Section 7

Challenges & Opportunities

Tutorial summary

- **Definition & preliminaries**

- **Retrieval-based LMs: Architecture**
  - What to retrieve: tokens, text blocks, entity mentions
  - How to use retrieval: input layer, intermediate layers, output layers
  - When to retrieve: once, every n tokens, every token

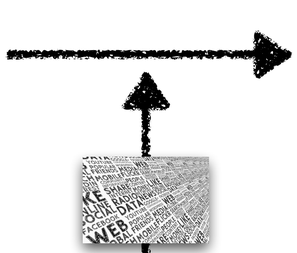
- **Retrieval-based LMs: Training**
  - Independent training
  - Sequential training
  - Joint training with asynchronous index update
  - Joint training with in-batch approximation

Tutorial summary (cont'd)

- **Applications**
  - What tasks: knowledge-intensive NLP, code generation, classification etc
  - How to adapt: retrieval-based prompting, fine-tuning, reinforcement learning
  - When to use: long-tail, knowledge update, verifiability, parameter-efficiency, privacy, out of domain adaptations
- **Multilingual & multimodal extensions**
  - Multilingual: retrieve and generate cross-lingually to overcome scarcity of datastore
  - Multimodal: multimodal extensions of retrieval-based LMs for new modality

Challenge: Scaling retrieval-based LMs

A small LM + a large datastore ≈ a large parametric LM?



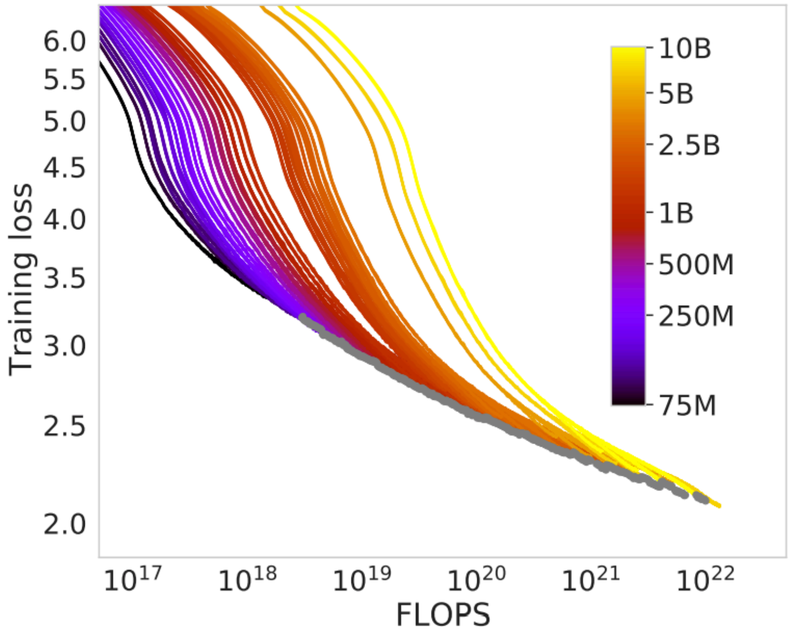
vs.

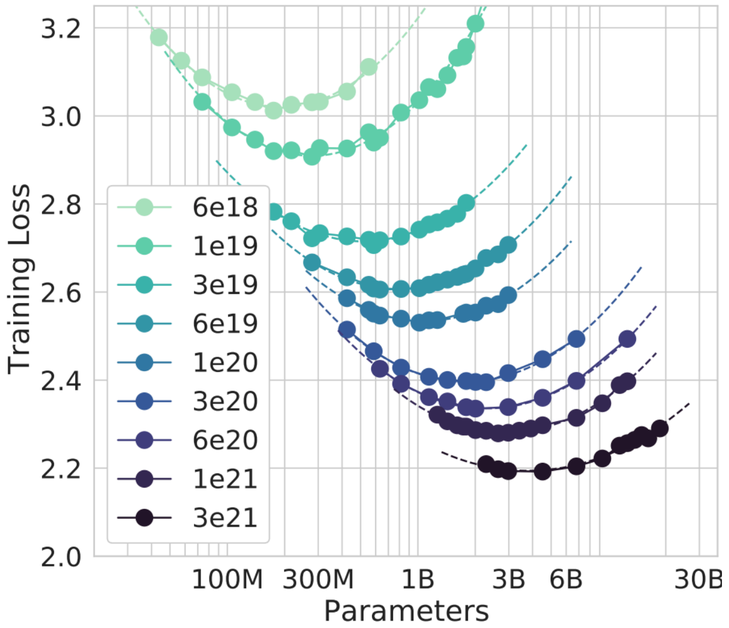
LM

	LM	Datastore
	# of parameters	# of tokens
kNN-LM (Khandelwal et al., 2020)	250M	≤ 3B
NPM (Min et al., 2023)	350M	1B
Atlas (Izacard et al., 2022)	11B	~30B
RETRO (Borgeaud et al., 2021)	7B	2T
REPLUG (Shi et al., 2023)	≤ 175B	~5B

Challenge: Scaling retrieval-based LMs

Scaling law?





Scaling law for parametric LMs (Kalpan et al., 2020; Hoffman et al., 2022)

Loss as a function of:

- Training tokens
- Model sizes
- Training FLOPs

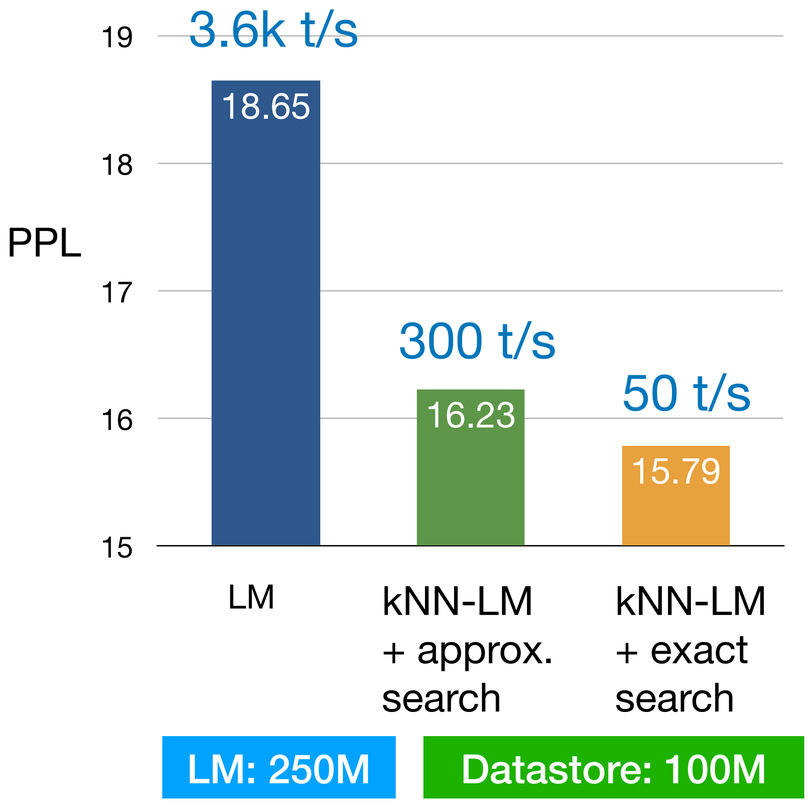
+ Datastore sizes?

Dependent on architecture and training methods!

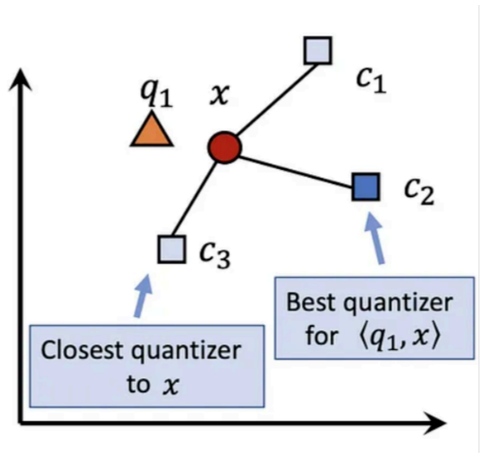
Challenge: Scaling retrieval-based LMs

Efficiency of similarity search

Measured on NVIDIA RTX 3090 GPU (Zhong et al., 2022) with a FAISS indexer (Johnson et al., 2021) with 32 CPUs



- >12 times slower even with **approximate** nearest neighbor search **implemented on GPUs**
- Efficient similarity search on GPUs/TPUs becomes the bottleneck of scaling retrieval-based LMs



Active research area!

Guo et al. 2020. "Accelerating Large-Scale Inference with Anisotropic Vector Quantization"

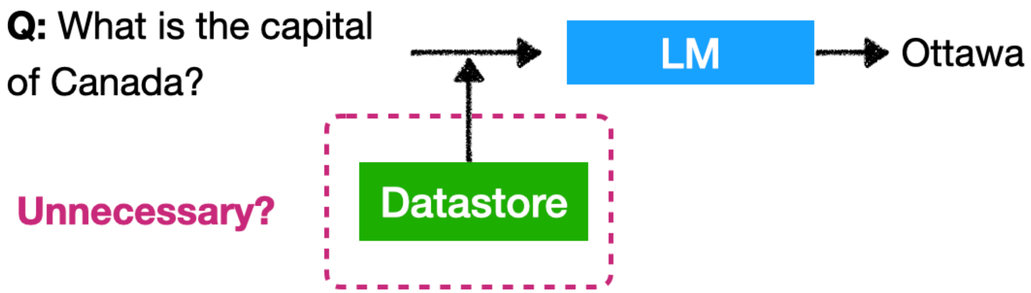
Challenge: Scaling retrieval-based LMs

Efficiency of similarity search

Efficient nearest neighbor search:

- Better loss for vector quantization (Guo et al., 2020)
- Optimization over different accelerators (Chern et al., 2022)
- Adaptive representations (Rege et al., 2023)

Middle ground: adaptively retrieve only when necessary (He et al., 2021; Mallen et al., 2023)



Space efficiency?

It also requires a large disk space for storage if the datastore becomes too large - storing **vectors** or **raw text**?

Challenge: Retrieval-based LMs for applications

Open-ended text generation?

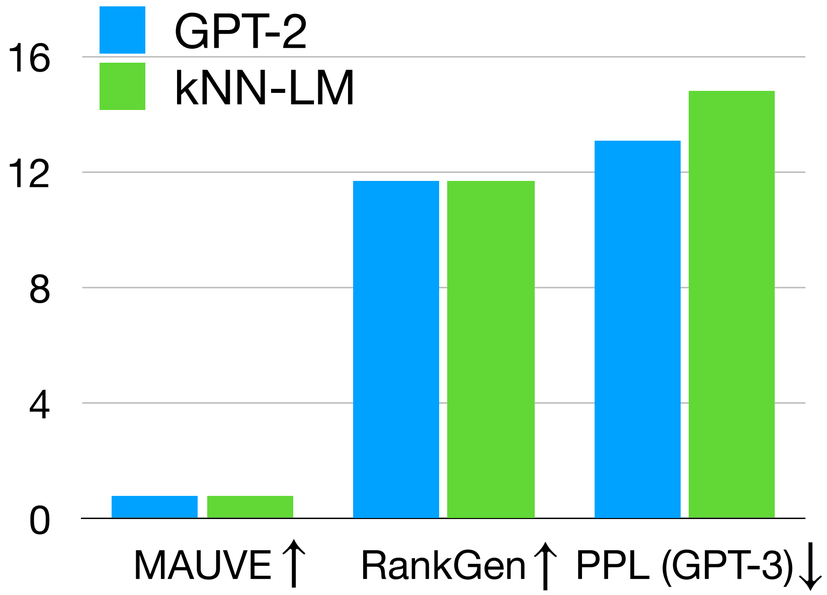
DA

Hey I am going to ACL in Toronto next week. Can you tell me how I should explore the conference?

Certainly! Attending a conference like ACL (Association for Computational Linguistics) can be a fantastic opportunity to learn, network, and engage with the research community. Here are some tips to help you make the most of your time at the conference:

1. Review the schedule: Take some time to familiarize yourself with the conference program in advance. Note the sessions, workshops, tutorials, and keynotes that interest you the most. Identify any must-attend sessions and plan your schedule accordingly.
2. Create a personalized agenda: Once you have a clear understanding of the sessions you want to attend, create a personalized agenda. Note down the session names, locations, and times so you can easily navigate the conference venue and stay organized.
3. Engage in networking: Conferences provide an excellent opportunity to connect with fellow researchers and professionals in your field. Be proactive and introduce yourself to others during coffee breaks, poster sessions, or social events. Don't be afraid to strike up conversations, exchange ideas, and ask questions. Networking can lead to valuable collaborations and partnerships.
4. Attend keynote speeches: Keynote speeches are typically delivered by renowned experts in

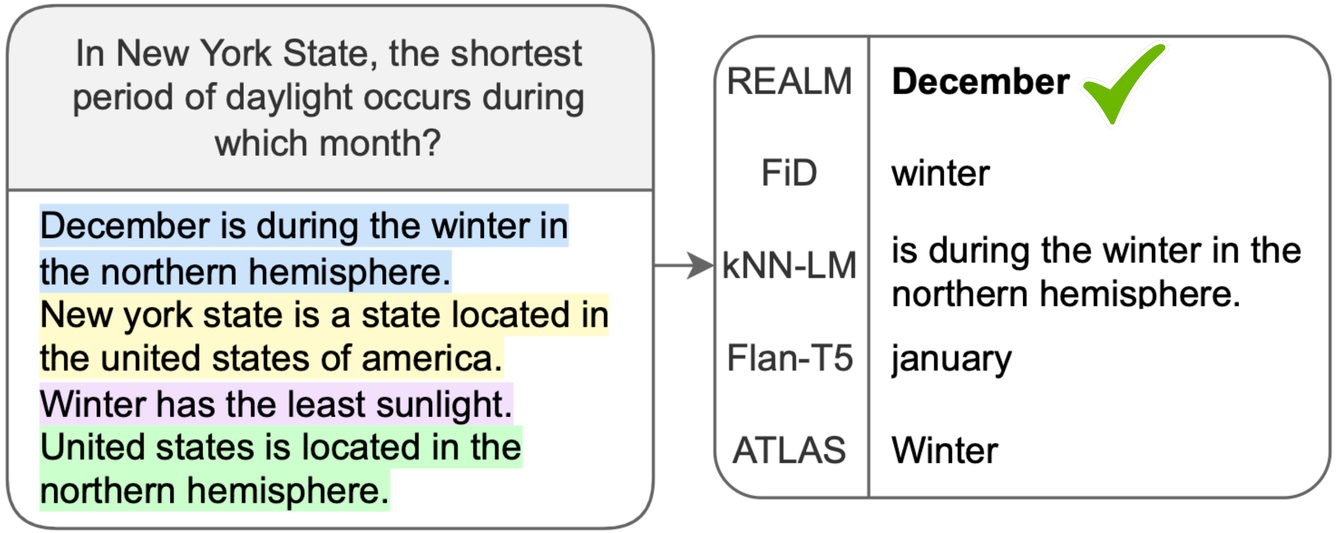
Exposure bias? Better decoding algorithms?



Wang et al. 2023. “kNN-LM Does Not Improve Open-ended Text Generation”

Challenge: Retrieval-based LMs for applications

Complex reasoning tasks?



Retrieval-based LMs struggle with **multi-step entailments** or **logical reasoning**:

- Retrieving based on similarity of queries is insufficient 😢

- LMs also struggle with synthesizing information from multiple statements 😢

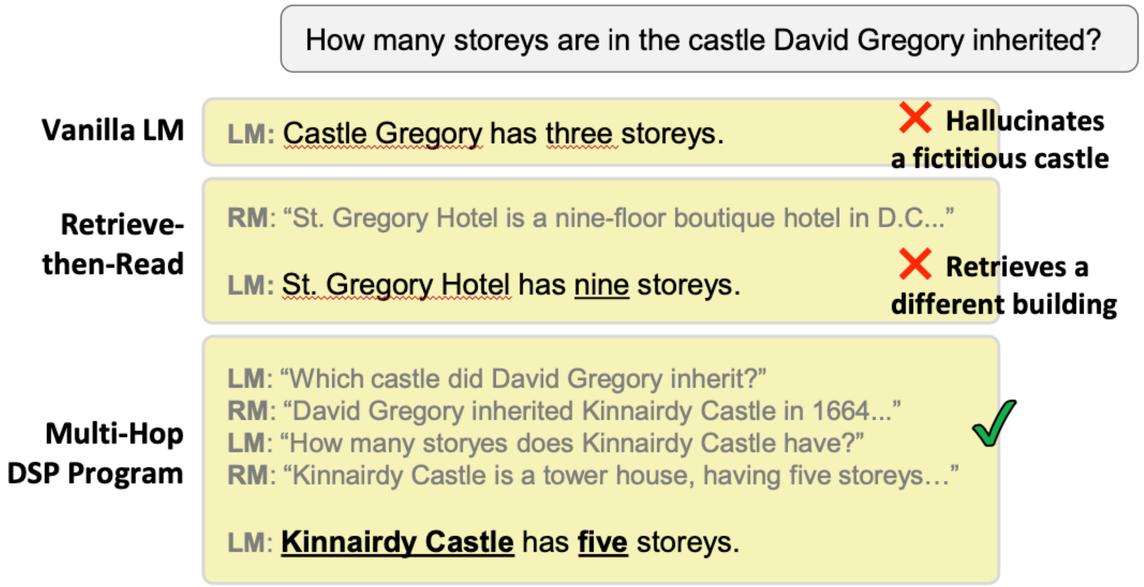
BehnamGhader et al. 2022. “Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model.”

Challenge: Retrieval-based LMs for applications

Complex reasoning tasks?

Potential solutions?

- **Iterative retrieval** & query reformulation
  - FLARE (Jiang et al., 2023), Self-Ask (Press et al., 2022), ReAct (Yao et al., 2023)
- Decompose the task into a **multi-hop program** for complex reasoning
  - DSP: Demonstrate-Search-Predict



LM: Language Model, RM: Retrieval Model (both frozen)

Khattab et al. 2022. "Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP"

Open questions

- What is the best **architecture** & **training method** for retrieval-based LMs in practice?
- We still don’t know yet how to best **scale up** these models - Scaling law?
- We may need to explore alternative **decoding or adaptation methods** in downstream tasks (e.g., open-ended text generation, complex reasoning)!

Q & A

Thank you for joining us today!

All the materials are at
https://acl2023-retrieval-lm.github.io/

Section 1: References (1)

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Omar Khattab, Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR 2020.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, Arnold Overwijk. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. ICLR 2021.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. TMLR 2022.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, Hannaneh Hajishirzi. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, John Schulman. WebGPT: Browser-assisted question-answering with human feedback. arXiv 2021.

Section 1: References (2)

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, Nat McAleese. Teaching Language Models to Support Answers with Verified Quotes. arXiv 2022.

Tianyu Gao, Howard Yen, Jiatong Yu, Danqi Chen. Enabling Large Language Models to Generate Text with Citations. arXiv 2023.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel. Extracting Training Data from Large Language Models. UNISEX Security 2021.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George 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 W. Rae, Erich Elsen, Laurent Sifre. Improving language models by retrieving from trillions of tokens. arXiv 2021.

Section 3: References (1)

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. ICML 2020.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, Yoav Shoham. In-Context Retrieval-Augmented Language Models. arXiv 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv 2023.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George 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 W. Rae, Erich Elsen, Laurent Sifre. Improving language models by retrieving from trillions of tokens. arXiv 2021.

Section 3: References (2)

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis. Nearest Neighbor Machine Translation. ICLR 2021.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, Graham Neubig. Active Retrieval Augmented Generation. arXiv 2023.

Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick. Efficient Nearest Neighbor Language Models. EMNLP 2021.

Andrew Drozdov, Shufan Wang, Razieh Rahimi, Andrew McCallum, Hamed Zamani, Mohit Iyyer. You can't pick your neighbors, or can you? When and how to rely on retrieval in the kNN-LM. EMNLP 2022 Findings.

Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, Graham Neubig. Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval. ICML 2022.

Section 3: References (3)

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, Tom Kwiatkowski. Entities as Experts: Sparse Memory Access with Entity Supervision. EMNLP 2020.

Michiel de Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, William Cohen. Mention Memory: incorporating textual knowledge into Transformers through entity mention attention. ICLR 2022.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, Christian Szegedy. Memorizing Transformers. ICLR 2022.

Amanda Bertsch, Uri Alon, Graham Neubig, Matthew R. Gormley. Unlimiformer: Long-Range Transformers with Unlimited Length Input. arXiv 2023.

Ohad Rubin, Jonathan Berant. Long-range Language Modeling with Self-retrieval. arXiv 2023.

Section 4: References (1)

Juan Ramos. Using TF-IDF to Determine Word Relevance in Document Queries. The first instructional conference on machine learning 2003.

Robertson and Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval 2009.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. TMLR 2022.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv 2023.

Section 4: References (2)

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, Yoav Shoham. In-Context Retrieval-Augmented Language Models. arXiv 2023.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, Edouard Grave. Atlas: Few-shot Learning with Retrieval Augmented Language Models. arXiv 2022.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, Edouard Grave. Atlas: Few-shot Learning with Retrieval Augmented Language Models. arXiv 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George 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 W. Rae, Erich Elsen, Laurent Sifre. Improving language models by retrieving from trillions of tokens. arXiv 2021.

Section 4: References (3)

Zexuan Zhong, Tao Lei, Danqi Chen. Training Language Models with Memory Augmentation. EMNLP 2022.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, Luke Zettlemoyer. Nonparametric Masked Language Modeling. ACL 2023 Findings.

Ohad Rubin, Jonathan Berant. Long-range Language Modeling with Self-retrieval. arXiv 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020.

Gautier Izacard, Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv 2020.

Section 5: References (1)

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-domain Question Answering. EMNLP 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. ICML 2020.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, Edouard Grave. Atlas: Few-shot Learning with Retrieval Augmented Language Models. Arxiv 2022.

Akari Asai, Matt Gardner, Hannaneh Hajishirzi. Evidentiality-guided Generation for Knowledge-Intensive NLP Tasks. NAACL 2022.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, Jason Weston. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. Arxiv 2022.

Section 5: References (2)

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis. Nearest Neighbor Machine Translation. ICLR 2021.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, Graham Neubig. Active Retrieval Augmented Generation. arXiv 2023.

Zexuan Zhong, Tao Lei, Danqi Chen. Training Language Models with Memory Augmentation. EMNLP 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George 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 W. Rae, Erich Elsen, Laurent Sifre. . Improving language models by retrieving from trillions of tokens. Arxiv 2022.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, Meng Jiang. Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022.

Weijia Shi, Julian Michael, Suchin Gururangan, Luke Zettlemoyer kNN-Prompt: Nearest Neighbor Zero-Shot Inference. EMNLP 2022.

Section 5: References (3)

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, Nat McAleese. GopherCite: Teaching language models to support answers with verified quotes. Arxiv 2022.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih. REPLUG: Retrieval-Augmented Black-Box Language Models. Arxiv 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, Hannaneh Hajishirzi. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhiruo Wang, Zhengbao Jiang, Graham Neubig. DocPrompting: Generating Code by Retrieving the Docs. ICLR 2023.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, Danqi Chen. Privacy Implications of Retrieval-Based Language Models. Arxiv 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, Hannaneh Hajishirzi FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. Arxiv 2023.

Section 6: References (1)

Akari Asai, Xinyan Yu, Jungo Kasai, Hannaneh Hajishirzi. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. NeurIPS 2021.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong. CL-ReLKT: Cross-lingual Language Knowledge Transfer for Multilingual Retrieval Question Answering. NAACL Findings 2022.

Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, Alessandro Moschitti. Cross-Lingual GenQA: A Language-Agnostic Generative Question Answering Approach for Open-Domain Question Answering. AACL 2022.

Kung-Hsiang Huang, ChengXiang Zhai, Heng Ji. CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval. COLING 2022.

Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, Michael R. Lyu. Retrieval-Augmented Multilingual Keyphrase Generation with Retriever-Generator Iterative Training. NAACL Findings 2022.

Section 6: References (2)

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, Alireza Fathi. REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. CVPR 2023.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, William W. Cohen. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. EMNLP 2022.

Shir Gur, Natalia Neverova, Chris Stauffer, Ser-Nam Lim, Douwe Kiela, Austin Reiter. Cross-Modal Retrieval Augmentation for Multi-Modal Classification. EMNLP Findings 2021.

Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, Anima Anandkumar. Retrieval-based Controllable Molecule Generation. ICLR 2023.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, Ziwei Liu. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. Arxiv 2023.

Section 7: References (1)

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei. Scaling Laws for Neural Language Models. arXiv 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, Laurent Sifre. Training Compute-Optimal Large Language Models. NeurIPS 2022.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, Sanjiv Kumar. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. ICML 2020.

Felix Chern, Blake Hechtman, Andy Davis, Ruiqi Guo, David Majnemer, Sanjiv Kumar. TPU-KNN: K Nearest Neighbor Search at Peak FLOP/s. NeurIPS 2022.

Aniket Rege, Aditya Kusupati, Sharan Ranjit S, Alan Fan, Qingqing Cao, Sham Kakade, Prateek Jain, Ali Farhadi. AdANNS: A Framework for Adaptive Semantic Search. arXiv 2023.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, Graham Neubig. Active Retrieval Augmented Generation. arXiv 2023.

Section 7: References (2)

Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, Mohit Iyyer. KNN-LM Does Not Improve Open-ended Text Generation. arXiv 2023.

Parishad BehnamGhader, Santiago Miret, Siva Reddy. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. arXiv 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, Mike Lewis. Measuring and Narrowing the Compositionality Gap in Language Models. arXiv 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, Matei Zaharia. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. arXiv 2022.