



[https://tinyurl.com/  
retrieval-lm-tutorial](https://tinyurl.com/retrieval-lm-tutorial)  
Join sli.do for Q & A!



ACL 2023 Tutorial:

# Retrieval-based Language Models and Applications

Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen

<https://acl2023-retrieval-lm.github.io/>

July 9, 2023

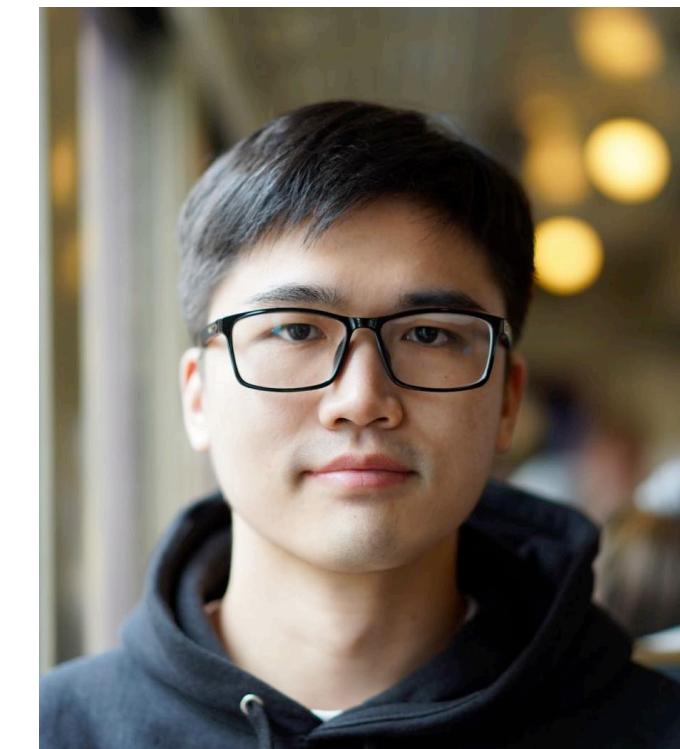
# About presenters



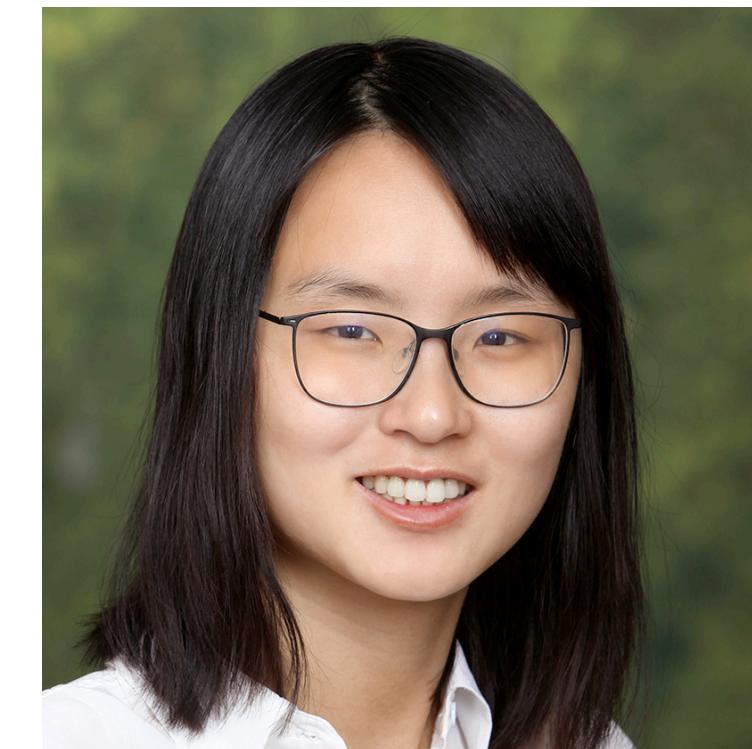
Akari Asai  
PhD student  
@UW



Sewon Min  
PhD student  
@UW



Zexuan Zhong  
PhD student  
@Princeton



Danqi Chen  
Faculty  
@Princeton

# Participation and Q & A

- All tutorial slides and reading list are available at:

<https://acl2023-retrieval-lm.github.io/>

- Throughout the tutorial, we will provide Q & A on sli.do (also on the website):

<https://tinyurl.com/retrieval-lm-tutorial>



The screenshot shows the sli.do interface for the "ACL 2023 Tutorial: Retrieval-based Methods". The top navigation bar includes a menu icon, the title, and a "Q&A" button. Below the title, event details are listed: "ACL 2023 Tutorial: Retrieval-based Methods", "Jul 6–10, 2023", and "#1430 562". On the left, there are three buttons: "Live interaction", "Switch event", and "Dark mode" with a toggle switch. On the right, there is a large input field with a user icon and the placeholder text "Type your question".

Ask & upvote questions anytime!  
(+10 minutes Q & A in 1st and 2nd half)

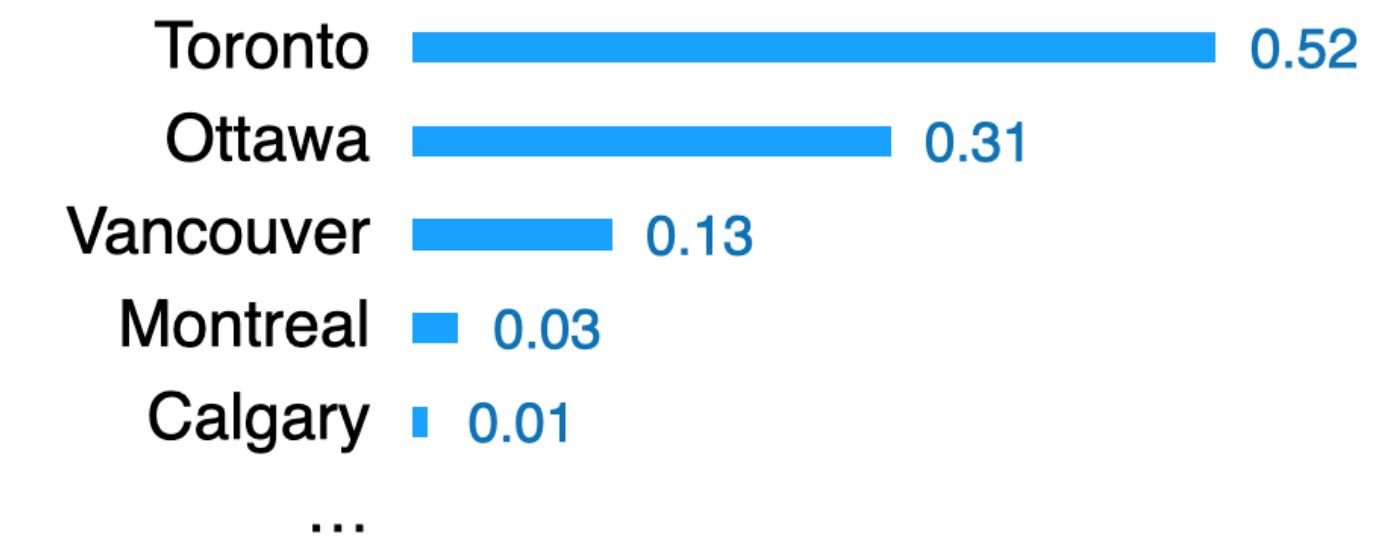
# Retrieval-based language models (LMs)

Retrieval-based LMs = Retrieval + LMs

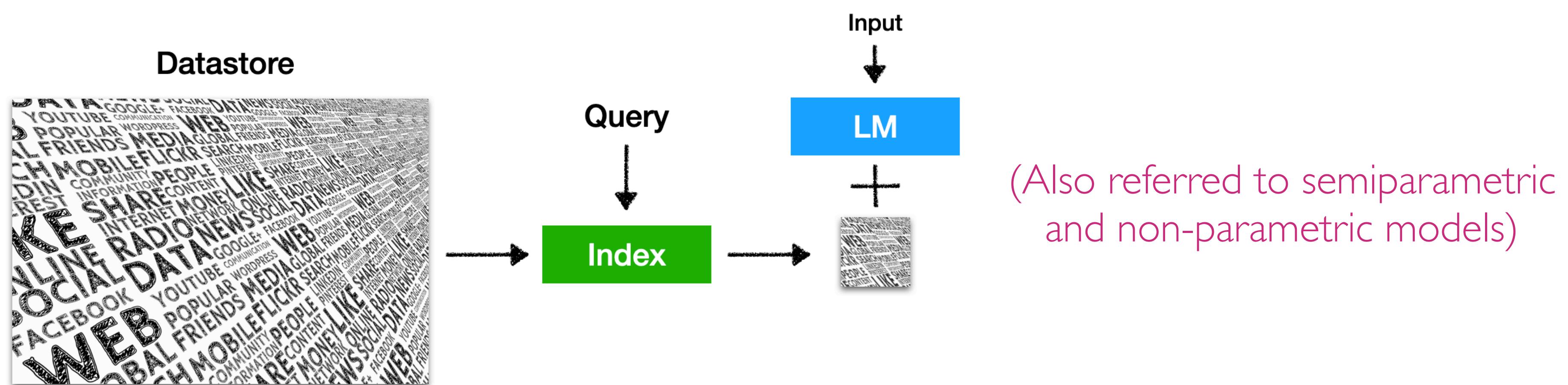
- It is a **language model**  $P(x_n | x_1, x_2, \dots, x_{n-1})$

The capital city of Ontario is \_\_\_\_\_

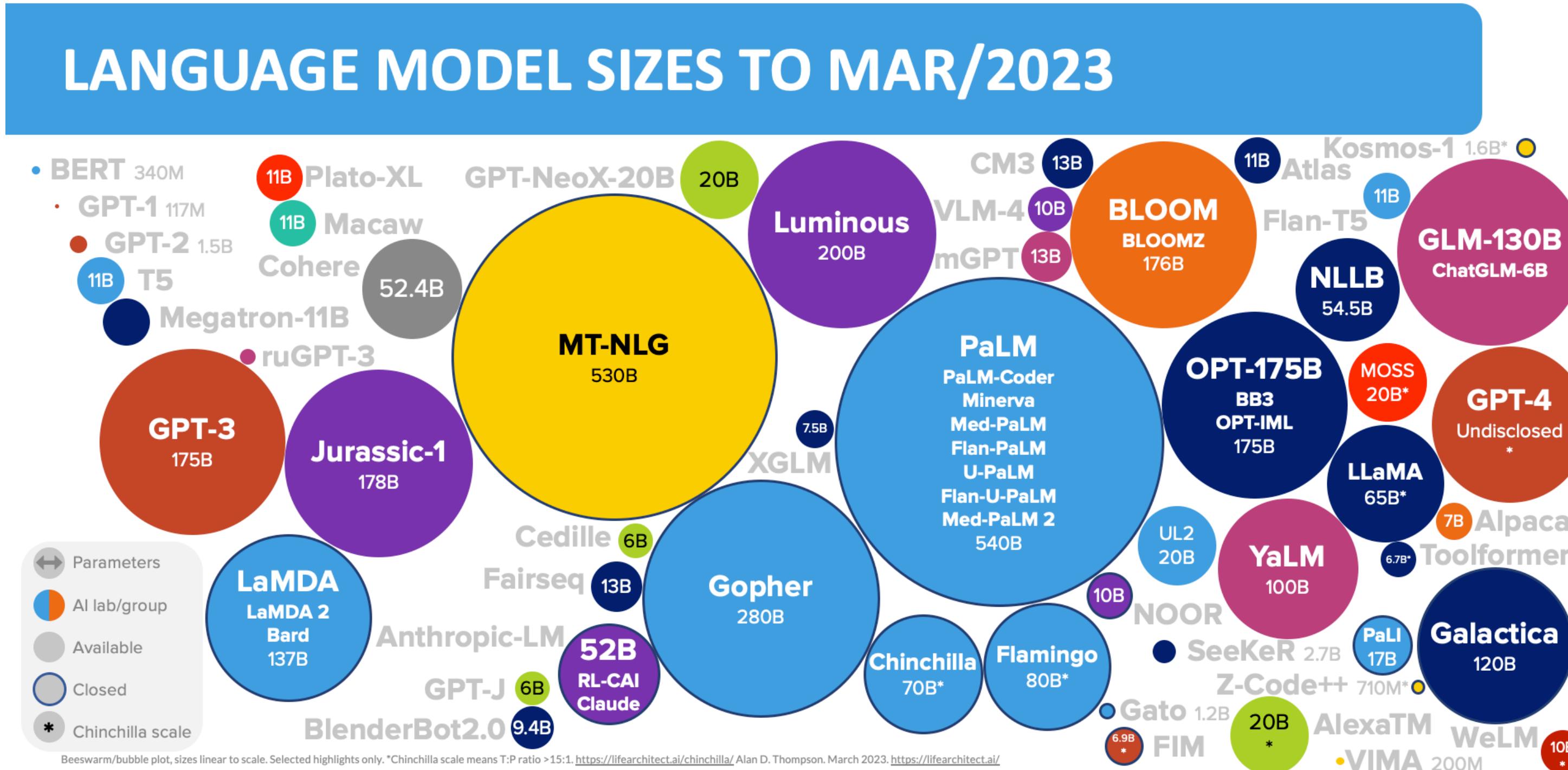
(can be broadly extended to masked language models or encoder-decoder models)



- It retrieves from an **external datastore** (at least during inference time)



# The age of large language models (LLMs)



- Transformers-based, **fully parametric**
  - Trained on next-token prediction tasks (+ RLHF; not the focus today)
  - **Model size  $\uparrow$ , data size  $\uparrow$**

Image: <https://lifearchitect.ai/models/>

# Retrieval for knowledge-intensive NLP tasks

**Representative tasks:** open-domain QA, fact checking, entity linking, ..

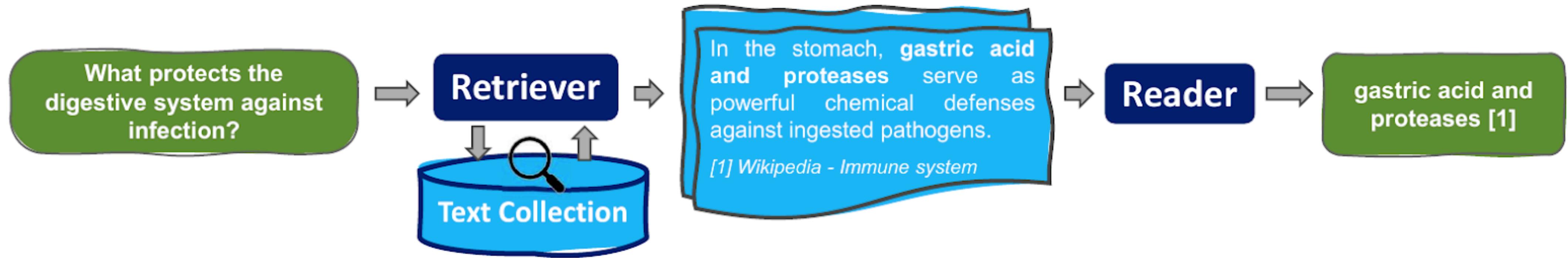


Image: <http://ai.stanford.edu/blog/retrieval-based-NLP/>

Drives a lot of research on better algorithms for **dense retrieval**, e.g., **DPR** (Karpukhin et al., 2020), **ColBERT** (Khattab and Zaharia, 2020), **ANCE** (Xiong et al., 2021), **Contriever** (Izacard et al., 2022), ...

Why retrieval → LMs?

# Why retrieval-based LMs?

LLMs can't memorize all (long-tail) knowledge in their parameters

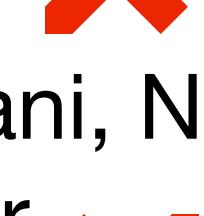


List 5 important papers authored by Geoffrey Hinton

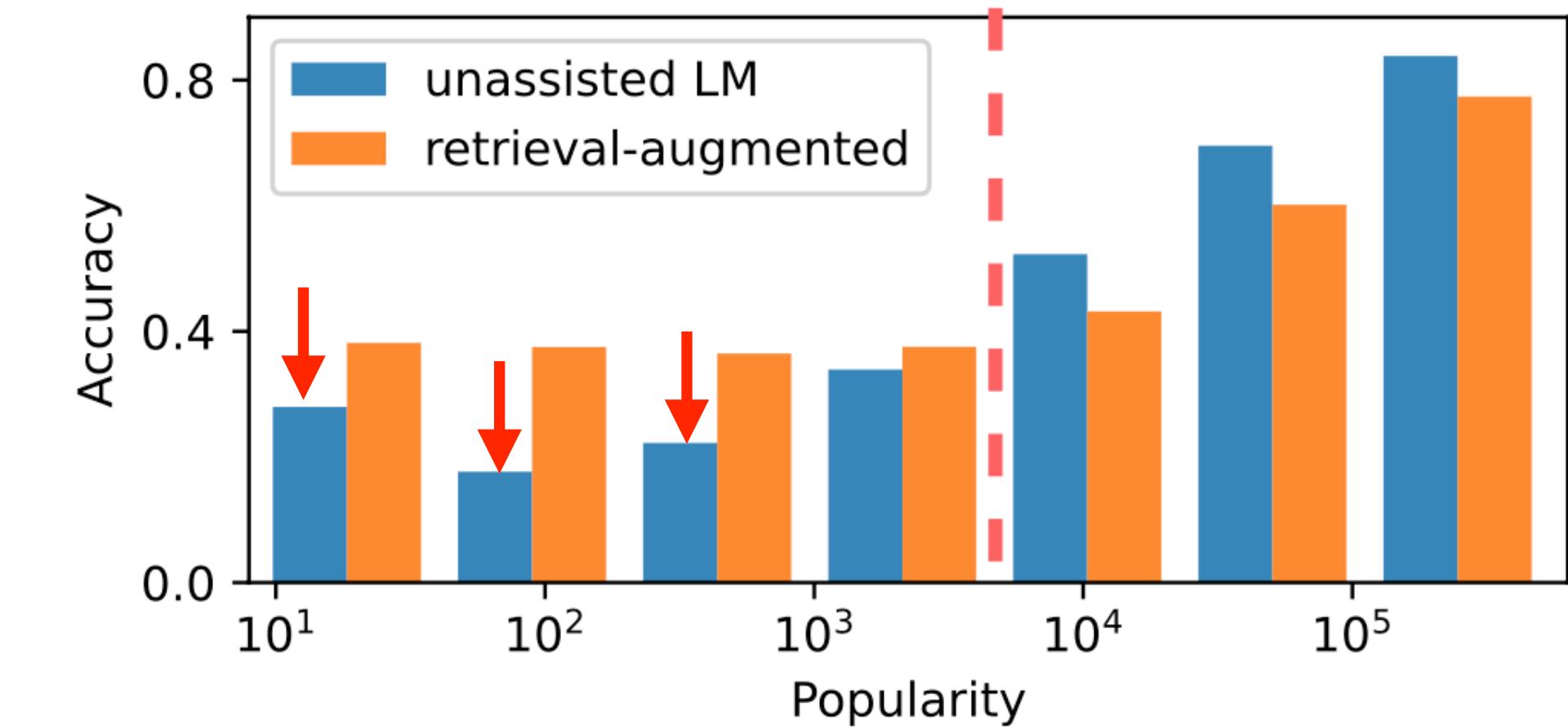


ChatGPT

Geoffrey Hinton is a renowned computer scientist ...  
Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, ... 
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, ... 
3. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,... 
4. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,... 

What is Kathy Saltzman's occupation?

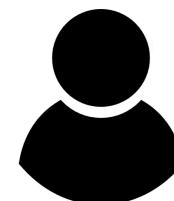


(Mallen et al., 2023)

GPT-3 davinci-003: 20%-30% accuracy

# Why retrieval-based LMs?

LLMs' knowledge is easily outdated and hard to update



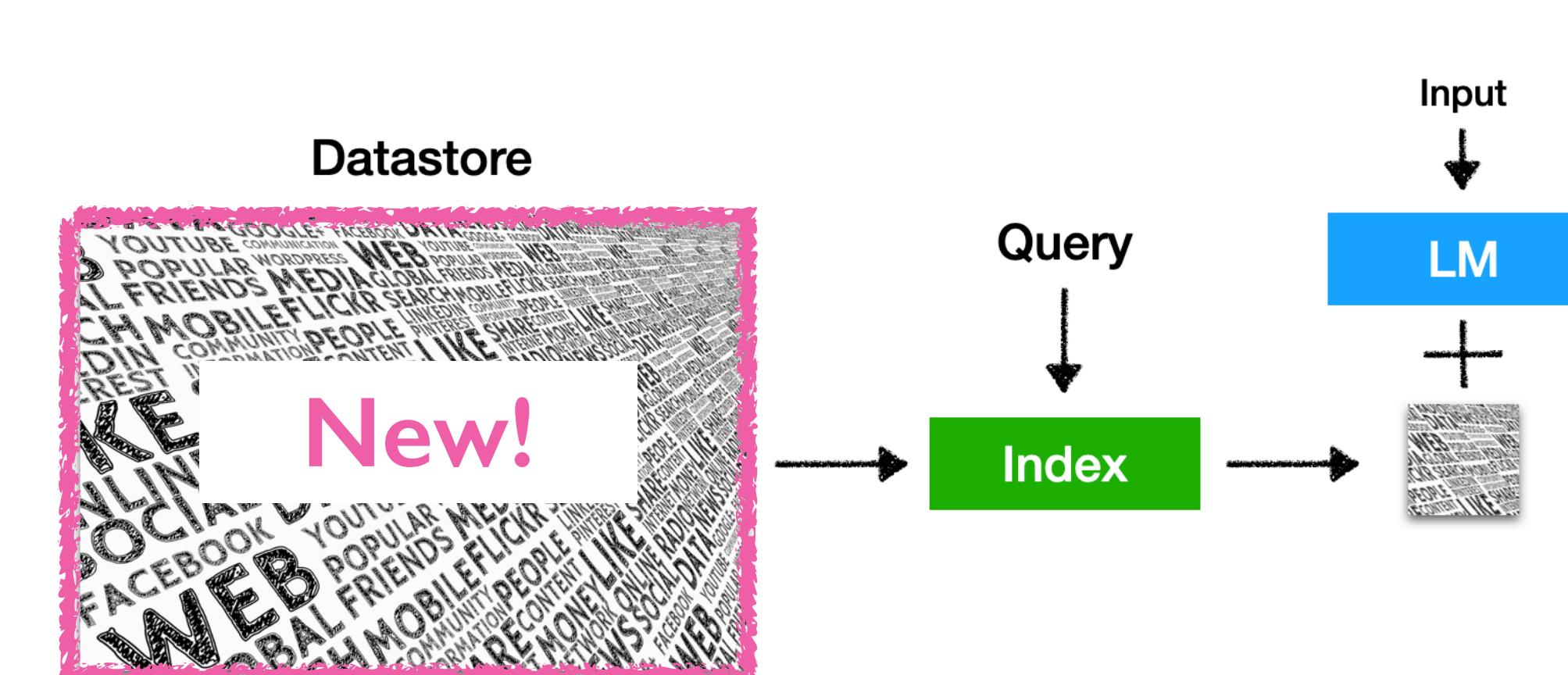
Who is the CEO of Twitter?



As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....

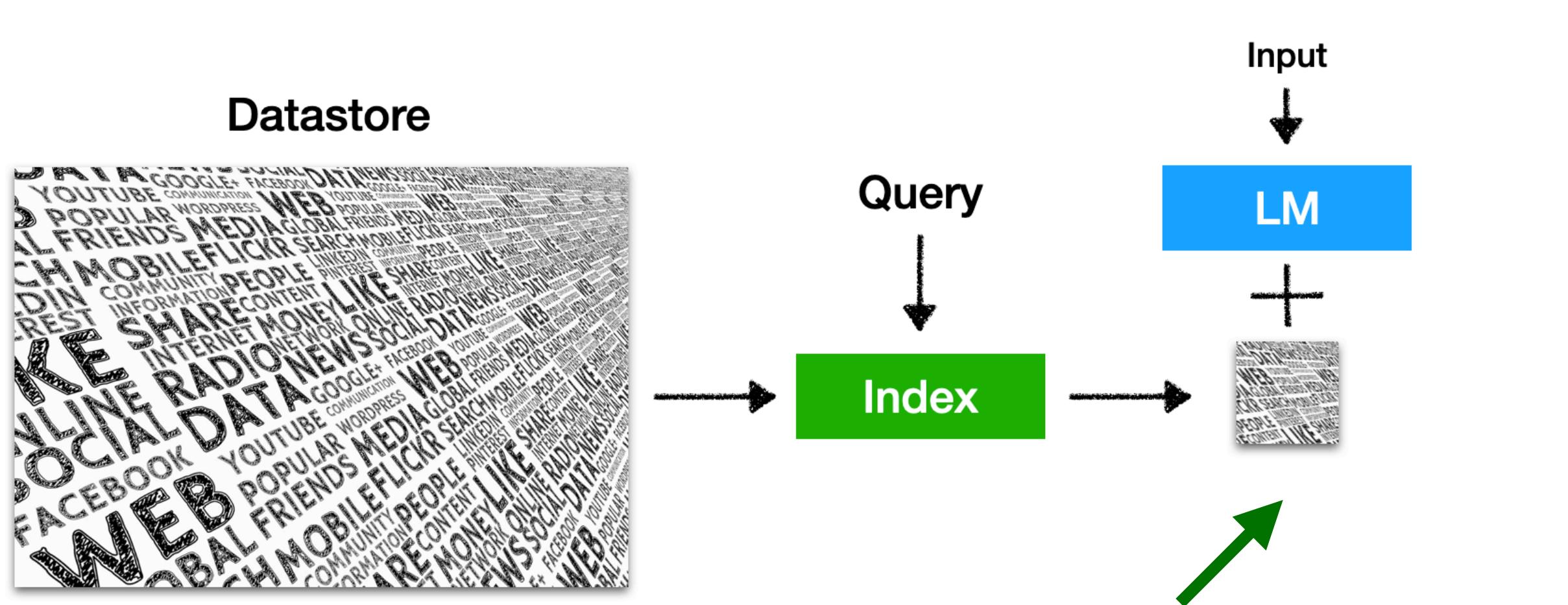
The screenshot shows a Google search results page. The search bar contains the query "Who is the CEO of Twitter?". Below the search bar are navigation links for "All", "News", "Images", "Shopping", "Videos", and "More". A status message indicates "About 1,090,000,000 results (0.45 seconds)". The top result is a snippet for "Twitter / CEO" featuring a photo of Linda Yaccarino, with the text "Linda Yaccarino" and "Jun 5, 2023–". To the right of the search results is a small portrait of Linda Yaccarino.

- Existing **knowledge editing** methods are still NOT scalable (**active research!**)
- The datastore can be easily **updated** and **expanded** - even without retraining!



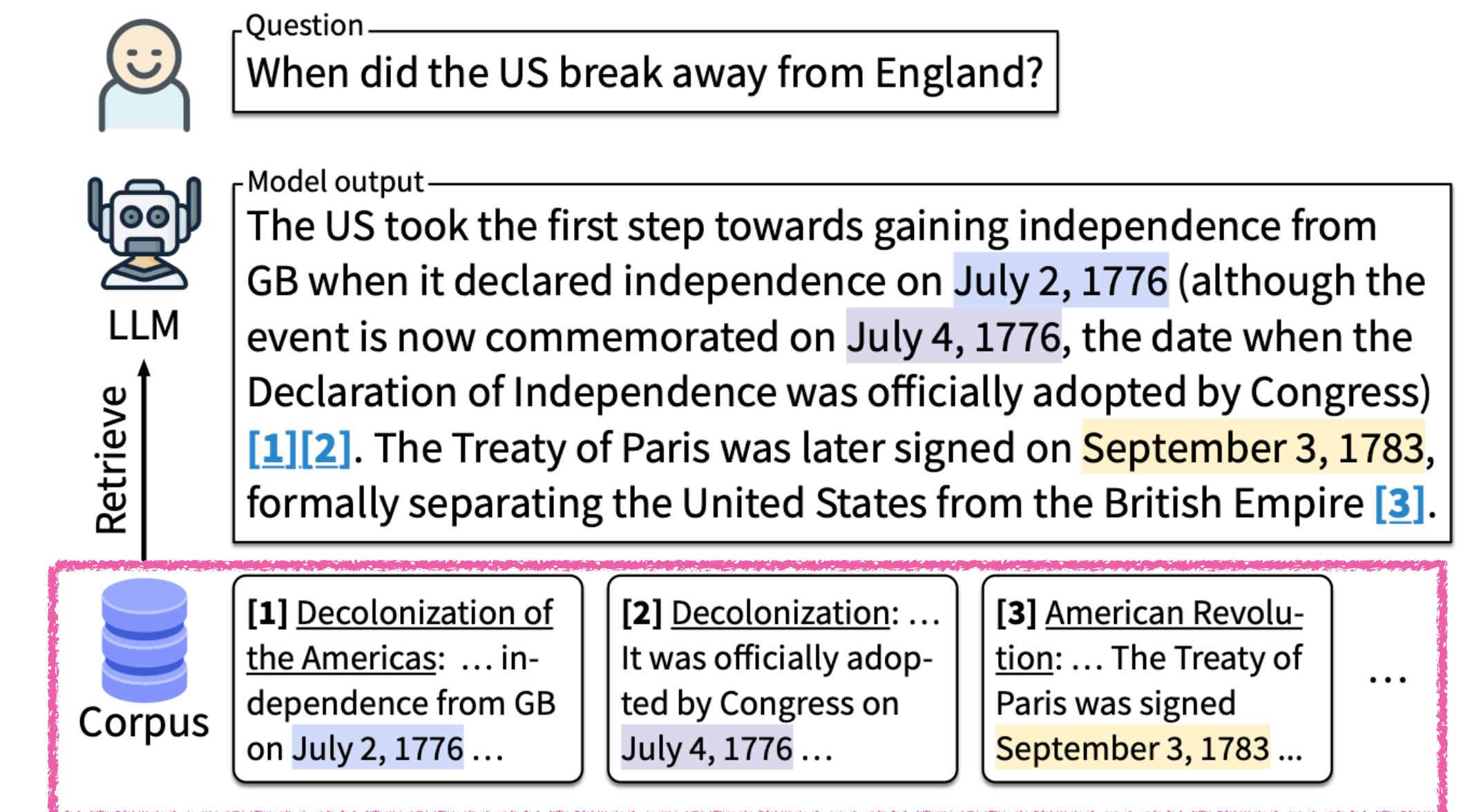
# Why retrieval-based LMs?

LLMs' output is challenging to interpret and verify



Can trace knowledge source  
from retrieval results - better  
interpretability & control

## Generating text with citations



(Nakano et al. 2021; Menick et al., 2022; Gao et al., 2023)

# Why retrieval-based LMs?

LLMs' output is challenging to interpret and verify

 Create an itinerary for exploring Toronto over a span of two days.  



Toronto is a vibrant and diverse city with many attractions and activities to suit different tastes and preferences. Here are some suggestions for creating an itinerary for exploring Toronto over two days:

- **Day 1:** Start your day with a visit to the **CN Tower**, the iconic landmark that offers panoramic views of the city and beyond <sup>1</sup>. You can also experience the thrill of walking on the glass floor or the EdgeWalk, the world's highest hands-free walk <sup>1</sup>. Next, head to the nearby **St. Lawrence Market**, one of the world's best food markets, where you can sample a variety of cuisines and local specialties <sup>2</sup>. After lunch, take a stroll along **Queen West**, a trendy neighborhood with eclectic shops, galleries, cafes and street art <sup>3</sup>. In the evening, enjoy a show at one of the many theaters or comedy clubs in the **Entertainment District**, or catch a game at the **Scotiabank Arena** if you're a sports fan.
- **Day 2:** Explore the history and culture of Toronto at the **Royal Ontario Museum**, one of the largest and most comprehensive museums in North America, featuring exhibits on art, natural history, world cultures and more <sup>4</sup>. Then, hop on a ferry to the **Toronto Islands**, a group of islands that offer a relaxing escape from the city, with beaches, parks, trails and amusement rides <sup>3</sup> <sup>5</sup>. You can also rent a bike or kayak to explore the islands at your own pace. For dinner, head to **Chinatown**, one of the largest and most vibrant in North America, where you can find a variety of Asian cuisines and shops <sup>3</sup>.

I hope this helps you plan your trip to Toronto. Have fun! 😊

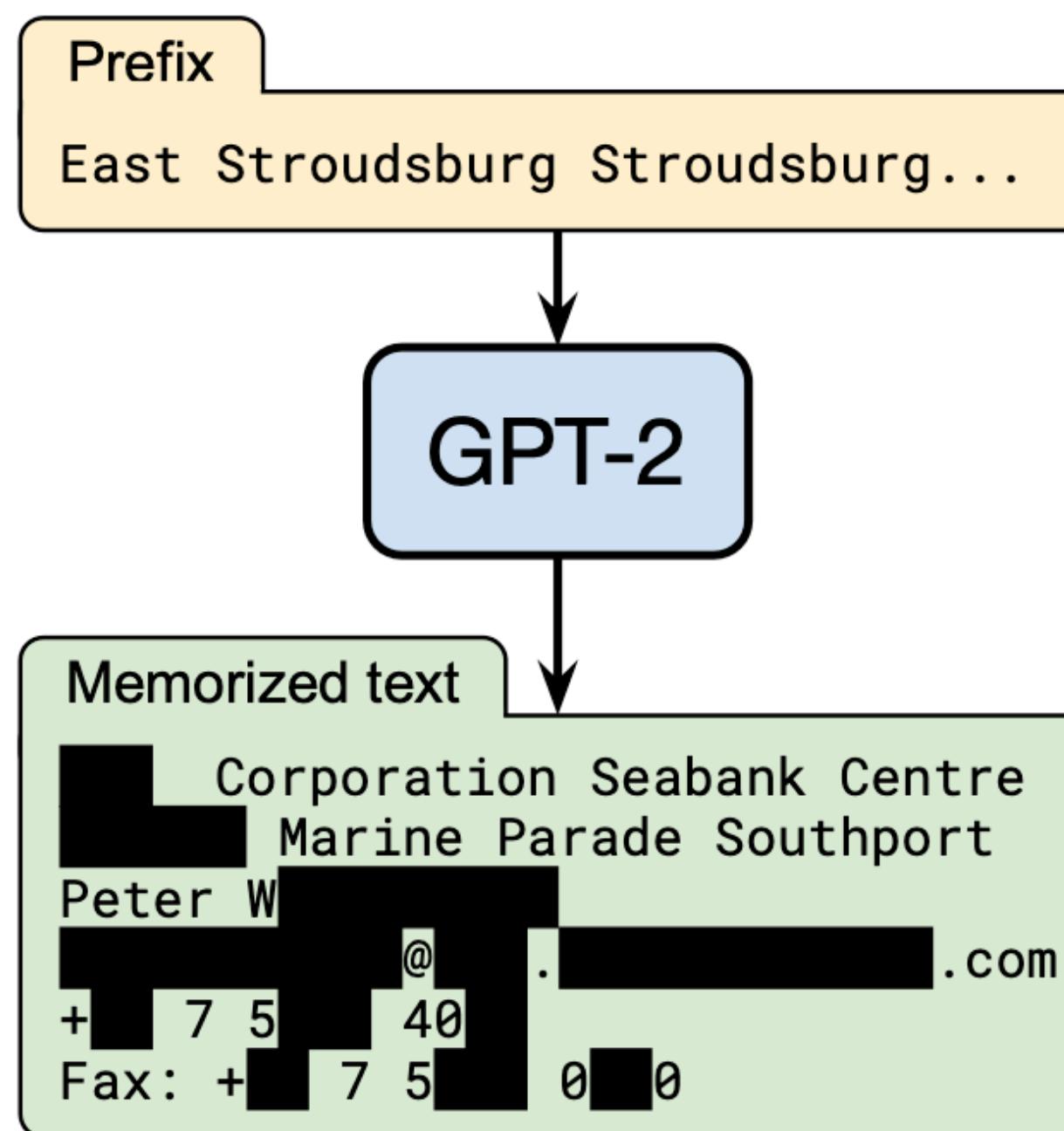
**Learn more:**

1. [cntower.ca](http://cntower.ca)   2. [travel.usnews.com](http://travel.usnews.com)   3. [bing.com](http://bing.com)

4. [rom.on.ca](http://rom.on.ca)   5. [tripadvisor.com](http://tripadvisor.com)

# Why retrieval-based LMs?

LLMs are shown to easily leak private training data

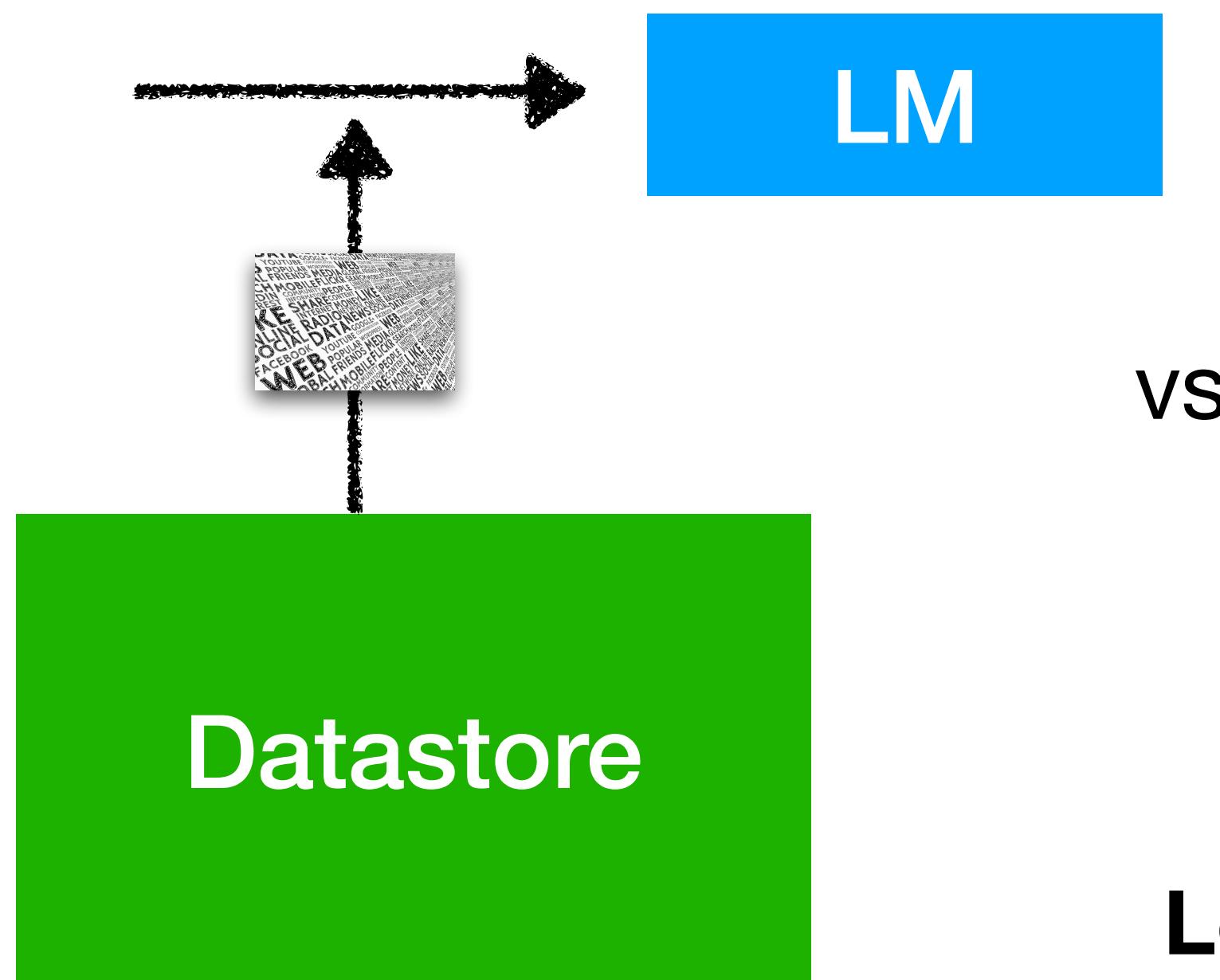


Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
<b>Named individuals (non-news samples only)</b>	<b>46</b>
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
<b>Contact info (address, email, phone, twitter, etc.)</b>	<b>32</b>
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

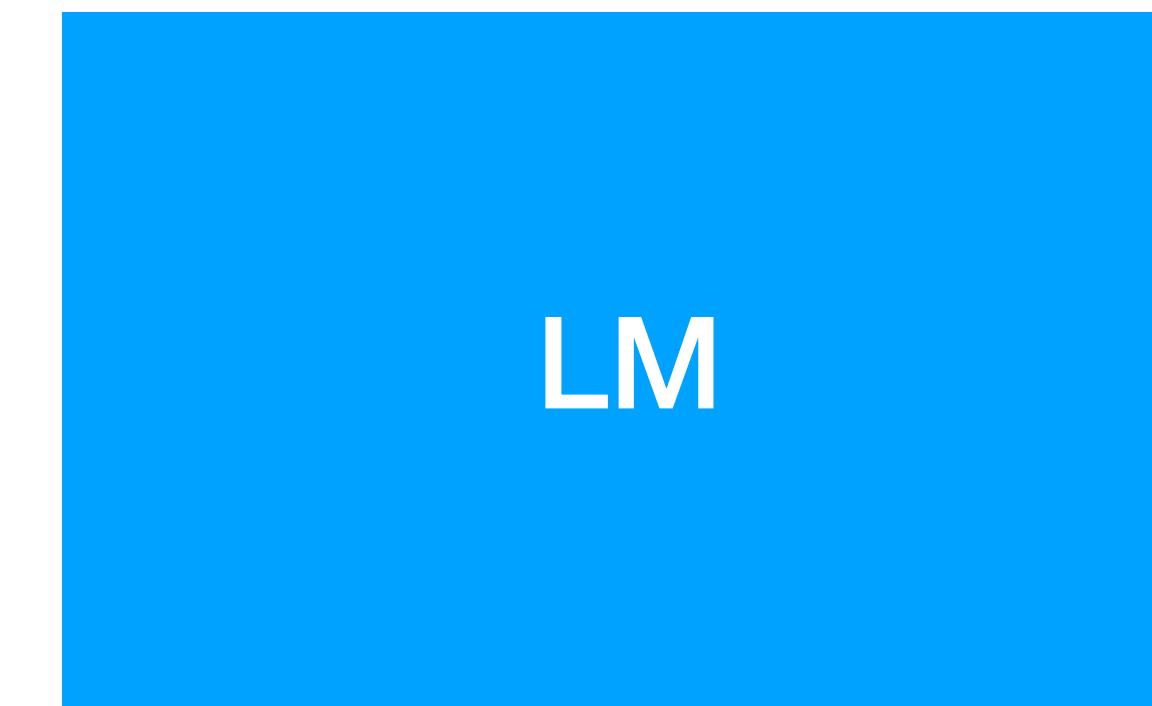
Individualization on private data by storing it in the datastore

# Why retrieval-based LMs?

LLMs are **\*large\*** and expensive to train and run



vs.



**Long-term goal:** can we possibly reduce the **training and inference costs**, and scale down the size of LLMs?

e.g., RETRO (Borgeaud et al., 2021): “obtains comparable performance to GPT-3 and Jurassic-1 on the Pile, despite using **25x fewer parameters**”

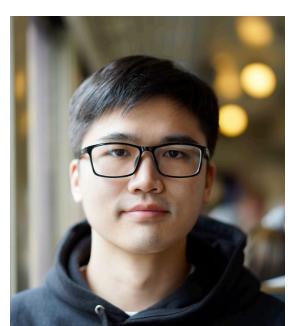
# Goals of the tutorial

- We will cover a number of key developments on retrieval-based LMs (mostly 2020–2023)
  - **Problem definition**
  - **Architecture design**
  - **Training methods**
  - **Applications and extensions**
- This tutorial is **cutting-edge**, and we are still far from understanding how to best develop retrieval-based LMs compared to parametric LLMs:
  - Taxonomies of existing research and key insights
  - Our perspectives on the current challenges & open problems



# Schedule

Time	Section	Presenter
14:00—14:15	Section 1: Introduction <a href="#">[Slides]</a>	Danqi
14:15—14:25	Section 2: Definition & Preliminaries <a href="#">[Slides]</a>	Sewon
14:25—15:05	Section 3: Retrieval-based LMs: Architecture <a href="#">[Slides]</a>	Sewon
15:05—15:20	Section 4: Retrieval-based LMs: Training <a href="#">[Slides]</a>	Zexuan
15:20—15:30	Q & A Session I	
 <b>30min coffee break</b>		
16:00—16:30	Section 4 (Cont'd): Retrieval-based LMs: Training <a href="#">[Slides]</a>	Zexuan
16:30—17:05	Section 5: Retrieval-based LMs: Applications <a href="#">[Slides]</a>	Akari
17:05—17:15	Section 6: Extension: Multilingual & Multimodal <a href="#">[Slides]</a>	Akari
17:15—17:20	Section 7: Challenges & Opportunities <a href="#">[Slides]</a>	Danqi
17:20—17:30	Q & A Session II	



**Q & A**

