



[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

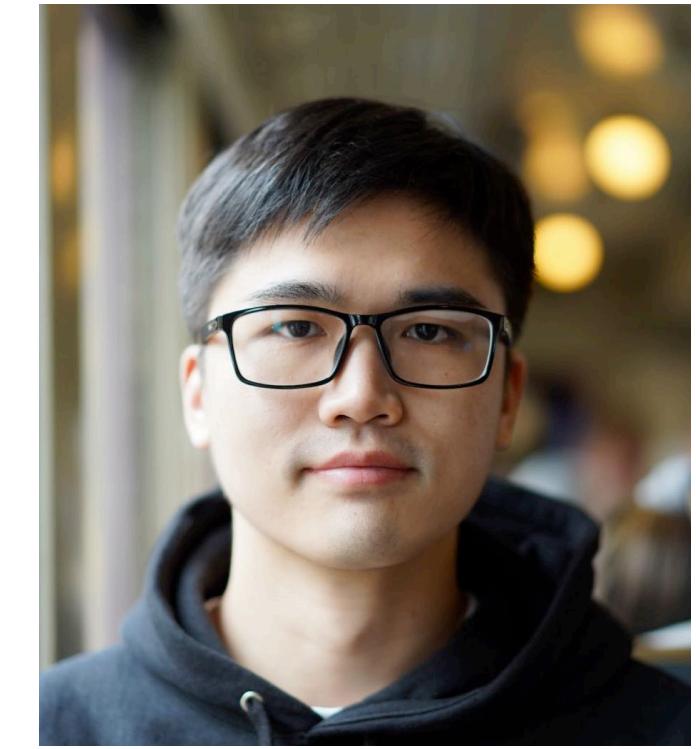
About presenters



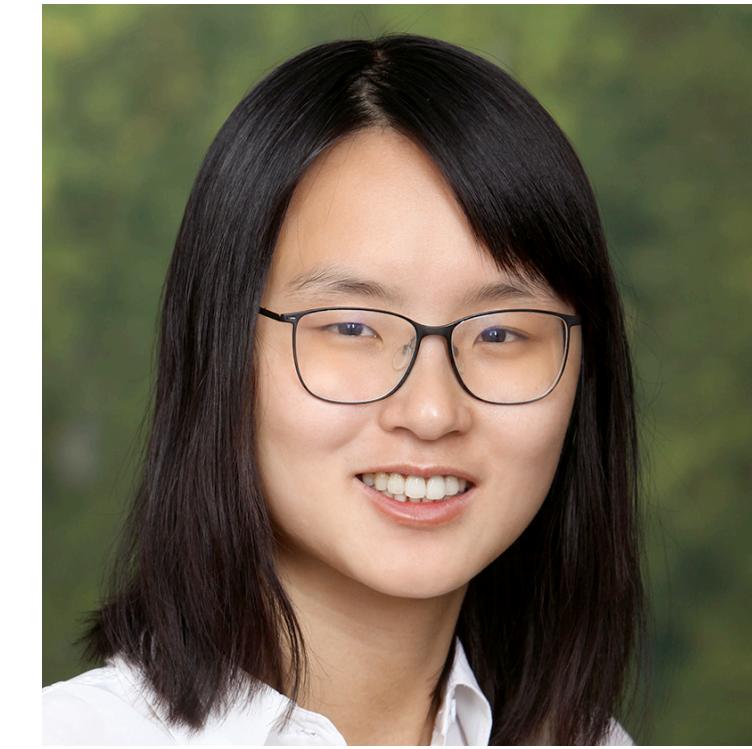
Akari Asai
PhD student
@UW



Sewon Min
PhD student
@UW



Zexuan Zhong
PhD student
@Princeton



Danqi Chen
Faculty
@Princeton

Participation and Q & A

- All the tutorial slides and references 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)

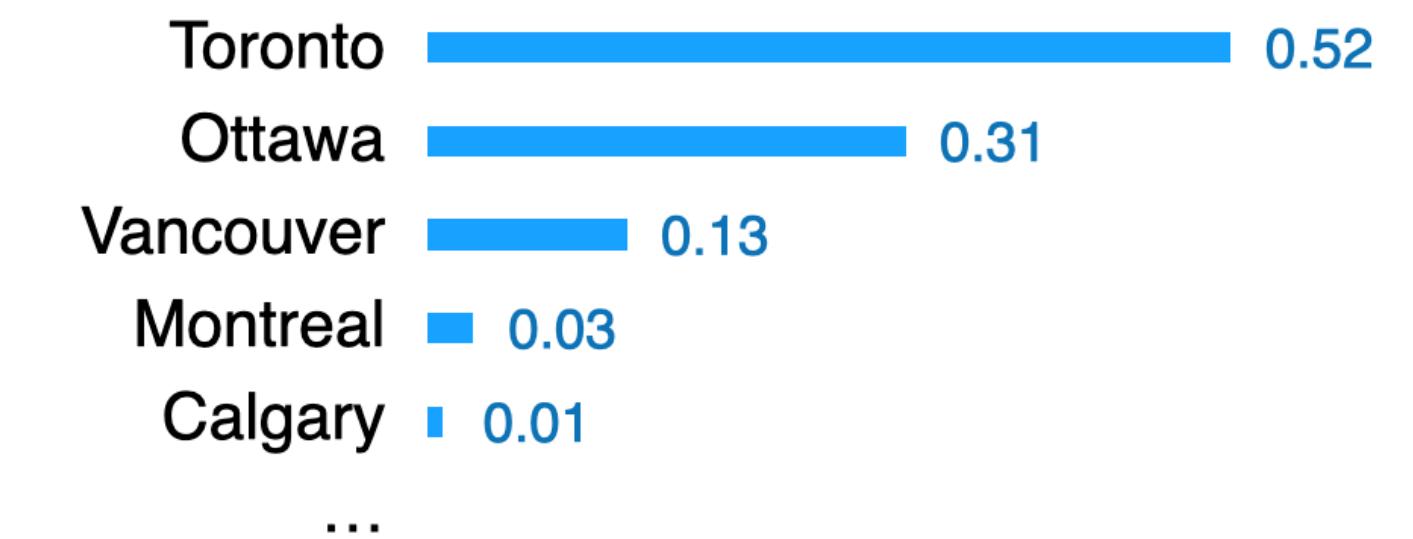
Retrieval-based language models (LMs)

Retrieval-based LMs = Retrieval + LMs

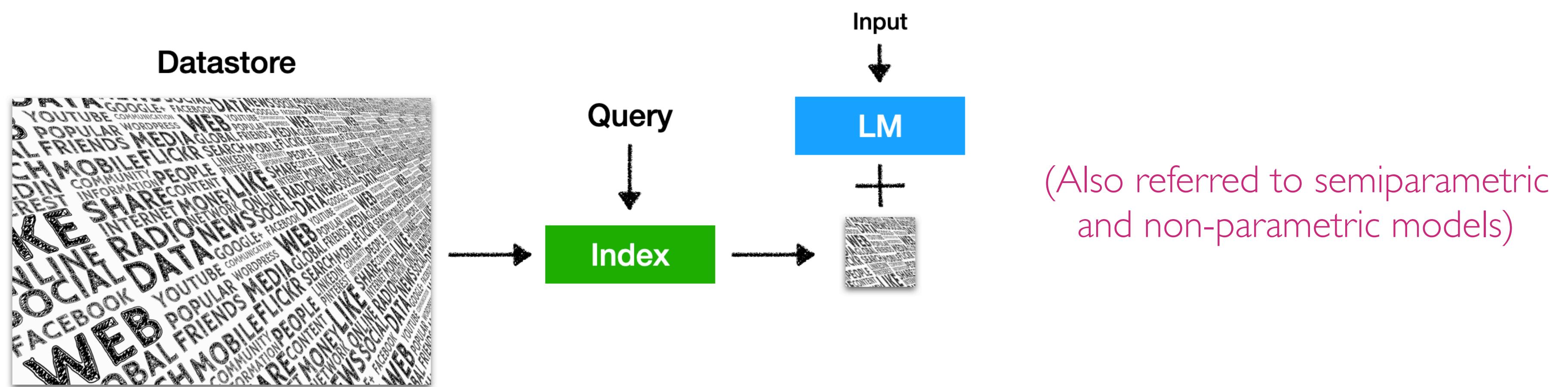
- It is a **language model**

$$P(x_n | x_1, x_2, \dots, x_{n-1})$$

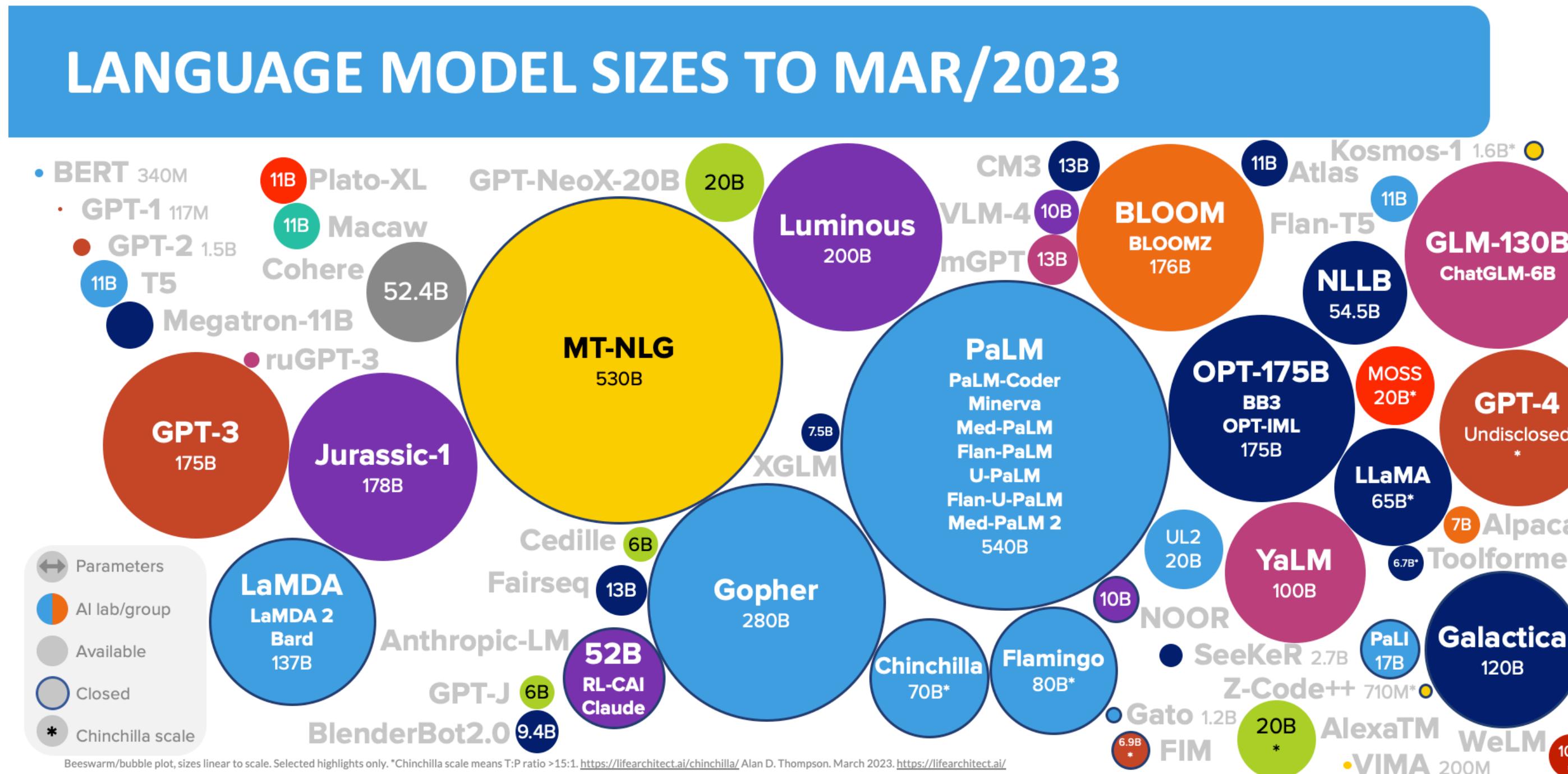
The capital city of Ontario is _____
(can be broadly extended to masked language
models or encoder-decoder models)



- 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 ↑, data size ↑

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

Retrieval for knowledge-intensive NLP tasks

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

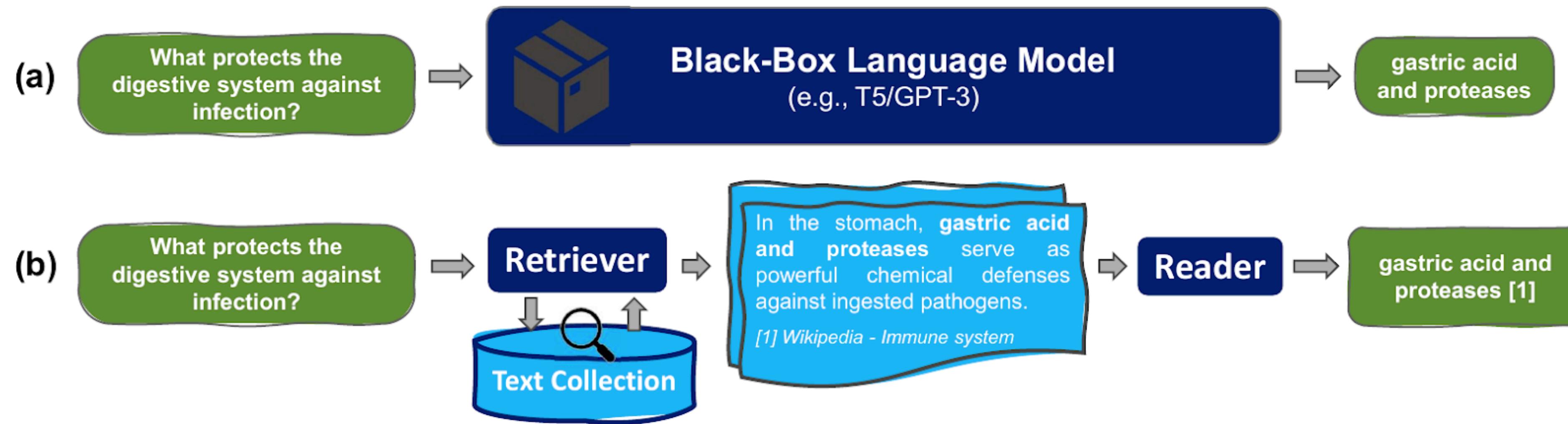


Image: <http://ai.stanford.edu/blog/retrieval-based-NLP/>

Drives a lot of research on better algorithms for **dense retrieval**, e.g., **DPR** (Karpukhin et al., 2020), **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

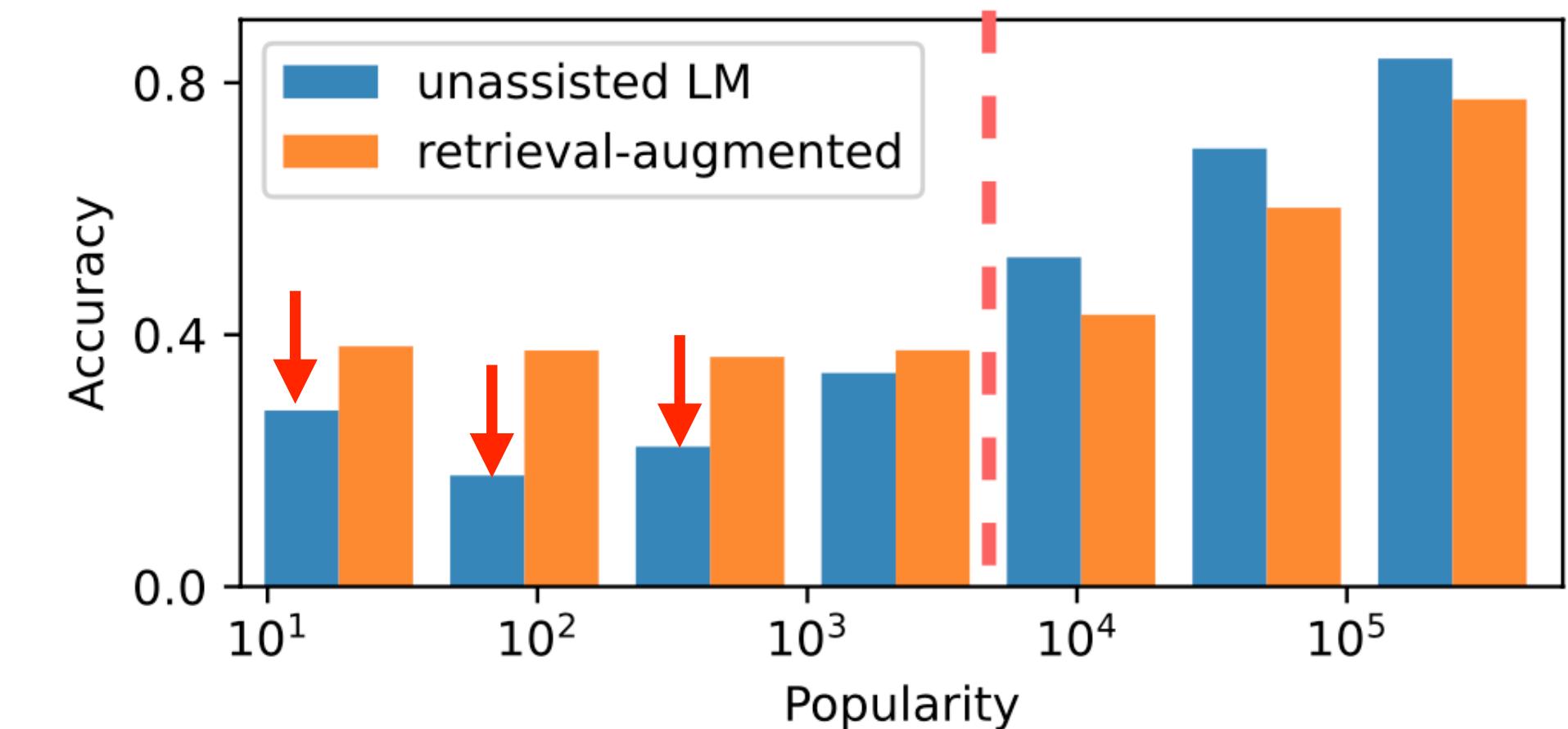


Geoffrey Hinton is a renowned computer scientist ...
Here are five important papers authored by him:

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. ...
4. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,...
5. "**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?

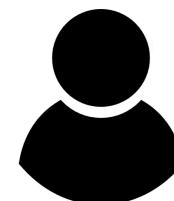
GPT-3 davinci-003: 20%-30% accuracy



(Mallen et al., 2023)

Why retrieval-based LMs?

LLMs' knowledge is easily outdated and hard to update



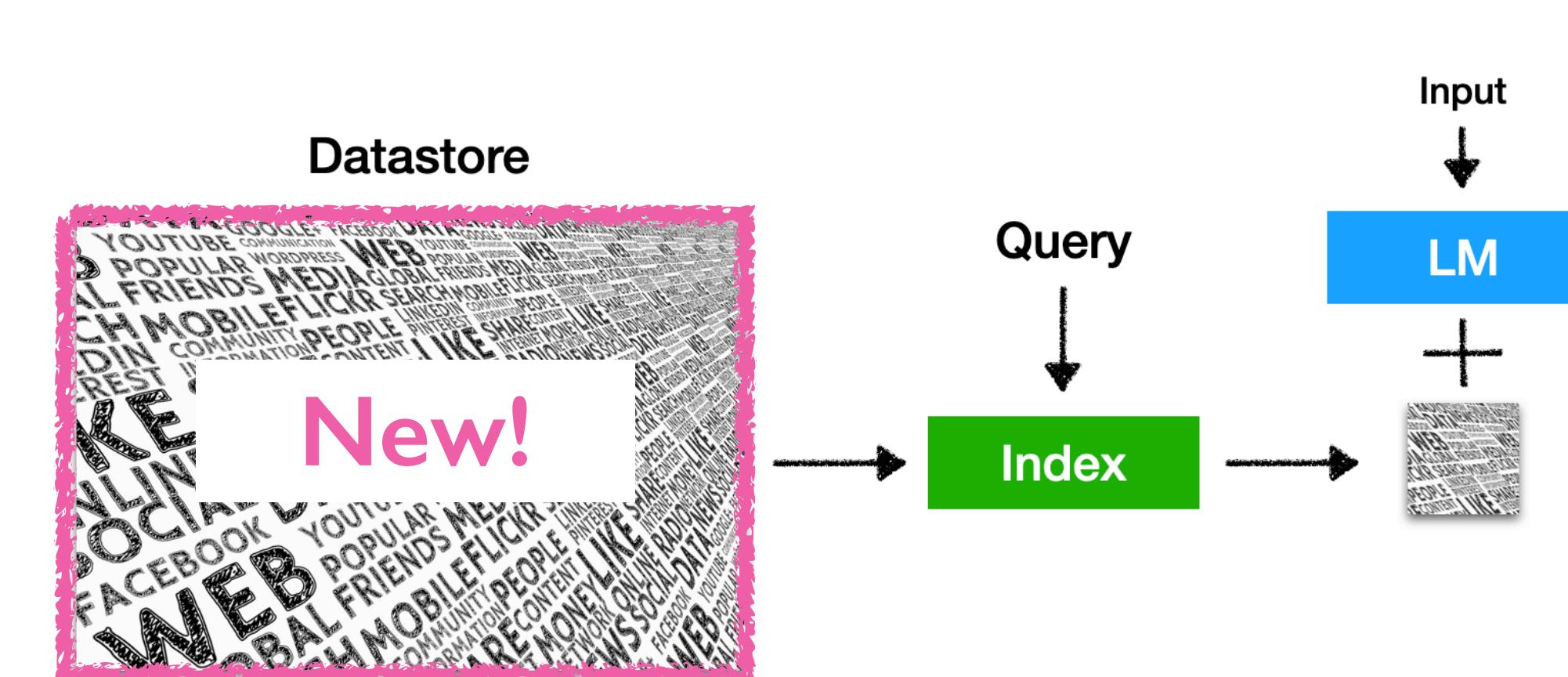
Who is the CEO of Twitter?



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

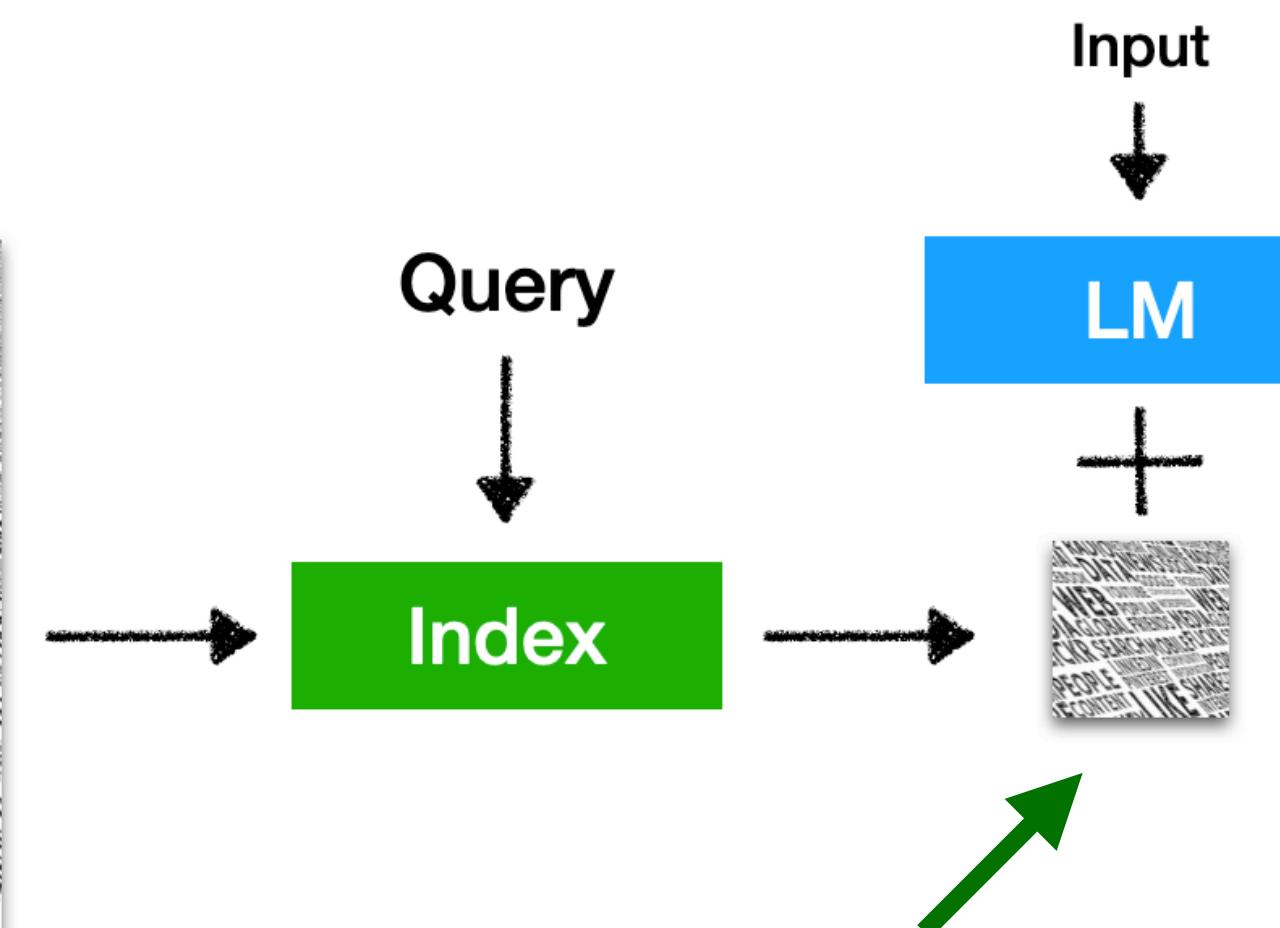
The screenshot shows a Google search results page. The search bar contains the query "Who is the CEO of Twitter?". Below the search bar are navigation links for "All", "News", "Images", "Shopping", "Videos", and "More". A status message indicates "About 1,090,000,000 results (0.45 seconds)". The top result is a snippet for "Twitter / CEO" followed by the name "Linda Yaccarino" and a small profile picture of a woman. The date "Jun 5, 2023" is shown next to the snippet.

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



Why retrieval-based LMs?

LLMs' output is challenging to interpret and verify



We can trace information source from retrieval results - better interpretability & control

Generating text with citations

Which river is longer, the Nile or the Yangtze?

The Nile is longer than the Yangtze River.^{[1][2][3]} The Nile runs 6,650 kilometers (4,130 miles) between Lake Victoria and the Mediterranean Sea.^[2] The Yangtze River runs 6,300 kilometers (3,915 miles) from its source in the Tanggula Mountains in Qinghai to the East China Sea.^[3]

1. [World's Longest Rivers | Amazon, Nile & Mississippi Rivers | Live Science \(www.livescience.com\)](#) ↗
2. [Nile - Wikipedia \(en.wikipedia.org\)](#) ↗
3. [Yangtze River Length - 6,300 Kilometers \(3,915 Miles\) \(www.travelchinaguide.com\)](#) ↗
4. [How Long is the Yangtze River? Yangtze River's Length \(www.chinatraveldiscovery.com\)](#) ↗

Retrieval comes into play!

Why retrieval-based LMs?

LLMs' output is challenging to interpret and verify

What link schemes are against Google's webmaster guidelines?

✓ Searching for: **link schemes against Google's webmaster guidelines**

✓ Generating answers for you...

Link schemes are against Google's Webmaster Guidelines. According to Google, link schemes are defined as "any links intended to manipulate PageRank or a site's ranking in Google search results may be considered part of a link scheme and a violation of Google's Webmaster Guidelines. This includes any behavior that manipulates links to your site or outgoing links from your site." [1](#) [2](#)

Examples of link schemes include buying or selling links that pass PageRank, exchanging money for links, excessive link exchanges, large-scale article marketing or guest posting campaigns with keyword-rich anchor text links, using automated programs or services to create links to your site, and more. [3](#) [4](#)

I hope this helps!

Learn more: [1. bing.com](#) [2. thecontentworks.uk](#) [3. tikkatwins.com](#) [4. seo-hacker.com](#) [5. support.google.com](#) [6. developers.google.com](#) [— see less](#)

1 of 20

Bing Chat

what is the capital of London? X ►

⚡ PERPLEXITY

London is the capital and largest city of England and the United Kingdom^{[1][2]}, with a population of just under 9 million^[1]. It has been the de facto capital since ancient times^{[3][4]}, and its status as capital was officially confirmed in the 16th century^[5].

Like Dislike

[] CITATIONS

1. London - Wikipedia
<https://en.wikipedia.org/wiki/london>
London is the capital and largest city of England and the United Kingdom, with a population of just under 9 million. ... It stands on the River Thames in south- ...

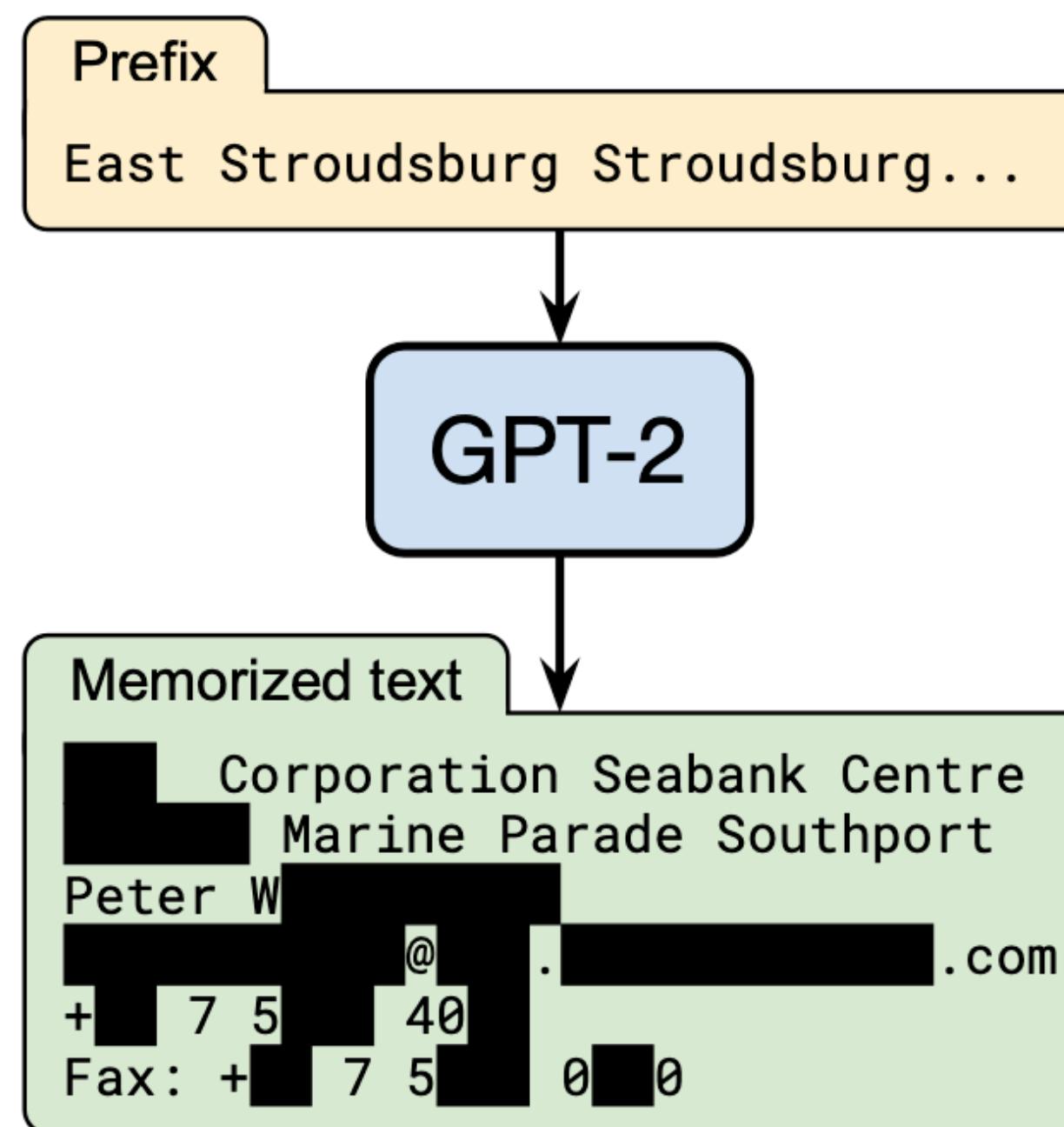
2. London | History, Maps, Population, Area, & Facts | Britannica
<https://www.britannica.com/place/london>
London, city, capital of the United Kingdom. It is among the oldest of the world's great cities—its history spanning nearly two ...

3. Has London always been the capital of the UK?
<https://evanevanstours.com/blog/has-london-always-been-the-capital-of-the-uk>
The UK has had more than one capital. While London is the UK's de facto capital, its status as capital has never

Perplexity.ai

Why retrieval-based LMs?

LLMs are shown to easily leak private training data

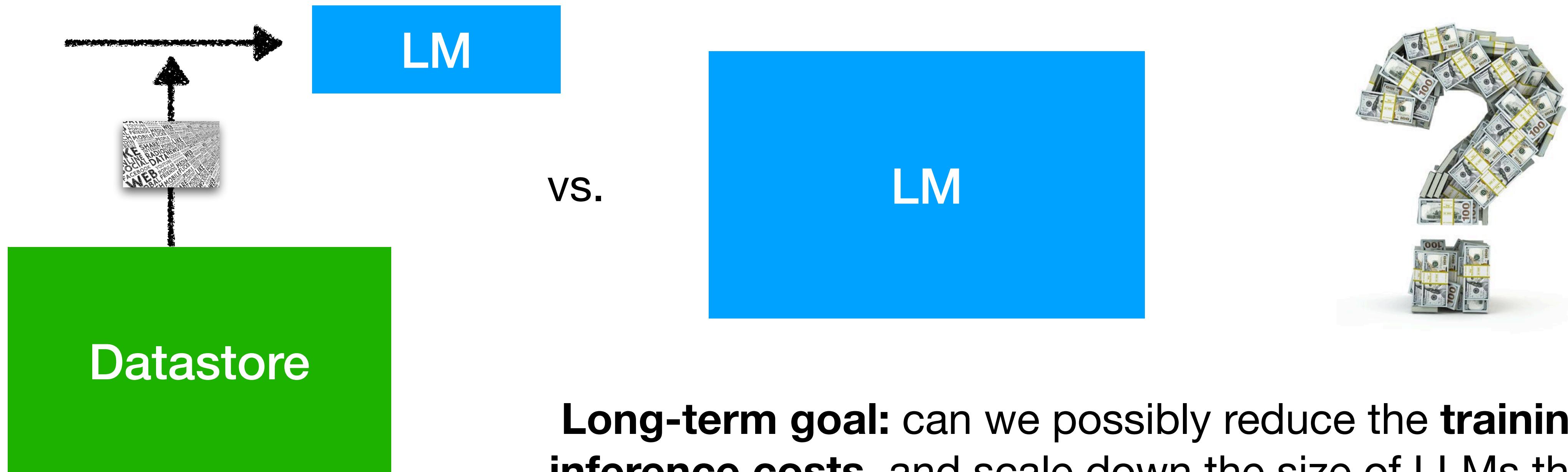


| Category | Count |
|--|-----------|
| US and international news | 109 |
| Log files and error reports | 79 |
| License, terms of use, copyright notices | 54 |
| Lists of named items (games, countries, etc.) | 54 |
| Forum or Wiki entry | 53 |
| Valid URLs | 50 |
| Named individuals (non-news samples only) | 46 |
| Promotional content (products, subscriptions, etc.) | 45 |
| High entropy (UUIDs, base64 data) | 35 |
| Contact info (address, email, phone, twitter, etc.) | 32 |
| 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



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

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:** how LMs and retrieval are combined?
 - **Training methods:** how to train LMs and retrievers?
 - **Applications and extensions** to a range of problems
- This tutorial is **cutting-edge**, and we are still far from understanding how to develop retrieval-based LMs to be competitive with parametric LLMs
 - Taxonomies of existing research and key insights
 - Our perspectives on the current challenges & open problems



Schedule

| Time | Section | Presenter |
|-------------|---|-----------|
| 14:00—14:10 | Section 1: Introduction | Danqi |
| 14:10—14:20 | Section 2: Definition & Preliminaries | Sewon |
| 14:20—15:00 | Section 3: Retrieval-based LMs: Architecture | Sewon |
| 15:00—15:20 | Section 4: Retrieval-based LMs: Training | Zexuan |
| 15:20—15:30 | Q & A Session I | |
| 16:00—16:25 | Section 4 (Cont'd): Retrieval-based LMs: Training | Zexuan |
| 16:25—17:00 | Section 5: Retrieval-based LMs: Application | Akari |
| 17:00—17:10 | Section 6: Extension: Multilingual & Multimodal | Akari |
| 17:10—17:20 | Section 7: Conclusion: Challenges & Opportunities | Danqi |
| 17:20—17:30 | Q & A Session II | |



30min coffee break

Q & A

