervised Dense Information Retrieval with Contrastive Learning (Izacard et al., 2022a)
- Task-aware Retrieval with Instructions (Asai et al., 2022)
- Atlas: Few-shot Learning with Retrieval Augmented Language Models (Izacard et al., 2022b)
- Improving language models by retrieving from trillions of tokens (Borgeaud et al., 2022)
- Mention Memory: incorporating textual knowledge into Transformers through entity mention attention (de Jong et al., 2022)
- Generalization through Memorization: Nearest Neighbor Language Models (Khandelwal et al., 2020)
- Nonparametric Masked Language Model (Min et al., 2022)
- Training Language Models with Memory Augmentation (Zhong et al., 2022)
- kNN-Prompt: Nearest Neighbor Zero-Shot Inference (Shi et al., 2022)
- Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval (Alon et al., 2022)

References

Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning (ICML)*, Baltimore, USA.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Michiel de Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William W. Cohen. 2022. Mention memory: incorporating textual knowledge into transformers through entity mention attention. In *International Conference on Learning Representations*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

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*.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language

models. *Advances in Neural Information Processing Systems*.

Patrick 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*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Nonparametric masked language modeling. *arXiv preprint arXiv:2212.01349*.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Panupong Pasupat, Yuan Zhang, and Kelvin Guu. 2021. Controllable semantic parsing via retrieval augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Zhenhailong Wang, Xiaoman Pan, Dian Yu, Dong Yu, Jianshu Chen, and Heng Ji. 2022. Zemi: Learning zero-shot semi-parametric language models from multiple tasks. *arXiv preprint arXiv:2210.00185*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.