

RAG

Instructor: Lara J. Martin (she/they)

TA: Omkar Kulkarni (he)

Slides adapted from an [ACL 2023 Tutorial by Akari Asai, Sewon Min, Zexuan Zhong, & Dr. Danqi Chen](#)

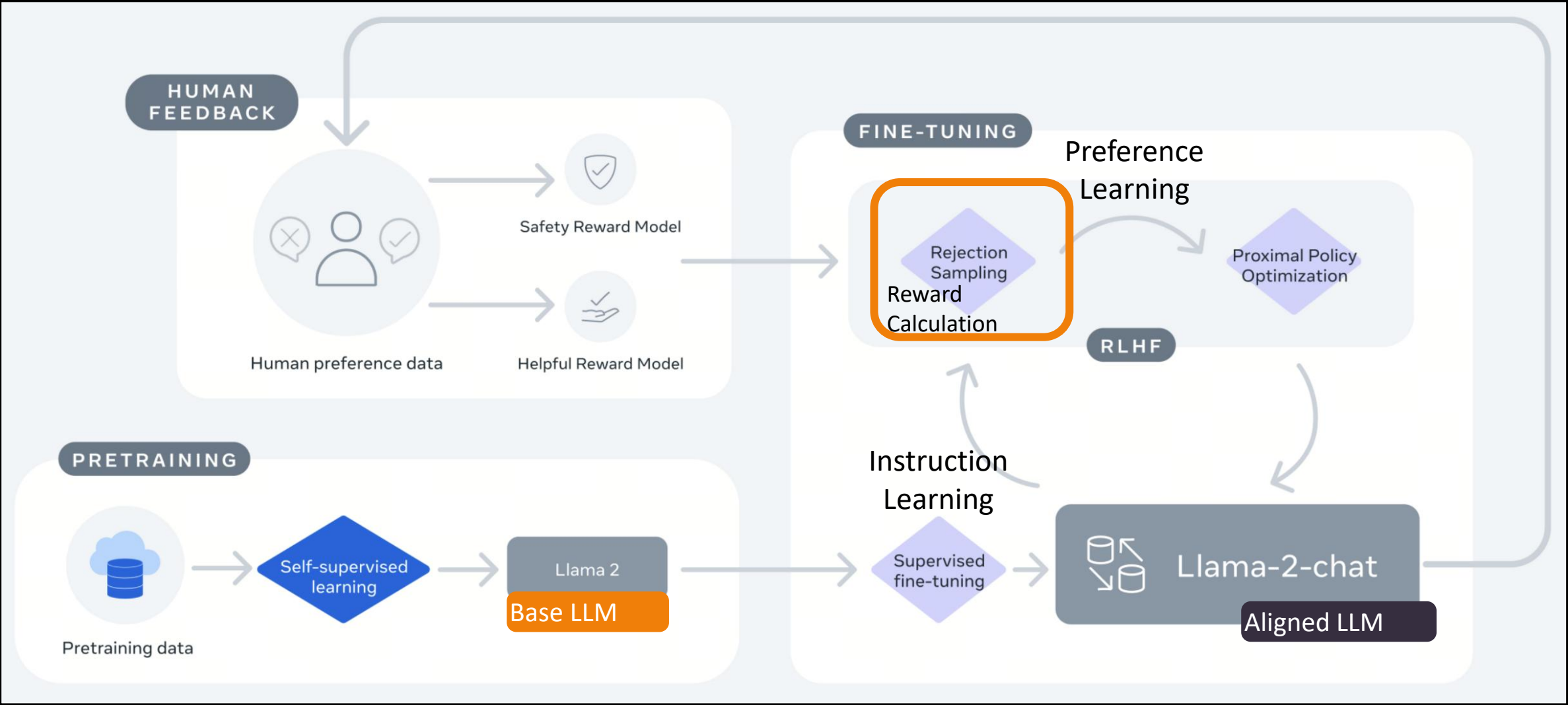
Learning Objectives

Understand the reasons why RAG was created

Explore how the retrieval component interacts with the LLM in RAG

Extract implementation details from papers and find different ways RAG is implemented

Example: Llama-2's alignment

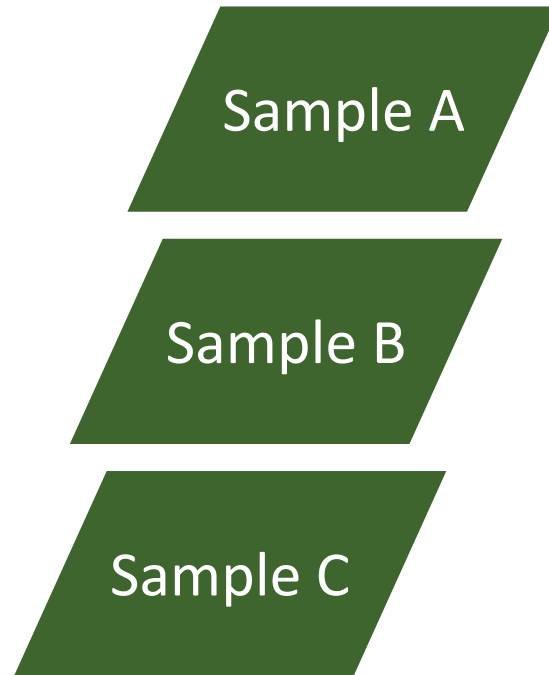


Review: Rejection Sampling

A set of sampled completions
for a prompt

Triples

Prompt



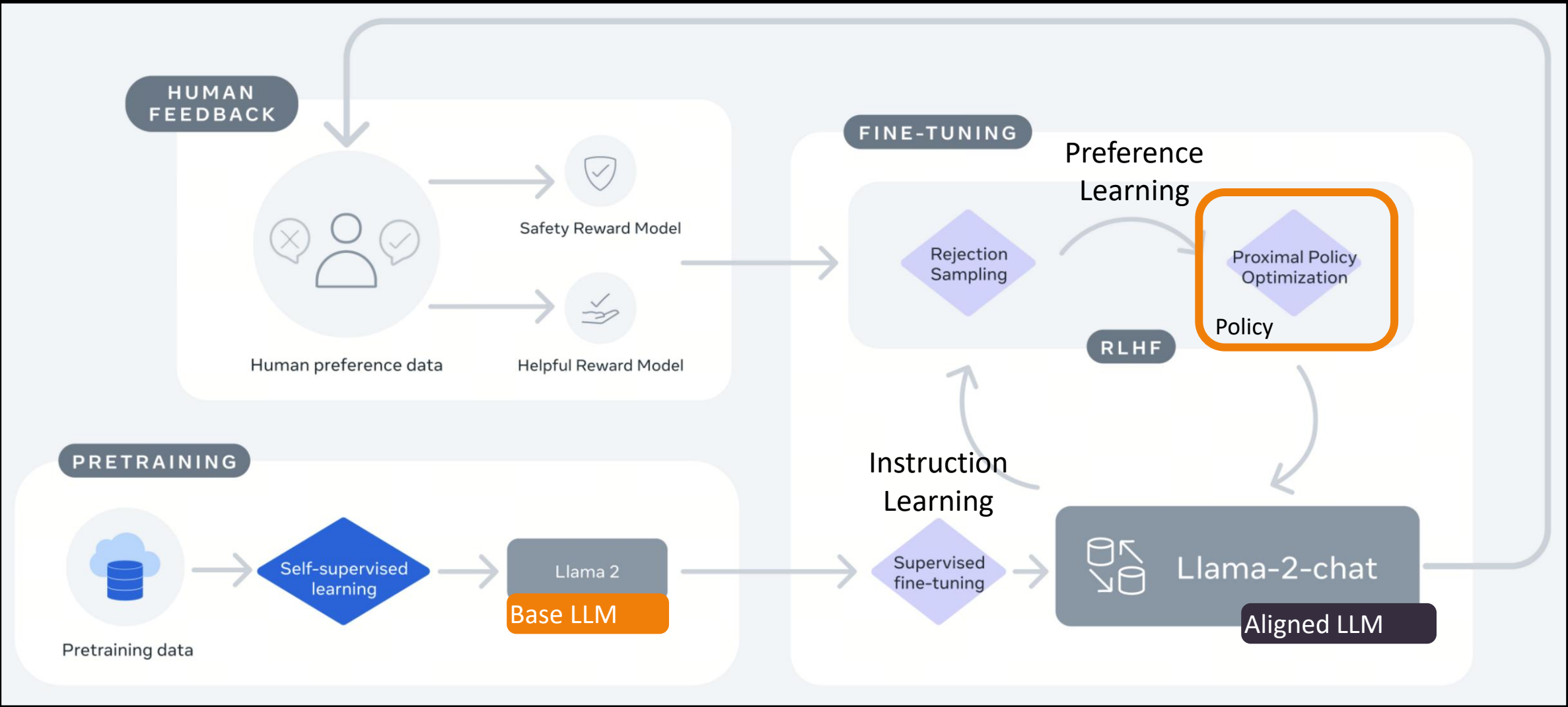
$$D = \{x^i, y_w^i, y_l^i\}$$

Prompt

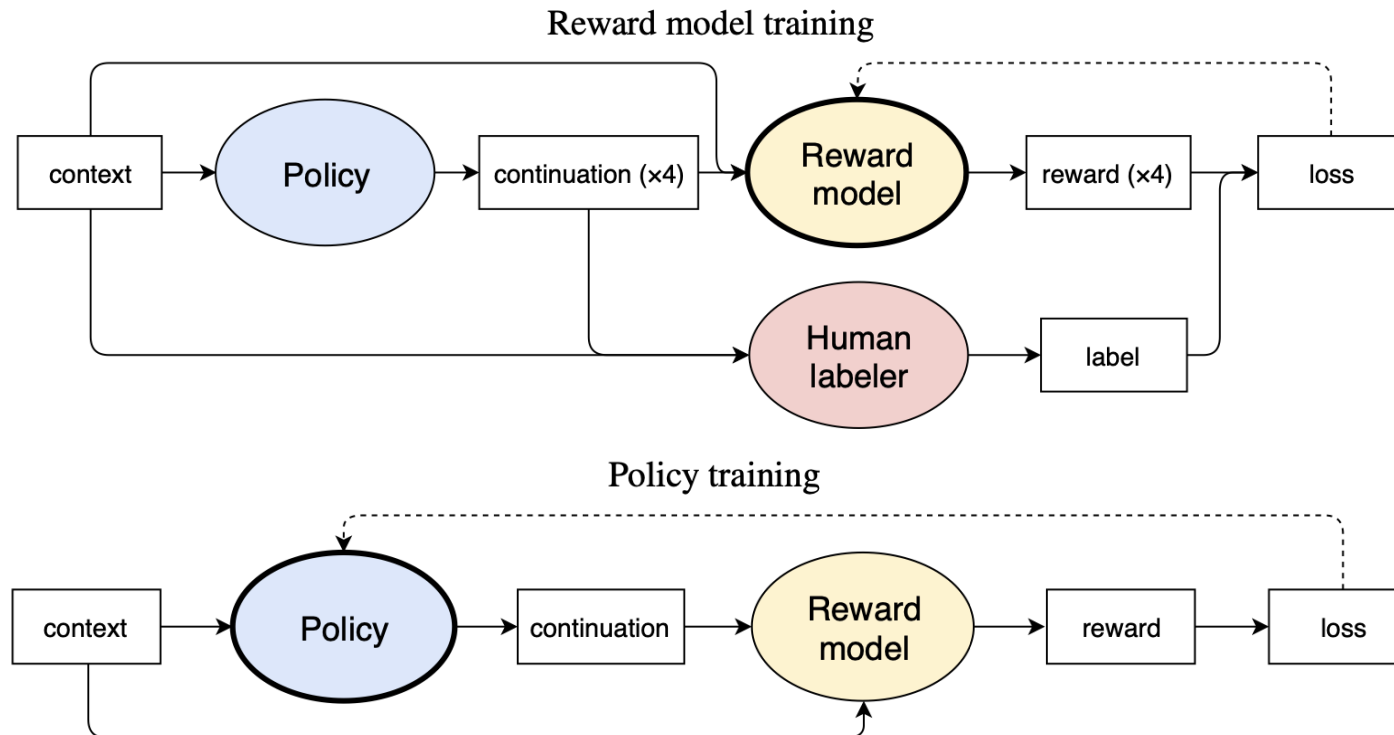
Preferred Response

Dispreferred Response

Example: Llama-2's alignment



Review: Policy Training with Reward Model



Ziegler et al., 2019 "Fine-Tuning Language Models from Human Preferences"

Review: Efficient LLMs

Methods for shrinking pre-existing models

Methods for mimicking pre-existing models with smaller models

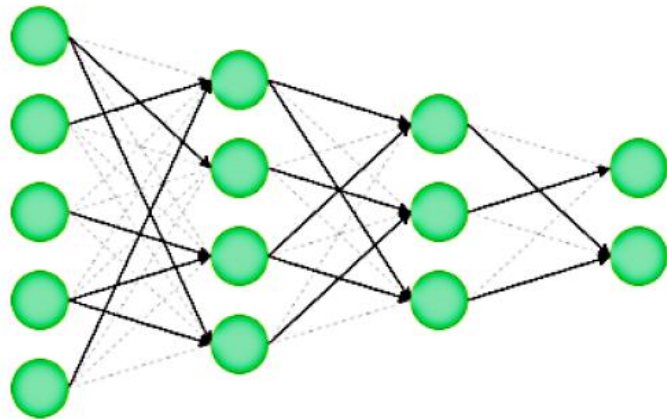
Methods for faster/smaller finetuning of pre-existing models

Methods for training new models more efficiently

Review: Pruning

Remove parts of the model

Unstructured Pruning



Structured Pruning

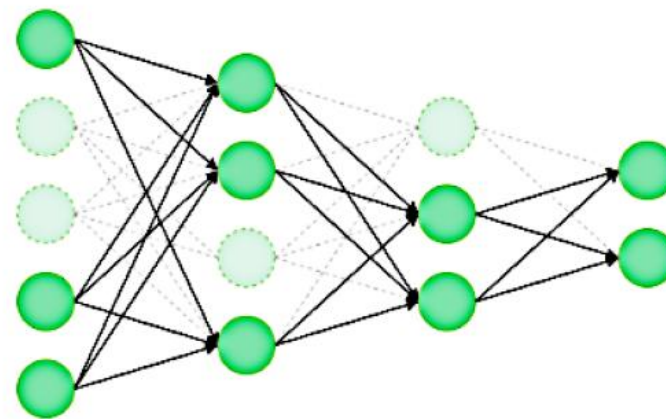
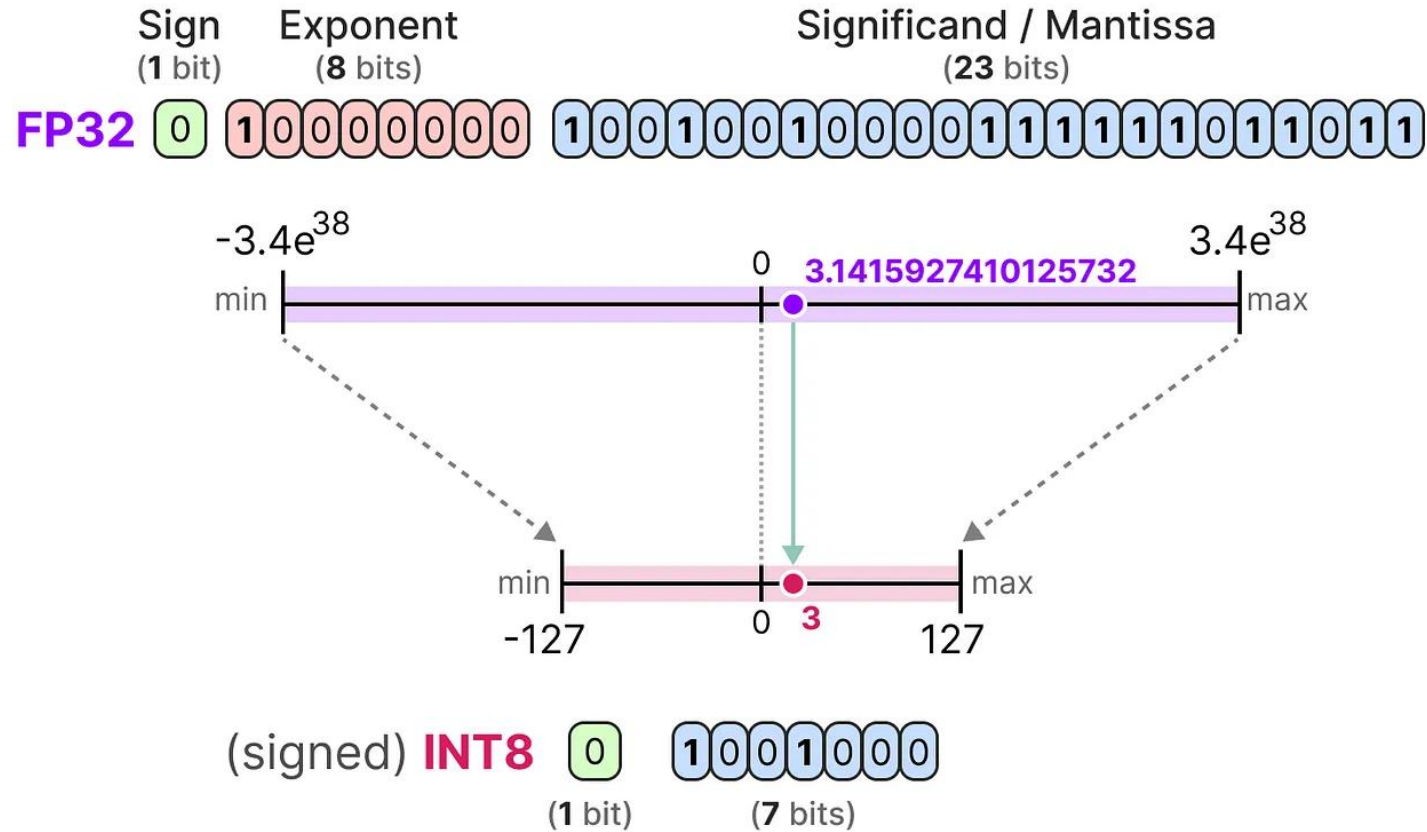


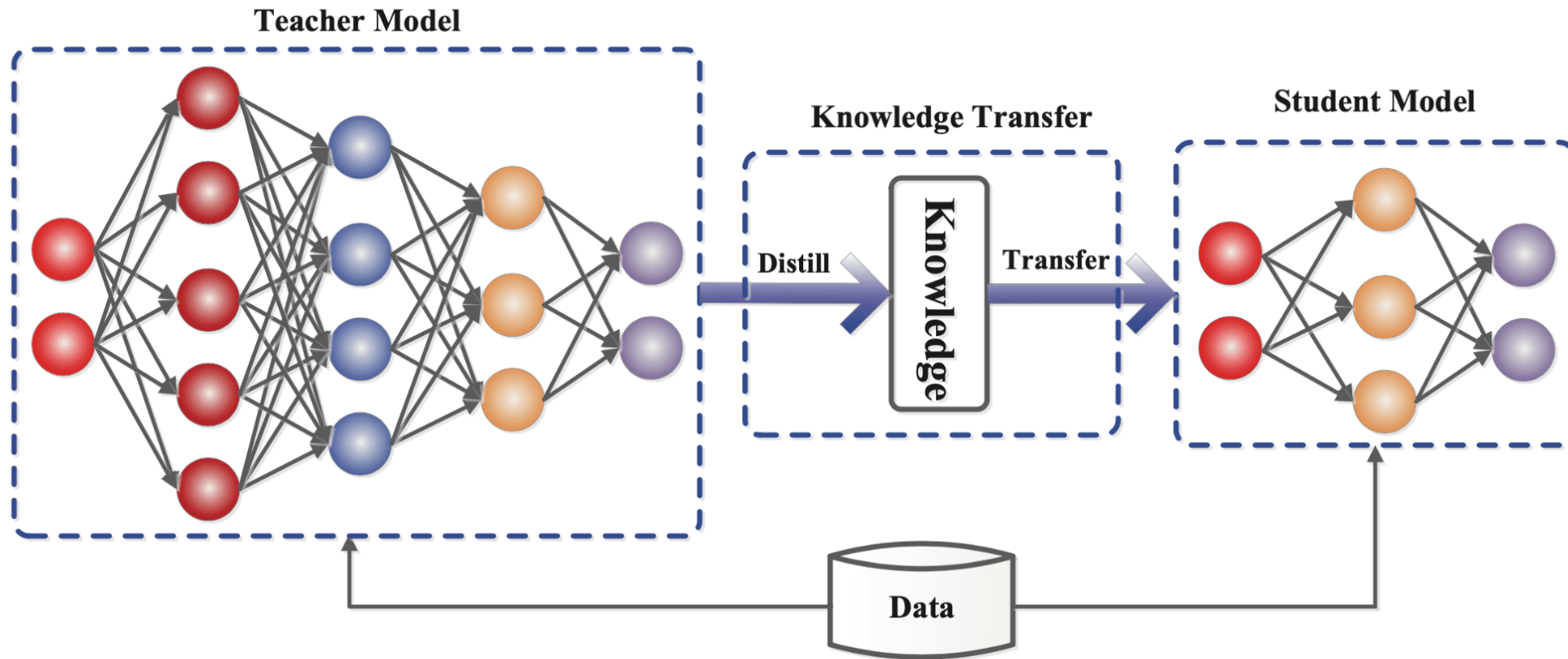
Image credits: neuralmagic.com

Implementation Tutorial: https://pytorch.org/tutorials/intermediate/pruning_tutorial.html

Review: Quantizing Models



Review: Knowledge Distillation



<https://neptune.ai/blog/knowledge-distillation>

Implementation Tutorial: https://pytorch.org/tutorials/beginner/knowledge_distillation_tutorial.html

Review: LoRA (Low-Rank Adaptation)

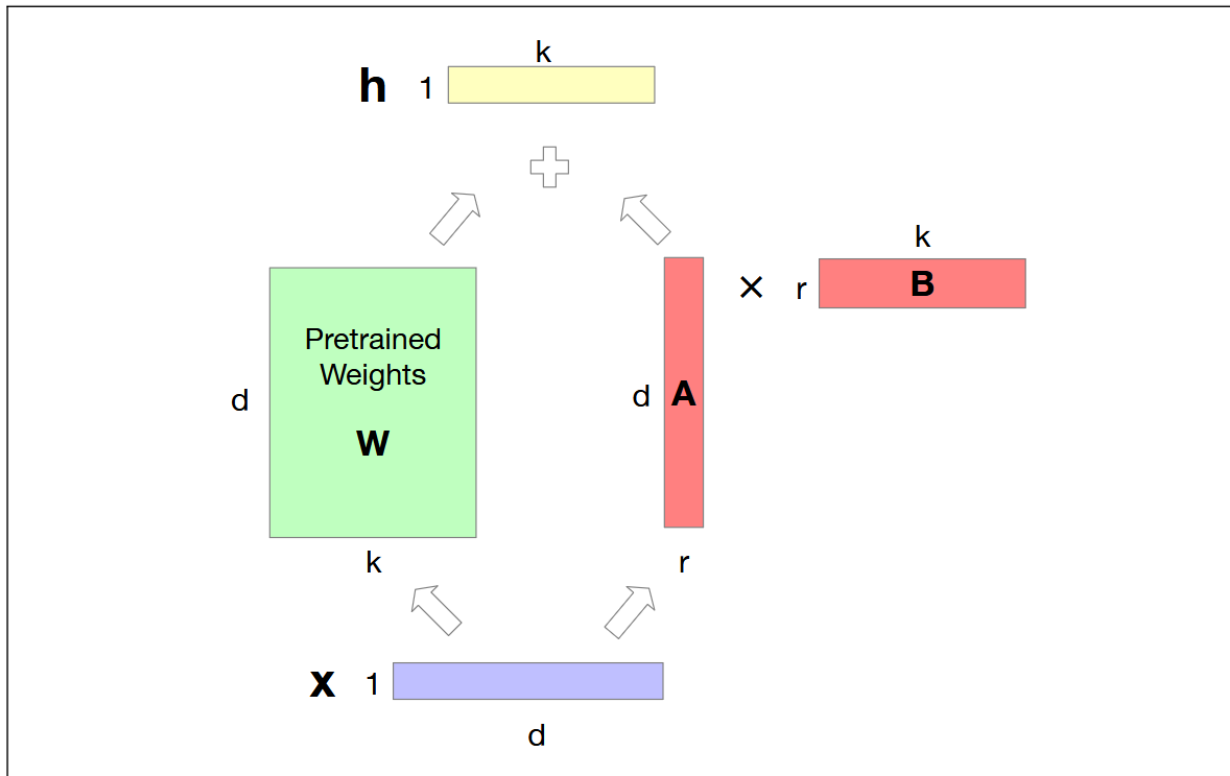


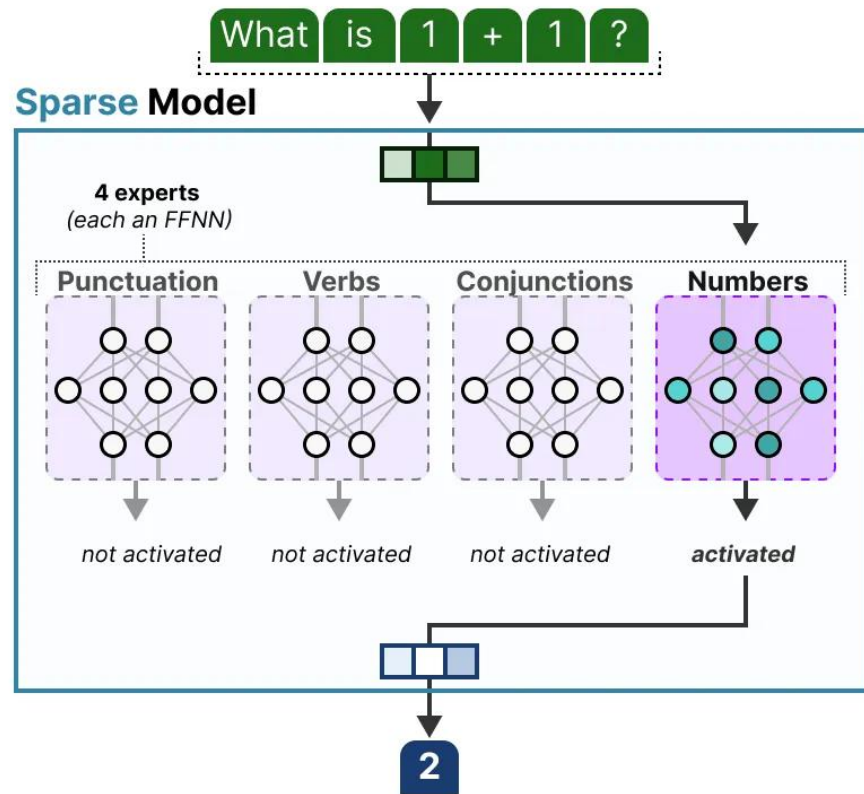
Figure 10.8 The intuition of LoRA. We freeze W to its pretrained values, and instead fine-tune by training a pair of matrices A and B , updating those instead of W , and just sum W and the updated AB .

Train a model using a pretrained LLM but give the new model fewer parameters \rightarrow a low-rank decomposition of the original weight matrix

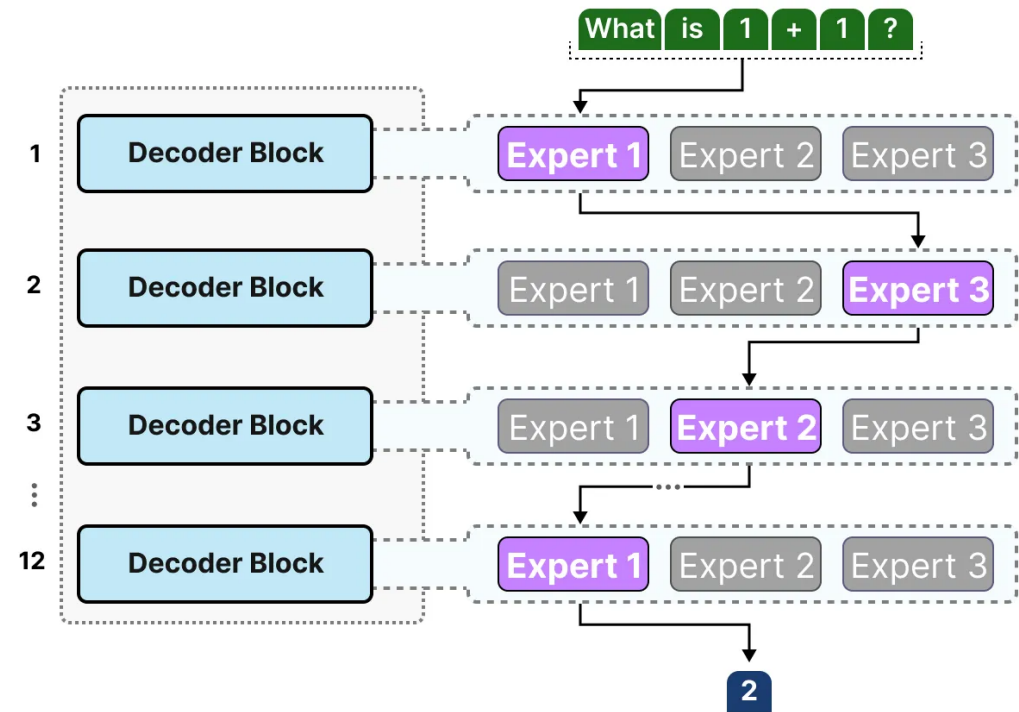
$$h = xW + xAB$$
$$r \ll \min(d, N)$$

From SLP book Chapter 10

Review: Mixture-of-Experts



Most transformers have multiple decoder blocks



Retrieval-Augmented Generation

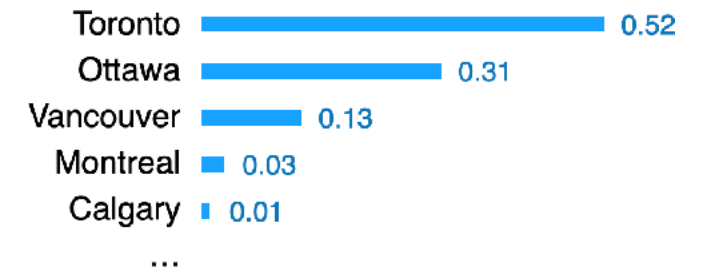
Retrieval-based language models (LMs)

Retrieval-based LMs = Retrieval + LMs

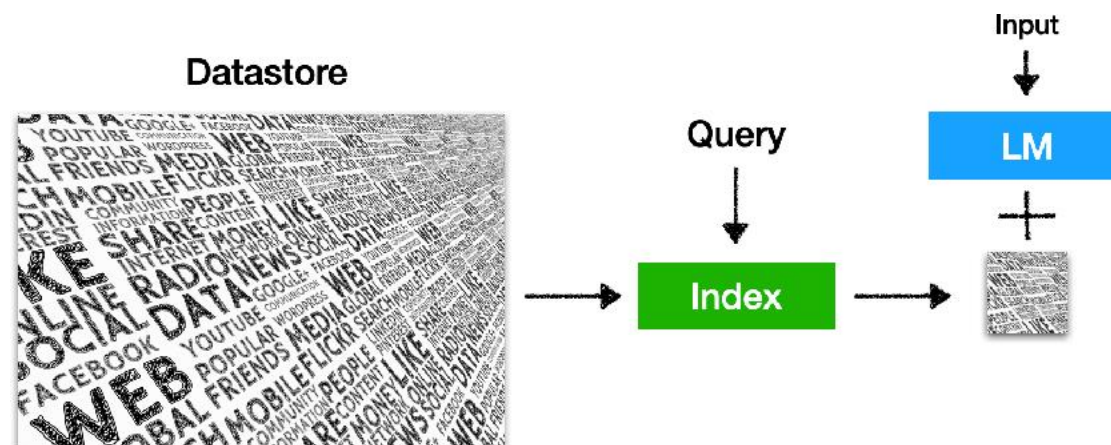
- It is a **language model** $P(x_n | x_1, x_2, \dots, x_{n-1})$

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)



Retrieval for knowledge-intensive NLP tasks

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

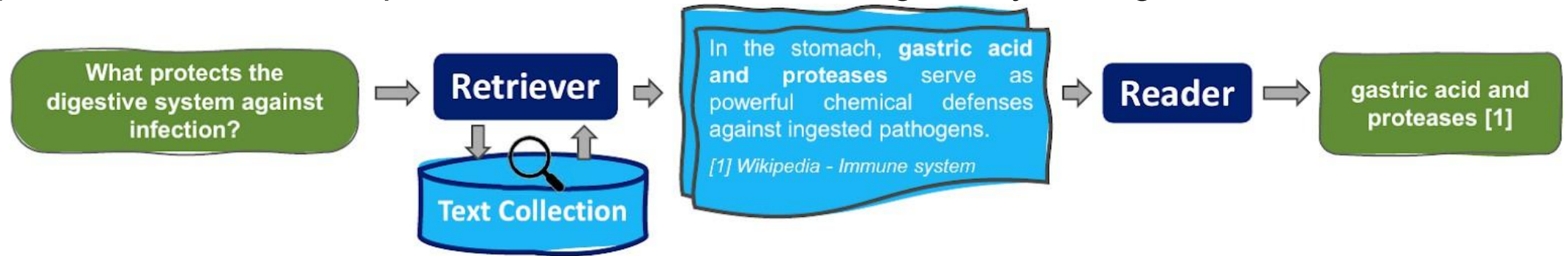


Image: <http://ai.stanford.edu/blog/retrieval-based-NLP/>

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

Why retrieval LMs?

Why retrieval-based LMs?



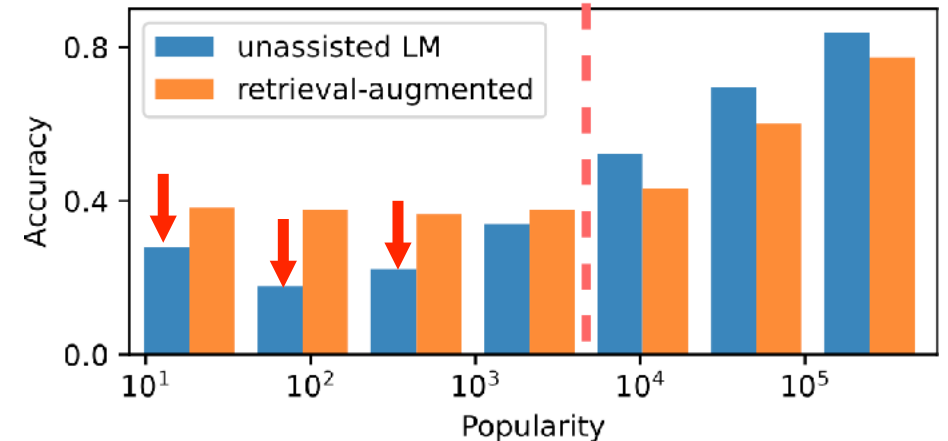
List 5 important papers authored by Geoffrey Hinton



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**, .. ✓
- ...
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?



(Mallen et al., 2023)

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

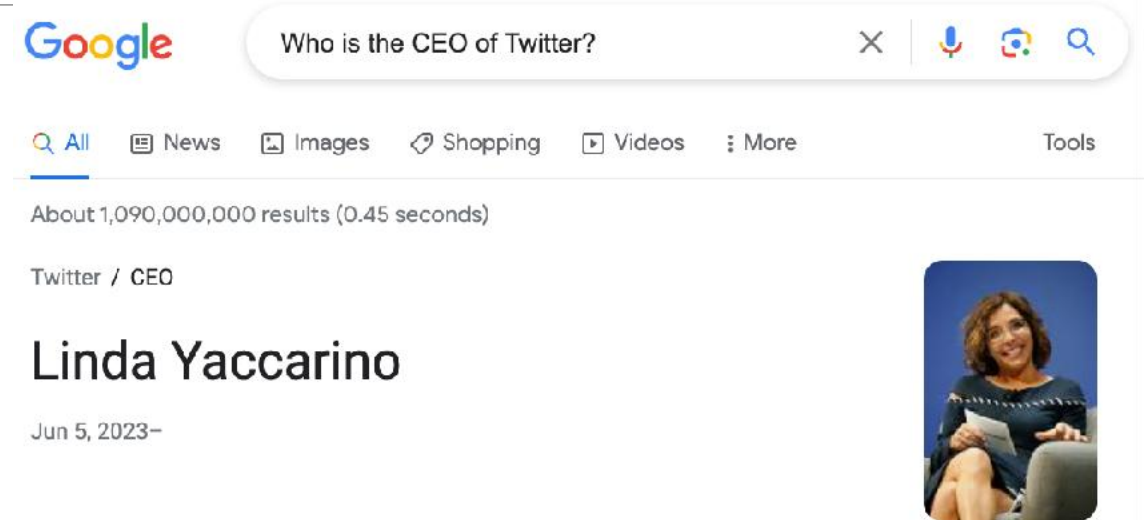
Why retrieval-based LMs?



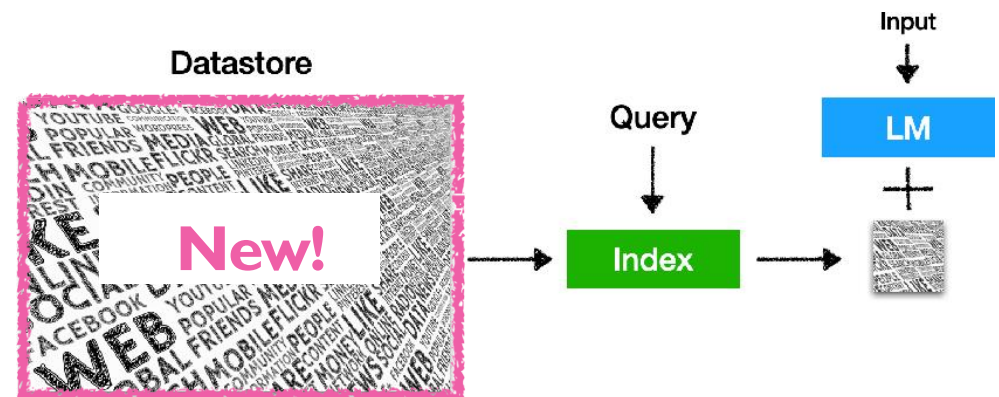
Who is the CEO of Twitter?



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



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



Actually
Google's AI still
gets it wrong...

◆ AI Overview

As of early 2026, the company formerly known as Twitter (now X) has experienced significant leadership changes following its 2022 acquisition by Elon Musk. [Linda Yaccarino](#) was appointed CEO in June 2023 to manage business operations. However, reports indicate Yaccarino stepped down in July 2025, with the platform's ownership shifting to Musk's AI company, xAI. [X +4](#)

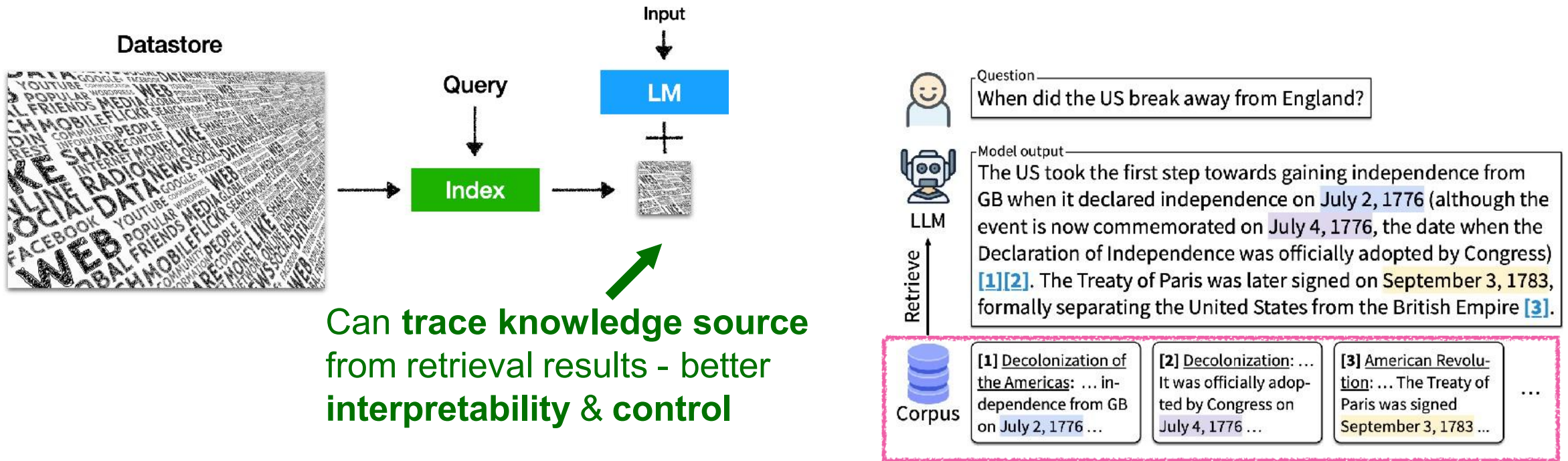
Key Leadership Timeline & Facts:

- **Owner/Executive Chair:** Elon Musk (2022–present), who took the company private and rebranded it to X, often holding CTO/CEO roles.
- **Previous CEO (2023–2025):** Linda Yaccarino, who focused on business and advertising.
- **Previous CEO (2021–2022):** Parag Agrawal, who was terminated upon Musk's acquisition.
- **Founder/Former CEO:** Jack Dorsey, who led the company over two separate terms. [Wikipedia +4](#)

Under Musk, the platform has seen extensive changes, including rebranding, layoffs, and a shift towards AI, with the company (X Corp.) being acquired by xAI in March 2025. [Wikipedia +1](#)

Why retrieval-based LMs?

Generating text with citations



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

Why retrieval-based LMs?

🔍 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 ¹. You can also experience the thrill of walking on the glass floor or the EdgeWalk, the world's highest hands-free walk ¹. 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 ². After lunch, take a stroll along **Queen West**, a trendy neighborhood with eclectic shops, galleries, cafes and street art ³. 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 ⁴. 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 ³ ⁵. 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 ³.

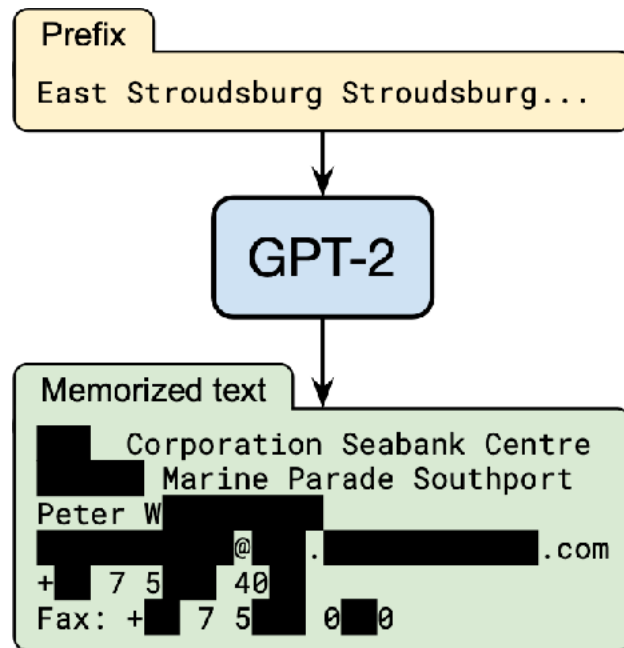
I hope this helps you plan your trip to Toronto. Have fun! 😊

Learn more:

- 1. cntower.ca
- 2. travel.usnews.com
- 3. bing.com
- 4. rom.on.ca
- 5. tripadvisor.com

LLMs are shown to easily leak private training data

Why retrieval-based LMs?

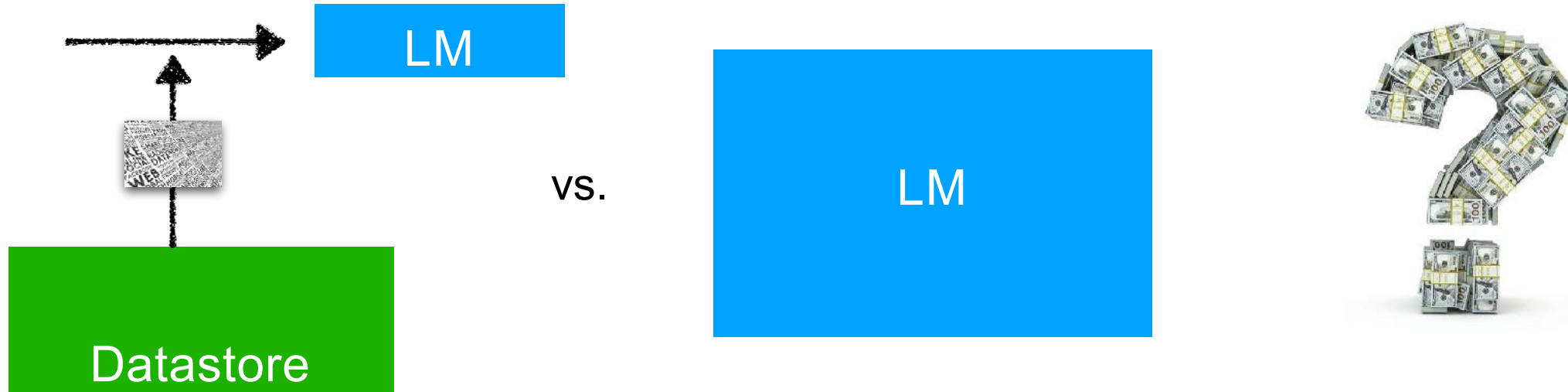


Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	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

LLMs are *large* and expensive to train and run

Why retrieval-based LMs?



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 on the Pile, despite using **25x fewer parameters**”

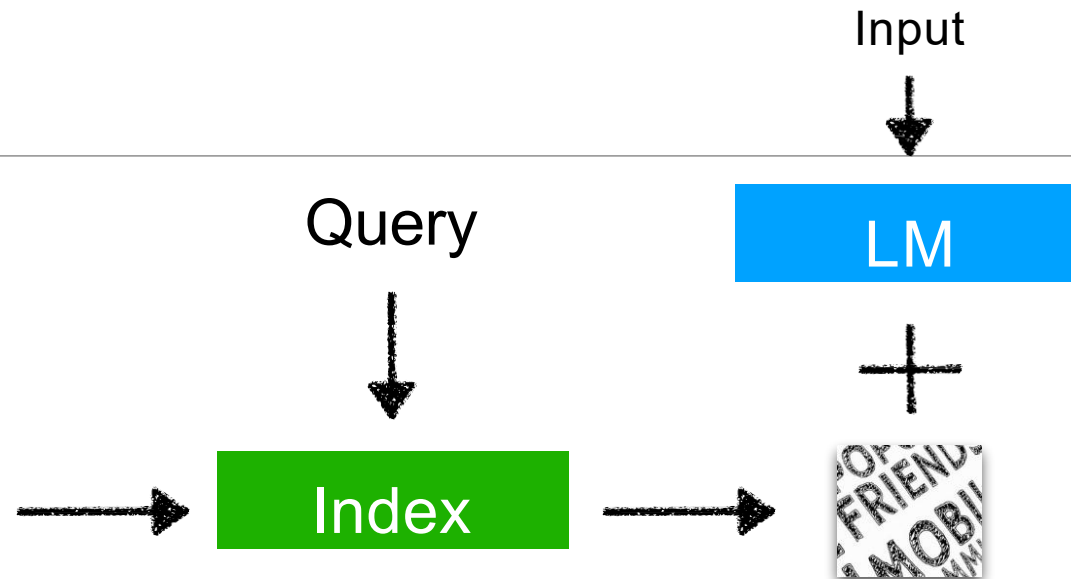
A Retrieval-based LM: Definition

A language model (LM) that uses
an external datastore at test time

Inference



Datastore



Inference: Index

Goal: find a small subset of elements in a datastore that are the most similar to the query

sim: a similarity score between two pieces of text

Inference: Index

Goal: find a small subset of elements in a datastore that are the most similar to the query

sim: a similarity score between two pieces of text

Example $\text{sim}(i, j) = \text{tf}_{i,j} \times \log \frac{N}{\text{df}_i}$

of occurrences of i in j # of total docs # of docs containing i

Example $\text{sim}(i, j) = \text{Encoder}(i) \cdot \text{Encoder}(j)$

Maps the text into an n -dimensional vector

Remember cosine similarity from our discussion of word embeddings

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}$$

Inference: Index

Goal: find a small subset of elements in a datastore that are the most similar to the query

sim: a similarity score between two pieces of text

Can be a totally separate research area on how to do this fast & accurate

Index: given q , return $\text{argTop-}k_{d \in \mathcal{D}} \text{sim}(q, d)$ through fast nearest neighbor search
 k elements from a datastore

Software: FAISS, Distributed FAISS, SCaNN, etc...

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive	Comments
Exact Search for L2	IndexFlatL2	"Flat"	d	4*d	yes	brute-force
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*d	yes	also for cosine (normalize vectors beforehand)
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M	$4*d + x * M * 2 * 4$	no	
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*d + 8	no	Takes another index to assign vectors to inverted lists. The 8 additional bytes are the vector id that needs to be stored.
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbbits/8)	yes	optimized by using random rotation instead of random projections
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes	4 and 6 bits per component are also implemented.
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQM"x"nbbits"	d, M, nbits	ceil(M * nbbits / 8)	yes	
IVF and scalar quantizer	IndexIVFScalarQuantizer	"IVFx,SQ4" "IVFx,SQ8"	quantizer, d, nlists, qtype	SQfp16: 2 * d + 8, SQ8: d + 8 or SQ4: d/2 + 8	no	Same as the IndexScalarQuantizer
IVFADC (coarse quantizer+PQ on residuals)	IndexIVFPQ	"IVFx,PQ"y"x"nbbits"	quantizer, d, nlists, M, nbits	ceil(M * nbbits/8)+8	no	
IVFADC+R (same as IVFADC with re-ranking based on codes)	IndexIVFPQR	"IVFx,PQy+z"	quantizer, d, nlists, M, nbits, M_refine, nbits_refine	M+M_refine+8	no	

Exact Search

Approximate Search
(Relatively easy to scale to ~1B elements)

More info: <https://github.com/facebookresearch/faiss/wiki>

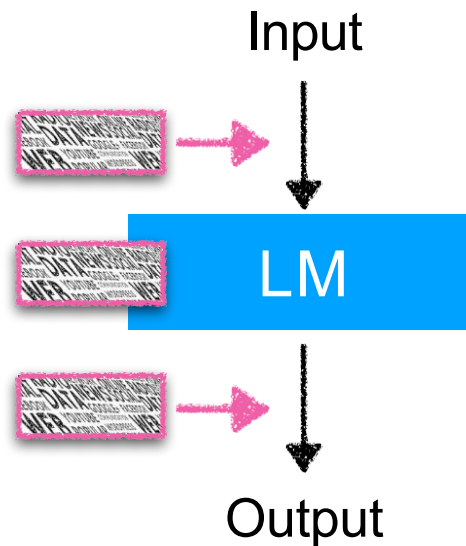
Variations of RAG

What to retrieve?



Text chunks (passages)?
Tokens?
Something else?

How to use retrieval?

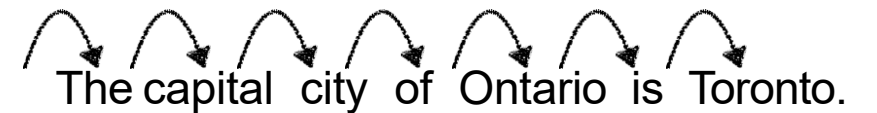


When to retrieve?

w/ retrieval



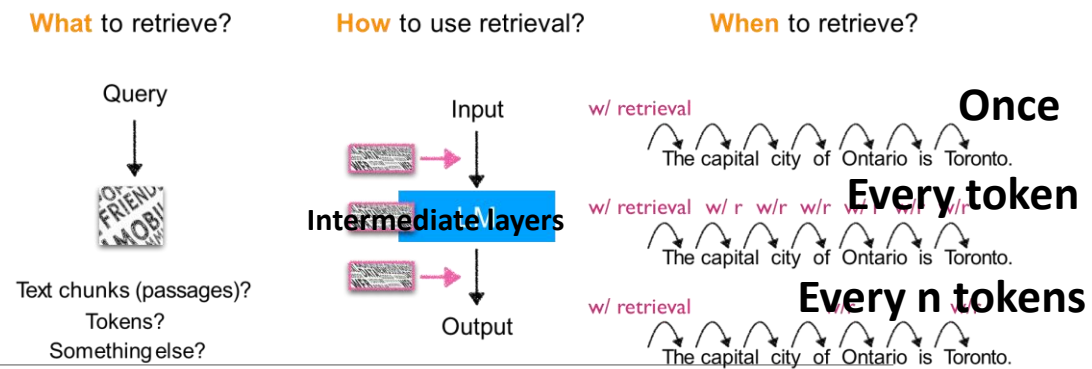
w/ retrieval w/ r w/r w/r w/ r w/r w/r



w/ retrieval



In-Class Activity



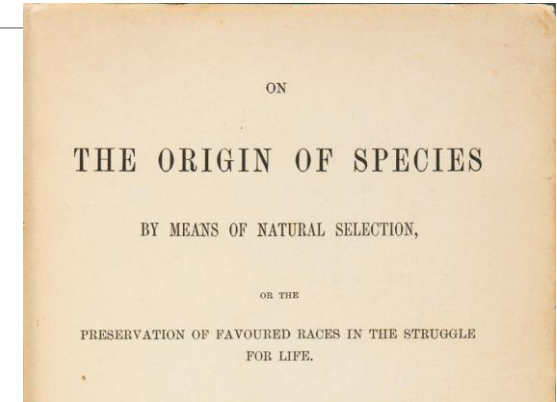
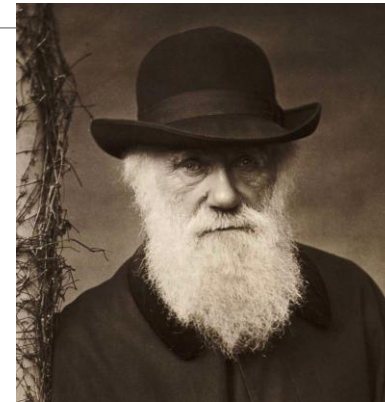
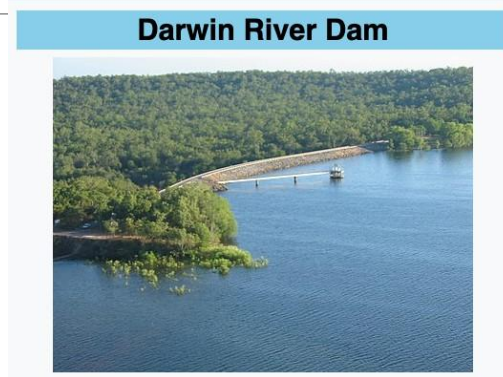
In the group you're in, everyone is assigned a different paper.

Skim your paper to find the answers to these questions:

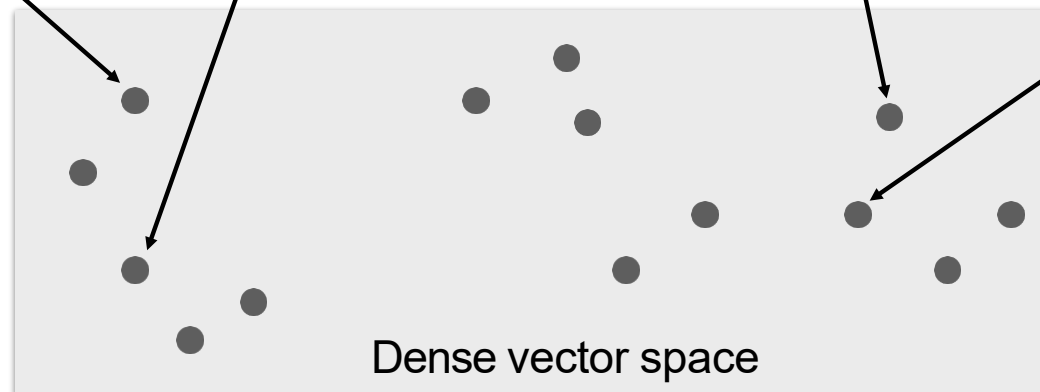
1. What is being retrieved (e.g., passages, tokens, some other level)?
2. How is the retrieval used in relation to the LLM (e.g., as part of the input, integrated into the intermediate layers of the model, added onto the output)?
3. When does the system retrieve (e.g., once, every token, every n tokens)?

Share what you learned with your group and post the summary of *your paper* to Blackboard (along with which paper it was).

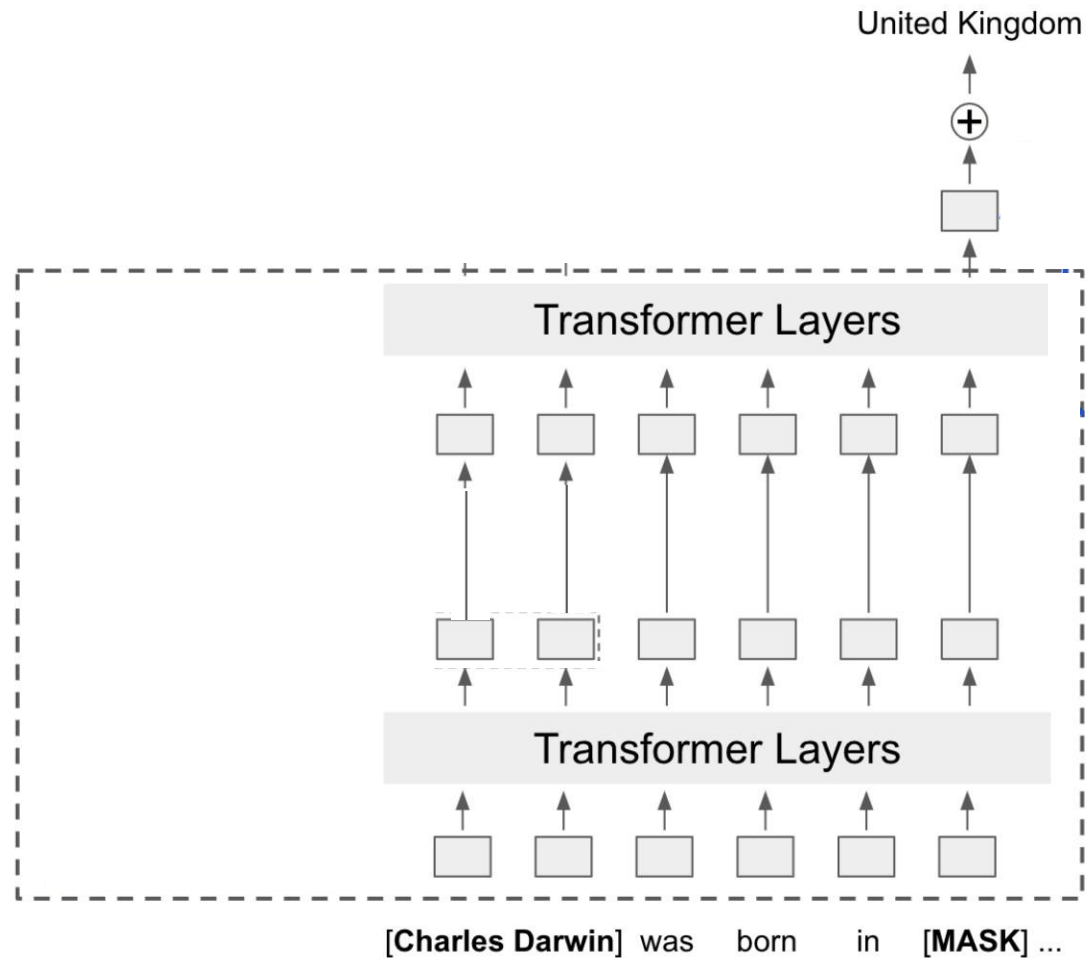
Entities as Experts



"Darwin"



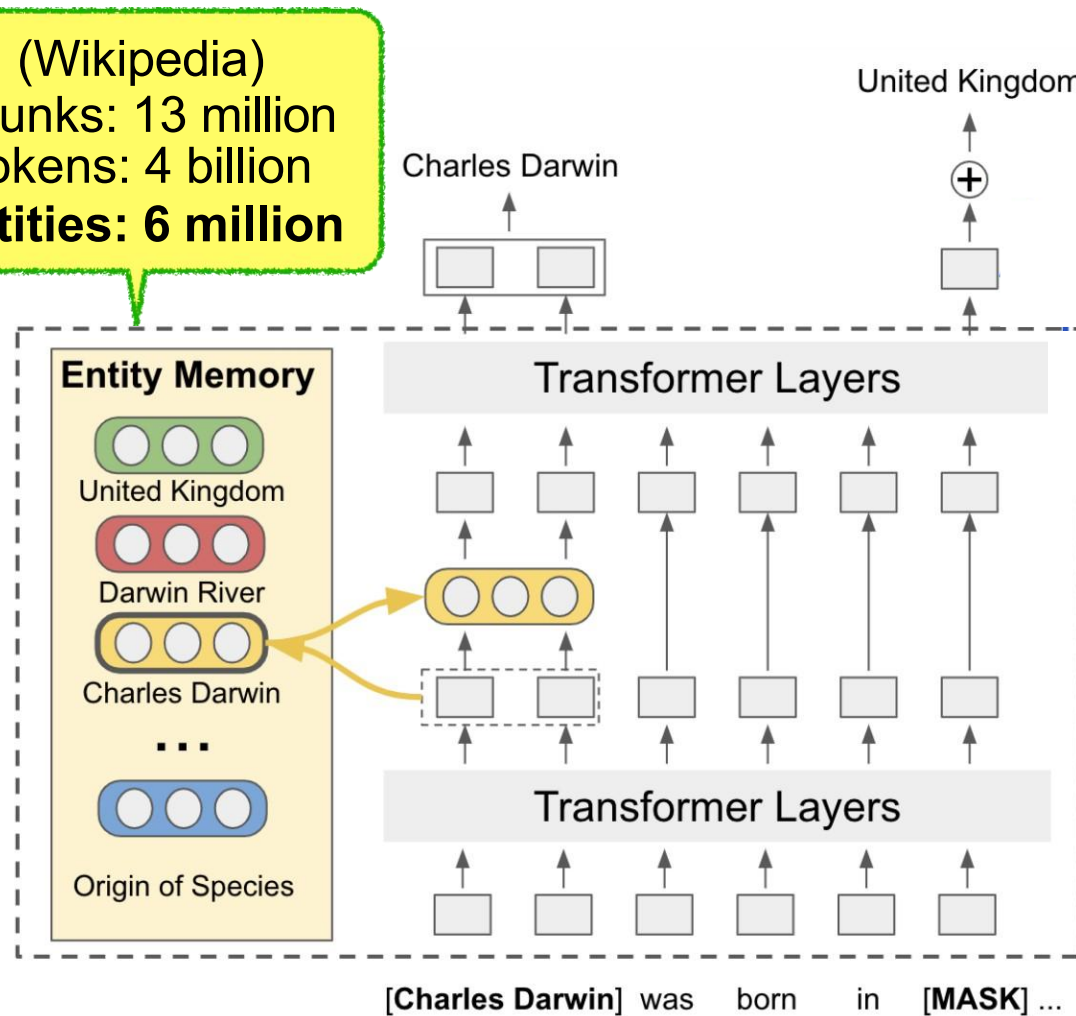
Entities as Experts



Entities as Experts

(Wikipedia)
chunks: 13 million
tokens: 4 billion
entities: 6 million

Need text with
entity detected



Need entity linker to disambiguate mentions

Mention Memory

One vector per entity → One vector per entity *mention*

(Wikipedia)
chunks: 13M
tokens: 4B
entities: 6M

entity mentions: 150M

[Perseus] was a great Greek hero ...
Perseus was a great [Greek] hero ...
...
... [Medusa] was slain by Perseus
... Medusa was slain by [Perseus]
[H Simpson] is a fictional character ...

