Retrieval-based Language Models: An Alternative LM Paradigm

Rulin Shao

CSE599J

Feb. 16, 2023

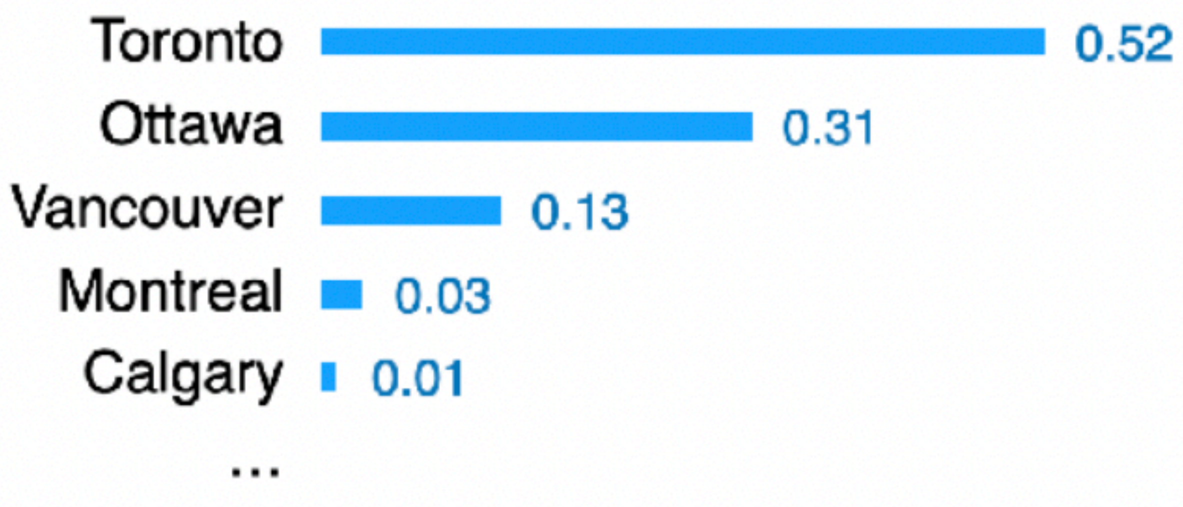
Retrieval-based language models (LMs)

Retrieval-based LMs = Retrieval + LMs

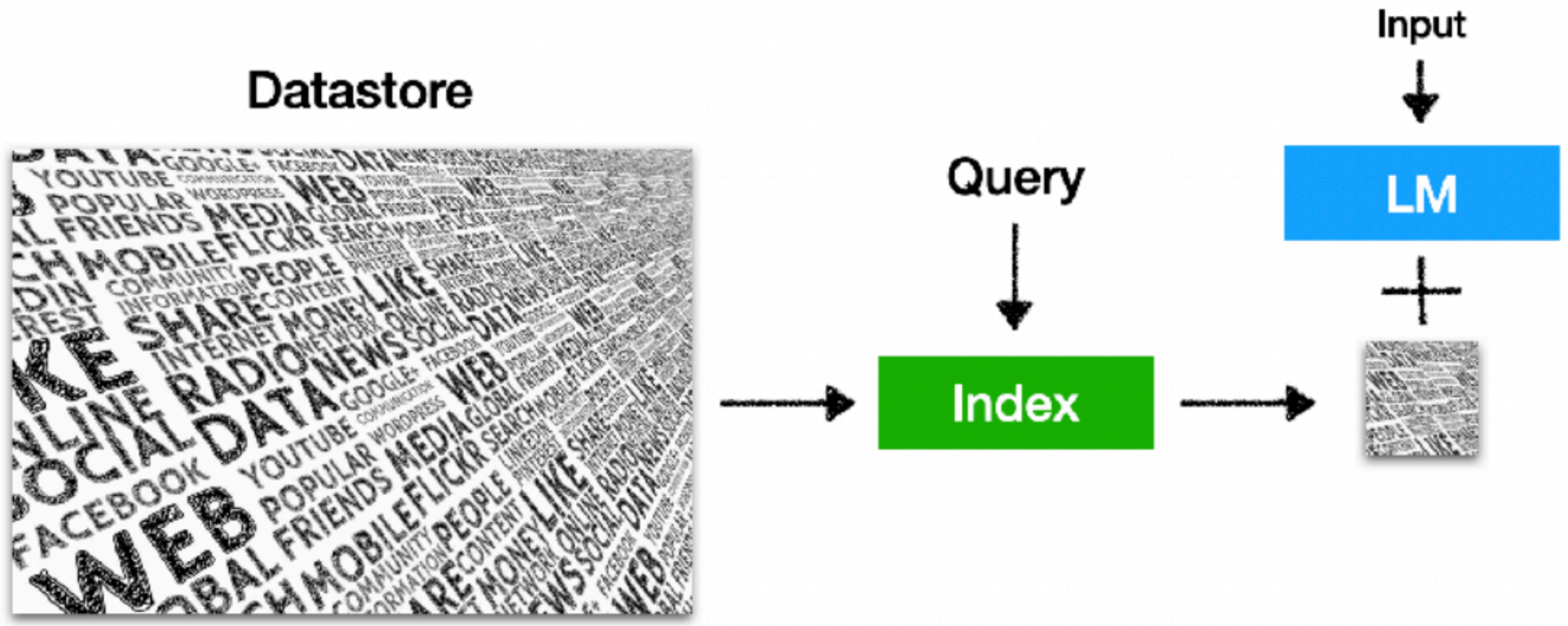
- It is a **language model** $P(x_n \mid x_1, x_2, \cdots, x_{n-1})$

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)



(Also referred to as semiparametric and non-parametric models)

ACL tutorial: https://acl2023-retrieval-lm.github.io/

Motivation

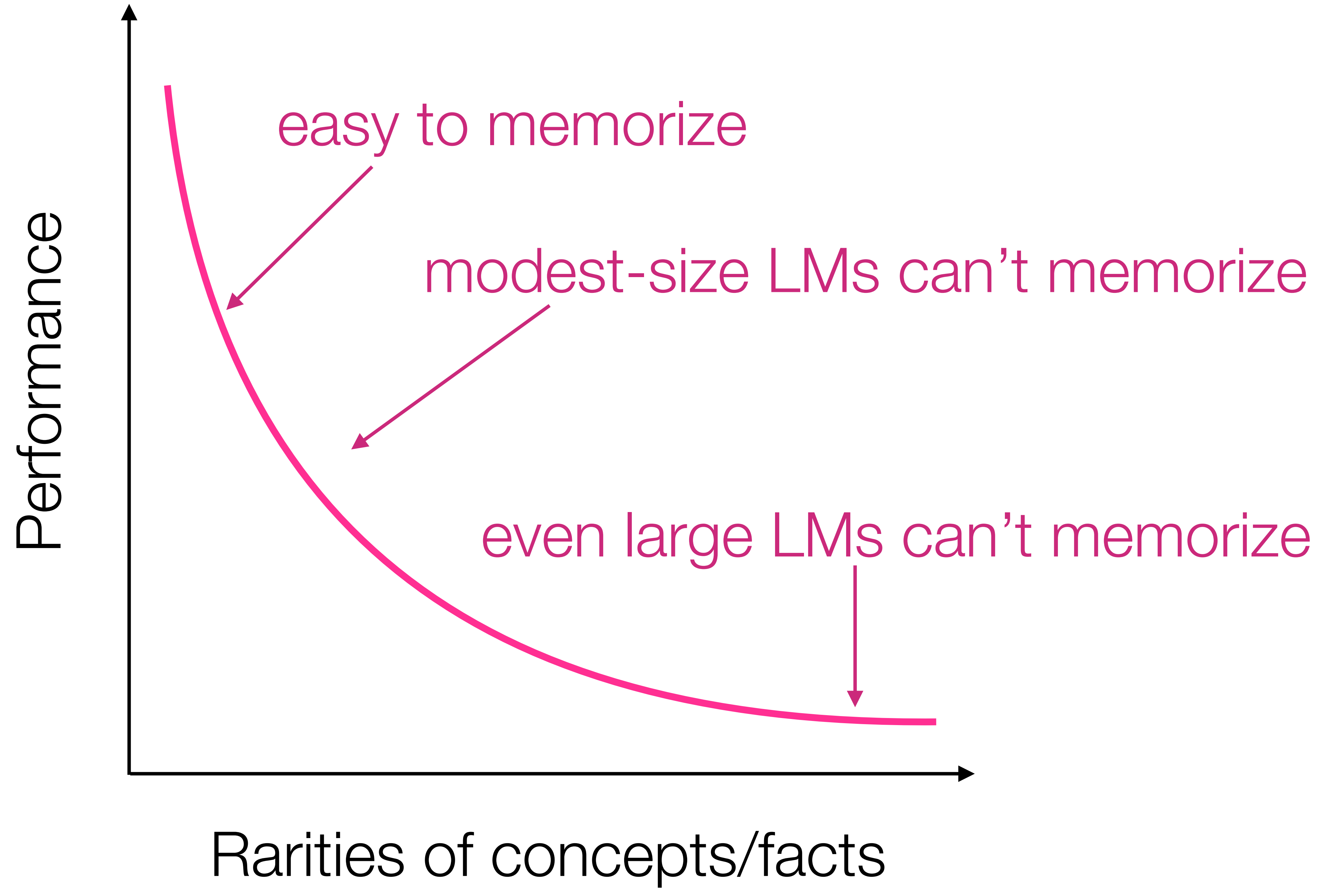
Why retrieval-based LMs?

Better at long-tail concepts & facts	Can grow & update w/o additional training	Provide data attribution

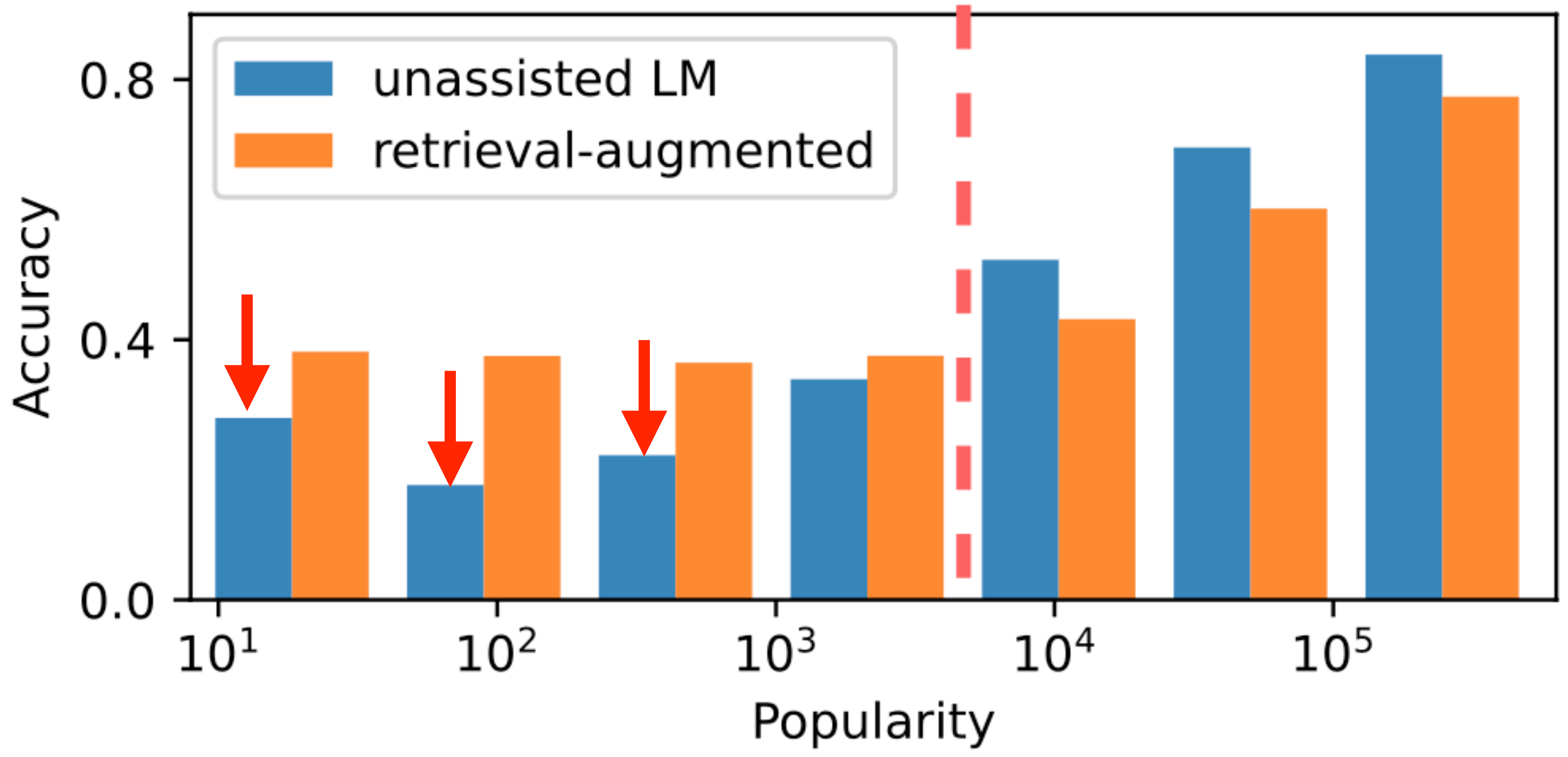
* Slide credit to Sewon Min

Why retrieval-based LMs?

Better at long-tail concepts & facts | Can grow & update w/o additional training | Provide data attribution



What is Kathy Saltzman's occupation?



(Mallen et al., 2023)

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

* Slide credit to Sewon Min

Why retrieval-based LMs?

Better at long-tail concepts & facts

PRESIDENTIAL ELECTIONS 2024

Is United Kingdom
a member of European Union?

Can grow & update w/o additional training

Twitter / CEO

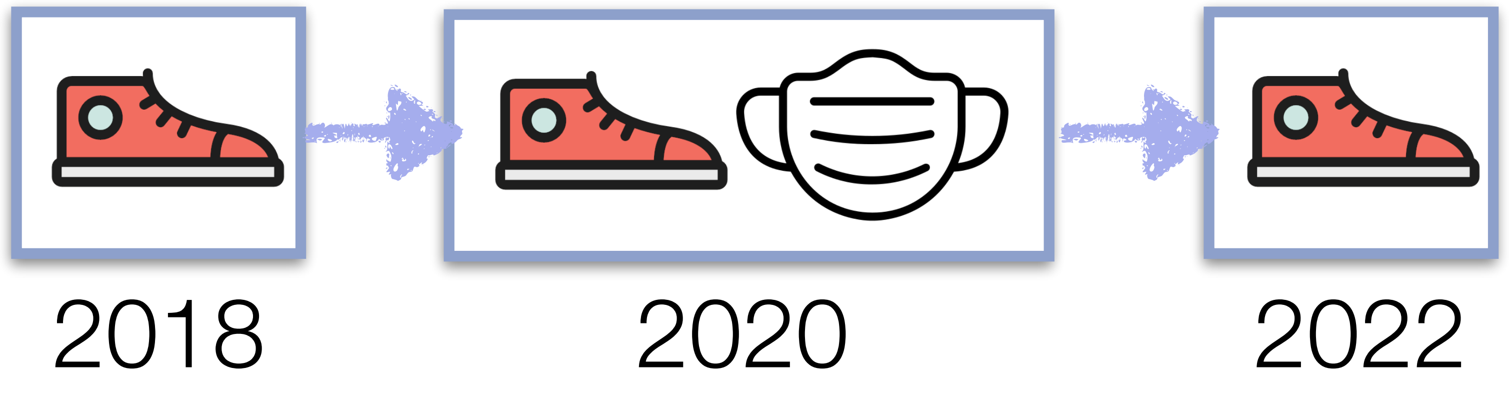
Linda Yaccarino

Jun 5, 2023–

facebook

Meta

Provide data attribution



Often Unpredictable!

* Slide credit to Sewon Min

Why retrieval-based LMs?

Better at long-tail concepts & facts | Can grow & update w/o additional training | **Provide data attribution**

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,…

* Slide credit to Sewon Min

Model Architecture

Model Design Considerations

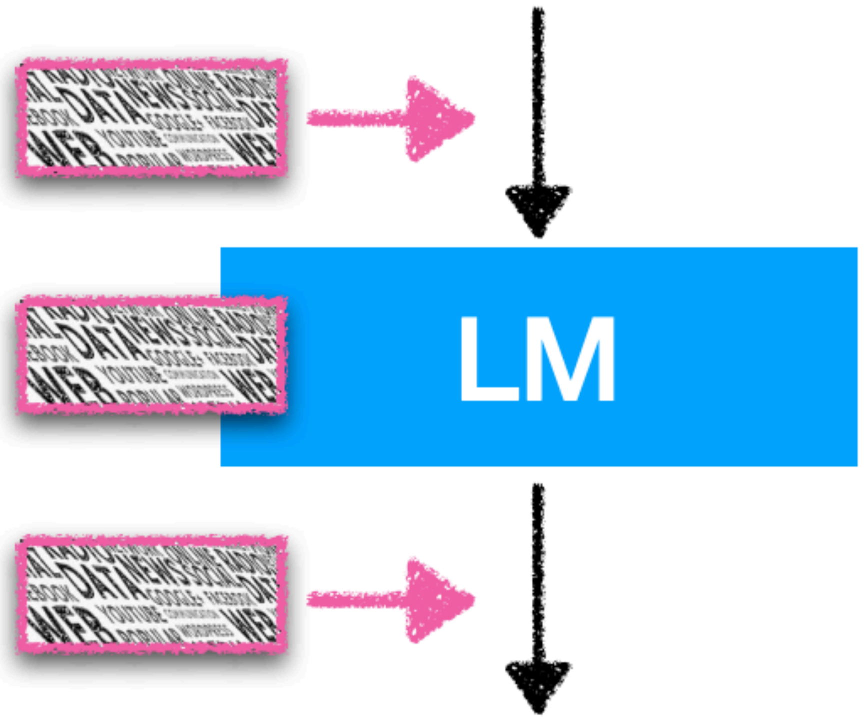
What to retrieve?

Query

Text chunks (passages)?
Tokens?
Something else?

How to use retrieval?

Input



Output

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

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

The capital city of Ontario is Toronto.

w/ retrieval w/r w/r

The capital city of Ontario is Toronto.

Model Design — Retrieve-in-context (RiC) LM

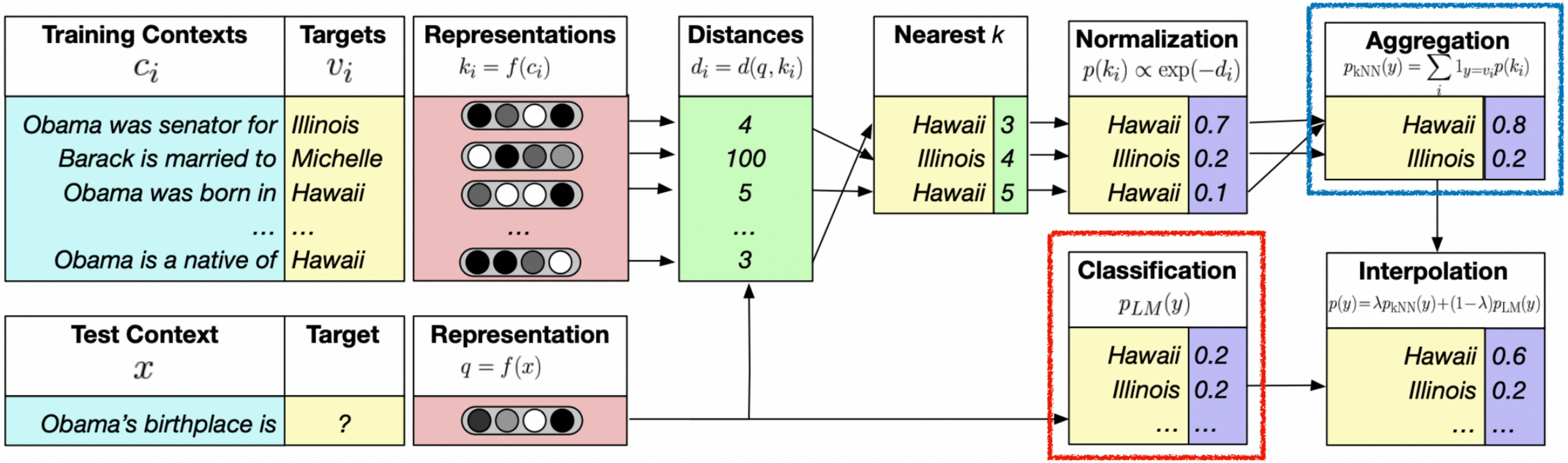
Using the retrieval results as a context

The model architecture used in today's papers.

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Model Design —kNN-LM

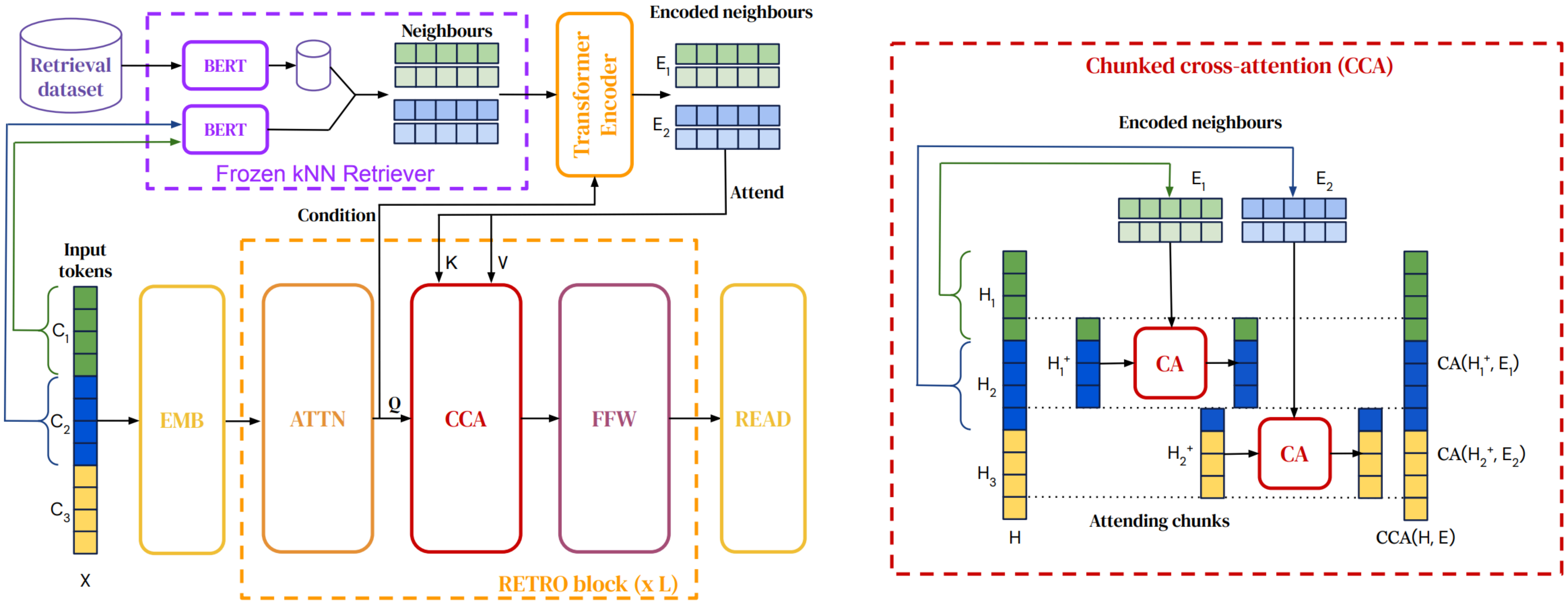
Using the retriever itself as a LM — kNN-LM (Khandelwal et al. 2020)



$$P_{k\text{NN-LM}}(y|x) = (1-\lambda)P_{\text{LM}}(y|x) + \lambda P_{k\text{NN}}(y|x)$$

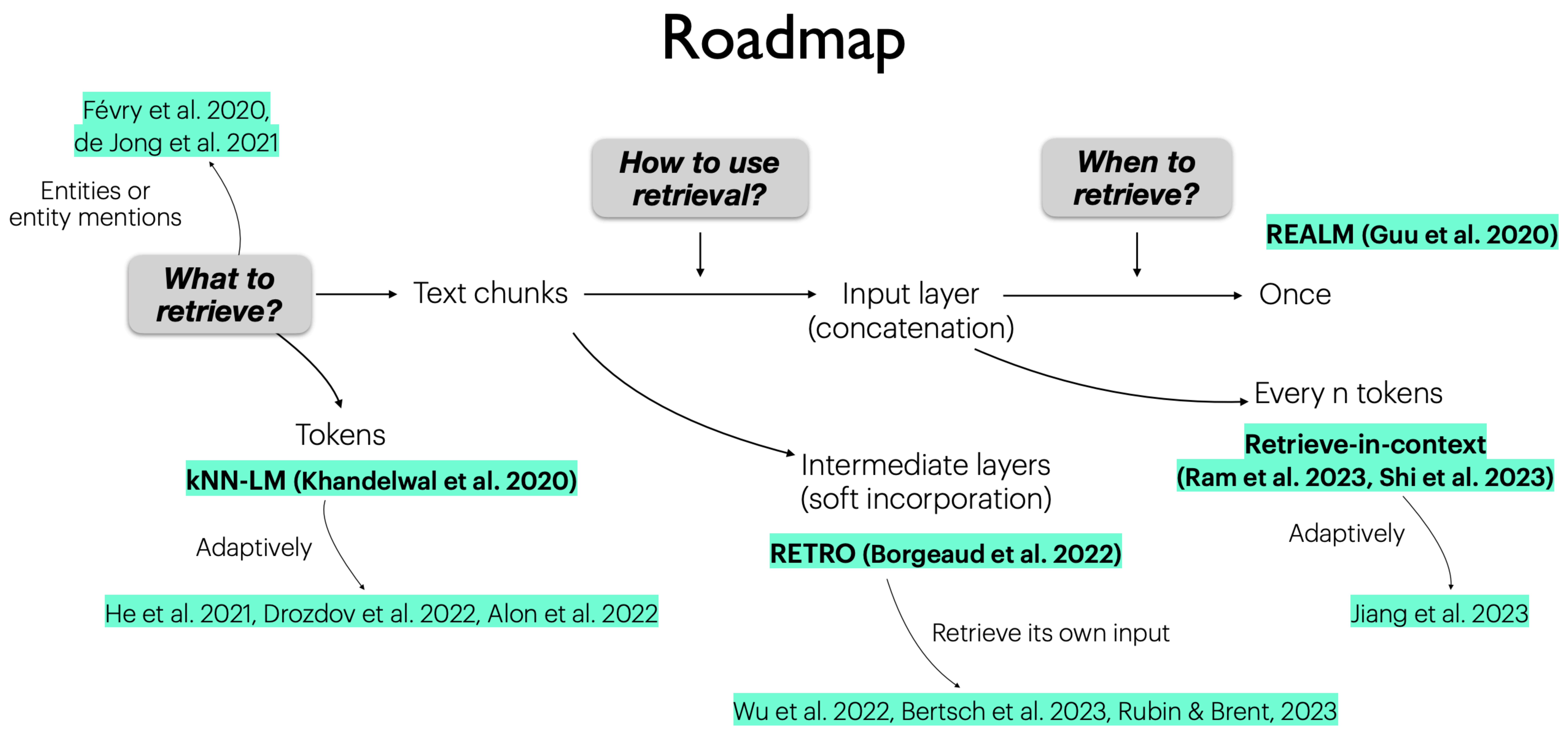
Model Design — RETRO

Feed the retrieval augmentation through cross-attention.



Borgeaud et al. 2022. "Improving language models by retrieving from trillions of tokens"

More Designs ...

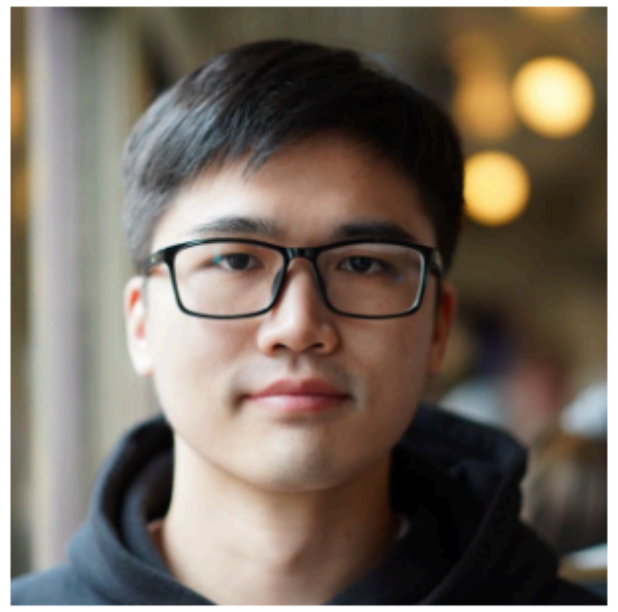
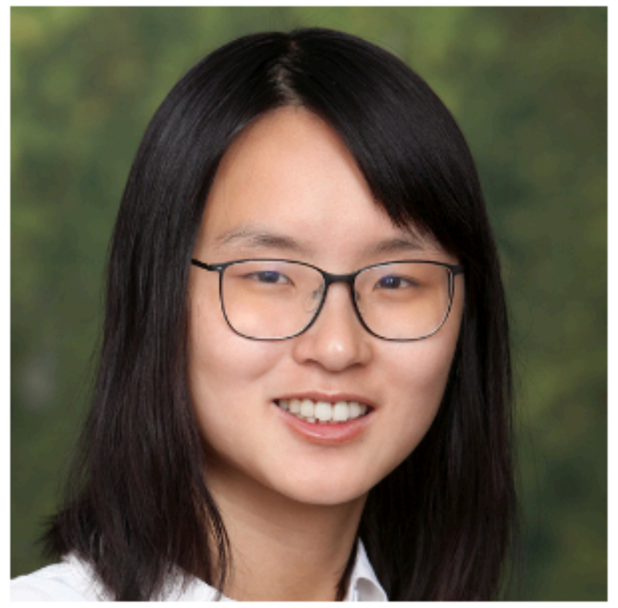


ACL tutorial: https://acl2023-retrieval-lm.github.io/

For more information

ACL 2023 Tutorial:
Retrieval-based Language Models and Applications



Akari Asai[1], Sewon Min[1], Zexuan Zhong[2], Danqi Chen[2]

[1]University of Washington, [2]Princeton University

ACL tutorial: https://acl2023-retrieval-lm.github.io/

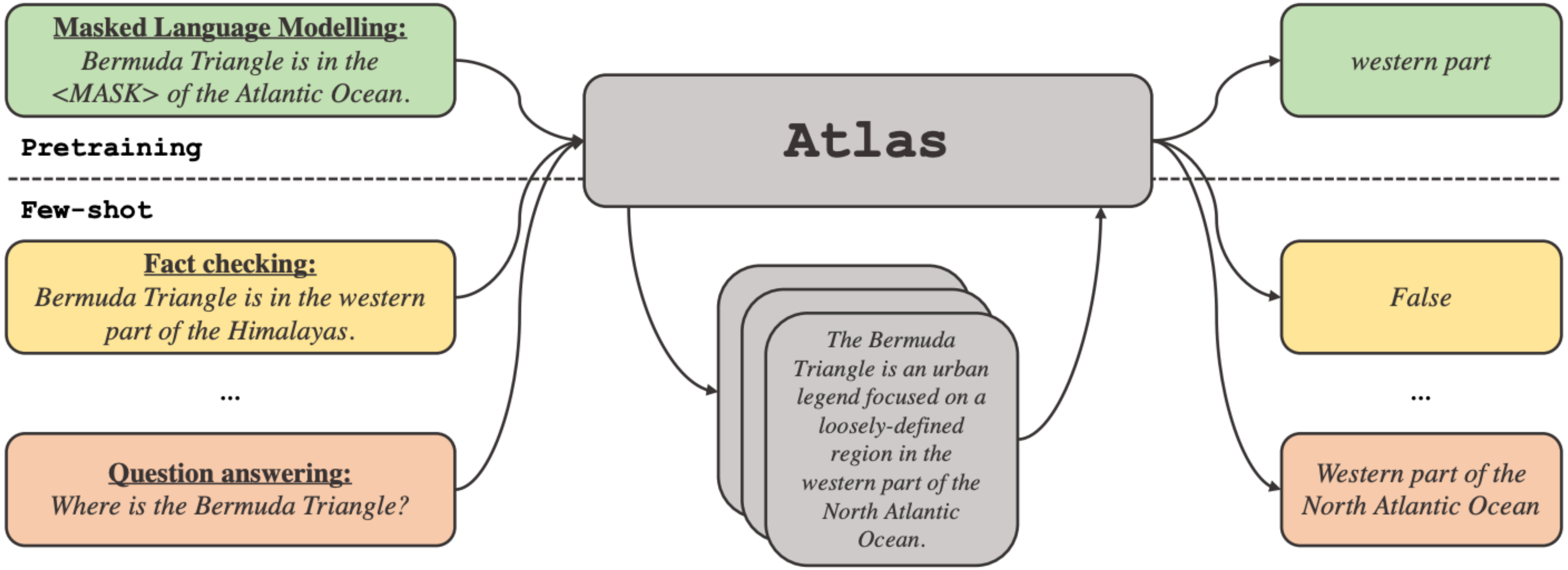
Atlas: Few-shot Learning with Retrieval Augmented Language Models

Gautier Izacard[*,◇,♣,♡] *gizacard@fb.com*
Patrick Lewis[*,◇] *plewis@fb.com*
Maria Lomeli[◇] *marialomeli@fb.com*
Lucas Hosseini[◇] *hoss@fb.com*
Fabio Petroni[◇] *fabiopetroni@fb.com*
Timo Schick[◇] *schick@fb.com*
Jane Dwivedi-Yu[◇] *janeyu@fb.com*
Armand Joulin[◇] *ajoulin@fb.com*
Sebastian Riedel[◇,♠] *sriedel@fb.com*
Edouard Grave[◇] *egrave@fb.com*
[◇] *Meta AI Research,* [♣] *ENS, PSL University,* [♡] *Inria,* [♠] *University College London*

- How to design and **train** retrieval-augmented language models, with a focus on downstream **few-shot learning** and sample efficiency.

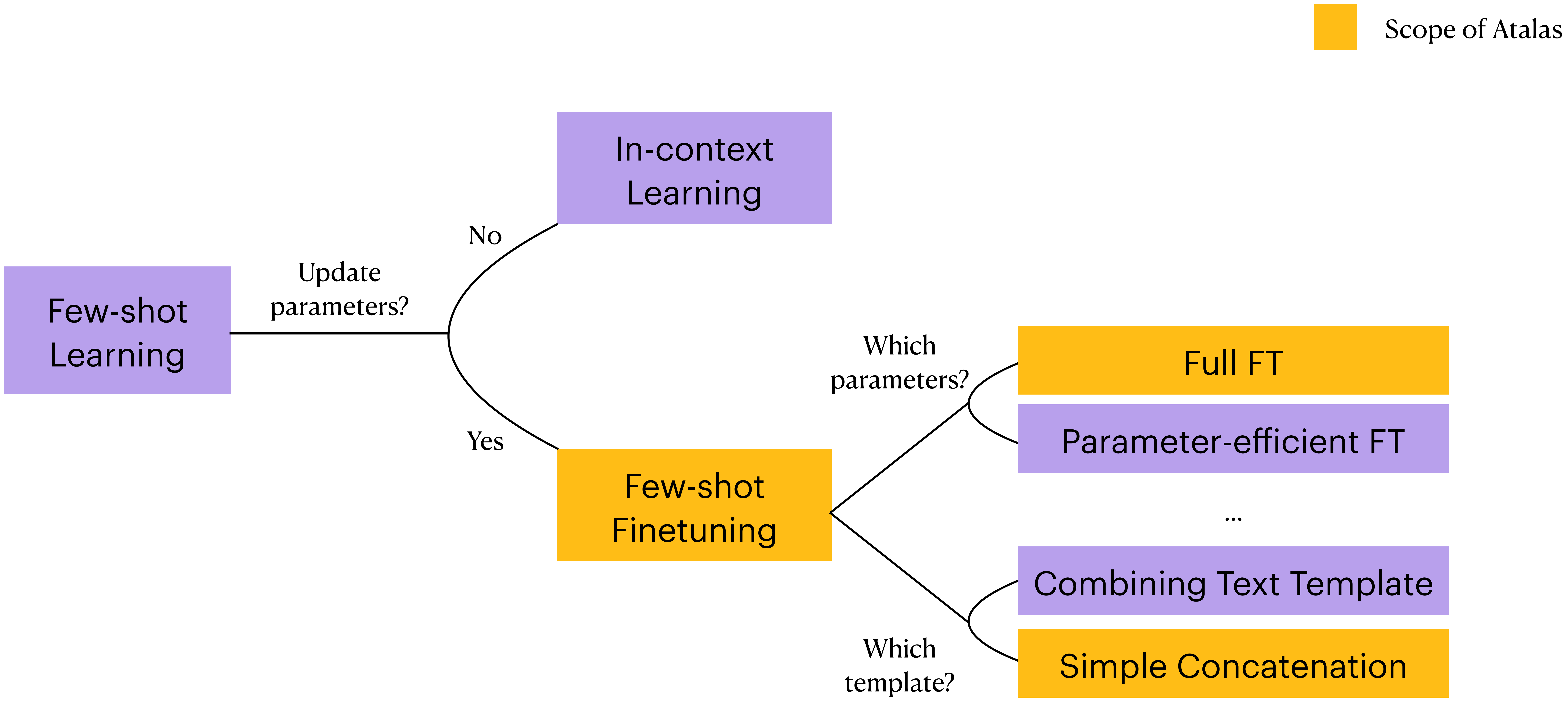
Few-shot Downstream Learning

- The task of learning from very few examples. Specifically, Atlas picks knowledge-intensive ones.



Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Few-shot Downstream Learning



Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

RIC Setting in Atlas

- Retriever: **Contriever (unsupervised pretrained)**
- Language model: T5 unsupervised pretrained model
- Pretraining & datastore data:
  - Dec. 20, 2021 Wikipedia dump (only this for ablation study)
  - 2020-10 common crawl dump

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Performance w/o Training

- Closed-book v.s. Vanilla RIC

		64-shot				1024-shot			
	MLM	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Closed-book	1.083	6.5	14.1	59.0	26.5	10.7	16.5	75.3	34.2
Vanila RIC	-	9.0	14.1	67.0	30.0	9.9	16.6	78.3	34.9

Retrieval augmentation improves the performance of these knowledge intensive tasks!

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

How to further improve the performance with few-shot learning?

Specifically, by co-training the retriever and the LM

- What we have:
  - Input-output pairs from the task of interest.
  - LM.
  - Datastore.
- What is desirable, but we don't have:
  - Annotations on the documentations, e.g., a gold document that contains the supporting fact for each query.

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

How to further improve the performance with few-shot learning?

Specifically, by co-training the retriever and the LM

- What we have:
  - Input-output pairs from the task of interest.
  - LM.
  - Datastore.
- What is desirable but we don't have:
  - Annotations on the documentations, e.g., a gold document that contains the supporting fact for each query.

Q: How to find useful signals to co-train the retriever?

Leverage the language model to provide supervisory signals!

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Training Objectives for the Retriever

- Attention Distillation (ADist)
- End-to-end training of Multi-Document Reader and Retriever (EMDR2)
- Perplexity Distillation (PDist)
- Leave-one-out Perplexity Distillation (LOOP)

Training Objectives for the Retriever

Table 1: **Retriever loss ablation.** We compare different loss functions to pre-train the retriever jointly with the language model. We use the prefix MLM task, and the December 2021 Wikipedia dump for both the index and pre-training data. Fine-tuning is performed with query-side fine-tuning and the loss used for pre-training. Best result is bold, second highest underlined.

		64-shot				1024-shot			
	MLM	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Closed-book	1.083	6.5	14.1	59.0	26.5	10.7	16.5	75.3	34.2
No Joint pre-training	-	9.0	14.1	67.0	30.0	9.9	16.6	78.3	34.9
Fixed retriever	0.823	39.9	14.3	72.4	42.2	45.3	<u>17.9</u>	90.0	<u>51.1</u>
ADist	<u>0.780</u>	40.9	14.4	73.8	43.0	<u>46.2</u>	17.2	**90.9**	**51.4**
EMDR^2	0.783	<u>43.3</u>	<u>14.6</u>	72.1	43.3	44.9	**18.3**	85.7	49.6
PDist	0.783	**45.0**	**15.0**	**77.0**	**45.7**	44.9	<u>17.9</u>	<u>90.2</u>	51.0
LOOP	**0.766**	41.8	**15.0**	<u>74.4</u>	<u>43.7</u>	**47.1**	<u>17.9</u>	87.5	50.8

More helpful with fewer shots

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Pretext Tasks

Used to jointly pre-train the retriever and the language model using only unsupervised data.

- Prefix language
- Masked language modeling
- Title to section generation

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Pretext Tasks

Used to jointly pre-train the retriever and the language model using only unsupervised data.

		64-shot				1024-shot			
	MLM	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Closed-book	1.083	6.5	14.1	59.0	26.5	10.7	16.5	75.3	34.2
No Joint pre-training	-	9.0	14.1	67.0	30.0	9.9	16.6	78.3	34.9
Fixed retriever	0.823	39.9	14.3	72.4	42.2	45.3	<u>17.9</u>	90.0	<u>51.1</u>

Huge improvement

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Pretext Tasks

Used to jointly pre-train the retriever and the language model using only unsupervised data.

		64-shot				1024-shot			
	MLM	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Closed-book	1.083	6.5	14.1	59.0	26.5	10.7	16.5	75.3	34.2
No Joint pre-training	-	9.0	14.1	67.0	30.0	9.9	16.6	78.3	34.9
Fixed retriever	0.823	39.9	14.3	72.4	42.2	45.3	<u>17.9</u>	90.0	<u>51.1</u>

Table 2: **Pretext task ablation.** We compare different pretext tasks, used to jointly pre-train our models. Examples are randomly sampled from the training set of the KILT version of the dataset. We report the exact match on NaturalQuestions, the F1 score on Wizard of Wikipedia and the accuracy on FEVER.

	64-shot				1024-shot			
	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Prefix Language Modelling	41.0	14.5	64.9	40.1	**44.7**	17.9	86.0	49.5
Masked Language Modelling	**42.7**	**14.9**	**69.7**	**42.4**	**44.7**	**18.3**	**88.8**	**50.6**
Title-to-section generation	41.1	15.2	66.1	40.8	45.4	17.9	84.6	49.3

Similarly helpful

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Efficient Retriever Fine-tuning

- Full index update (expensive)
- Re-ranking
- Query-side fine-tuning

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Efficient Retriever Fine-tuning

Table 4: **Retriever fine-tuning ablation.** Here, we compare different strategies to fine-tune the retriever in a few-shot setting.

	64-shot				1024-shot			
	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Standard fine-tuning	44.3	14.9	73.2	44.1	47.0	18.4	89.7	**51.7**
Top-100 re-ranking	44.2	14.6	75.4	**44.7**	**47.1**	**18.7**	88.9	51.6
Query-side fine-tuning	**45.0**	**15.0**	**77.0**	**45.7**	44.9	17.9	**90.2**	51.0
Fixed retriever	36.8	14.5	72.0	41.1	38.0	17.7	89.3	48.3

Both re-ranking and query-side fine-tuning preserve or even improve performance though drastically reduce computation

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Analysis

Beating larger models with retrieval augmentations.

Table 5: **Performance on MMLU as a function of model size.**

	5-shot			5-shot (multi-task)			Full / Transfer		
	770M	3B	11B	770M	3B	11B	770M	3B	11B
Closed-book T5	29.2	35.7	36.1	26.5	40.0	43.5	42.4	50.4	54.0
ATLAS	38.9	42.3	43.4	42.1	48.7	56.4	56.3	59.9	65.8
Δ	+9.8	+6.6	+7.3	+15.6	+8.7	+12.9	+13.9	+9.5	+11.8

Small model outperforms larger models with retrieval augmentation!

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Analysis

Beating larger models with retrieval augmentations.

Table 7: **Comparison to state-of-the-art on MMLU.** *For the 5-shot setting, ATLAS uses fine-tuning, while previous works use in-context learning. The ATLAS model uses de-biased inference. Train FLOPS refers to total the amount of computation necessary to train the model, including pre-training and/or fine-tuning.

Setting	Model	Params	Train FLOPS	All	Hum.	Soc. Sci.	STEM	Other
zero-shot	ATLAS	11B	3.5e22	47.1	43.6	54.1	38.0	54.4
5-shot	GPT-3	175B	3.1e23	43.9	40.8	50.4	36.7	48.8
	Gopher	280B	5.0e23	60.0	56.2	71.9	47.4	66.1
	Chinchilla	70B	5.0e23	**67.5**	**63.6**	**79.3**	**55.0**	73.9
	ATLAS*	11B	3.5e22	47.9	46.1	54.6	38.8	52.8
5-shot (multi-task)	ATLAS	11B	3.5e22	56.6	50.1	66.4	46.4	66.2
Full / Transfer	UnifiedQA	11B	3.3e22	48.9	45.6	56.6	40.2	54.6
	GPT-3	175B	3.1e23	53.9	52.5	63.9	41.4	57.9
	ATLAS	11B	3.5e22	66.0	61.1	77.2	53.2	**74.4**

Small model outperforms larger models with retrieval augmentation!

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Analysis

Impact of inference biases

Table 6: **Standard vs de-biased inference for MMLU** These results are reported for ATLAS-11B, using cyclic permutations for de-biasing, which increases inference costs by a factor of $4\times$.

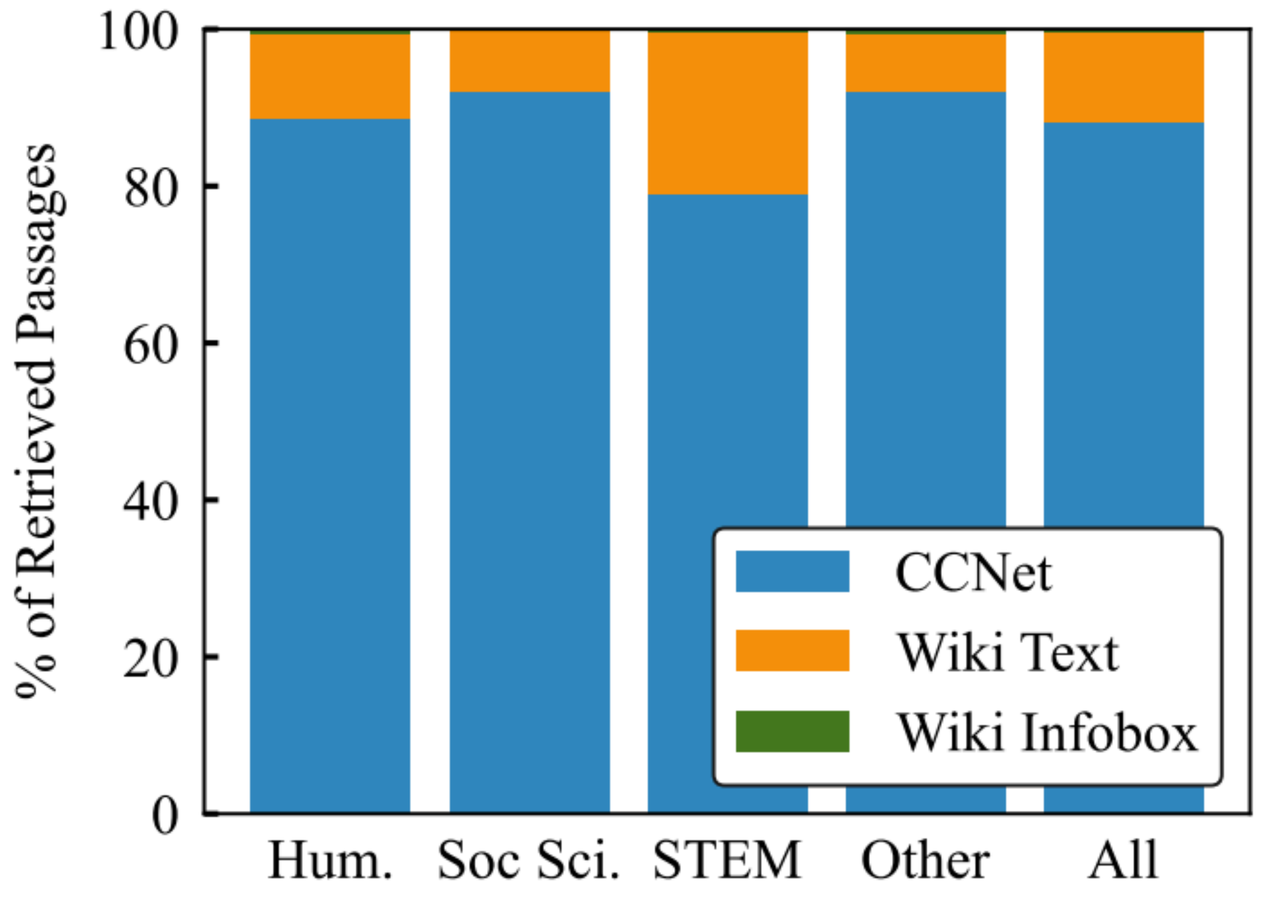
	Zero-shot	5-shot	5-shot (multi-task)	Full / Transfer
Standard Inference	36.8	43.4	56.4	65.8
De-biased Inference	47.1	47.9	56.6	66.0

De-biasing works effectively. With more training samples, the need decreases.

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Analysis

Composition of retrieved documents



Wikipedia makes up about 15% of retrieved passages, though it only makes up about 10% of index.

Retriever tends to retrieve from more relevant and higher-quality data.

Izacard et al. 2022, “Atlas: Few-shot Learning with Retrieval Augmented Language Models”

Analysis

Temporal sensitivity and updatability

Table 11: **Results on our TempLAMA-derived dataset.** We report performance for a static, closed-book T5-11B, as well as ATLAS-11B supplied with a test-time Wikipedia index from 2017 or 2020. We evaluate models finetuned on a small training set of 248 time-sensitive cloze-question-answer pairs, using answers either from 2017 or 2020. Good models should score highly when the test set year matches the year of the test-time index, and score low otherwise.

Temporally mismatched train set leads to worse closed-book performance.

		2017 Test Set Acc.		2020 Test Set Acc.	
Train Set	Test-time Index	Closed-book	ATLAS	Closed-book	ATLAS
2017 answers	2017	12.1	57.7	2.9	1.5
	2020	12.1	10.2	2.9	53.1
2020 answers	2017	4.8	50.1	3.6	4.2
	2020	4.8	3.5	3.6	60.5

Temporally mismatched index leads to inferior performance!

Table 12: **Impact of index data temporality on NaturalQuestions.** We report exact match performance on NaturalQuestions using different Wikipedia dumps in the index. We observe that the dump from December 2018, commonly used for NaturalQuestions, leads to the best result.

	Dec. 2017	Dec. 2018	Aug. 2019	Dec. 2020	Dec. 2021
64-shot	44.7	**45.1**	44.1	44.0	41.3
Full	63.2	**64.0**	62.4	61.1	59.6

Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

References

- ACL tutorial: https://acl2023-retrieval-lm.github.io/
- Ram, Ori, et al. "In-context retrieval-augmented language models."
- Shi, Weijia, et al. "Replug: Retrieval-augmented black-box language models."
- Khandelwal, Urvashi, et al. "Generalization through memorization: Nearest neighbor language models."
- Borgeaud et al. 2022. "Improving language models by retrieving from trillions of tokens"
- Izacard, Gautier, et al. "Unsupervised dense information retrieval with contrastive learning."
- Izacard et al. 2022, "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

🔍

Retrieval-based Language Models: Where They Help & Where They Don’t

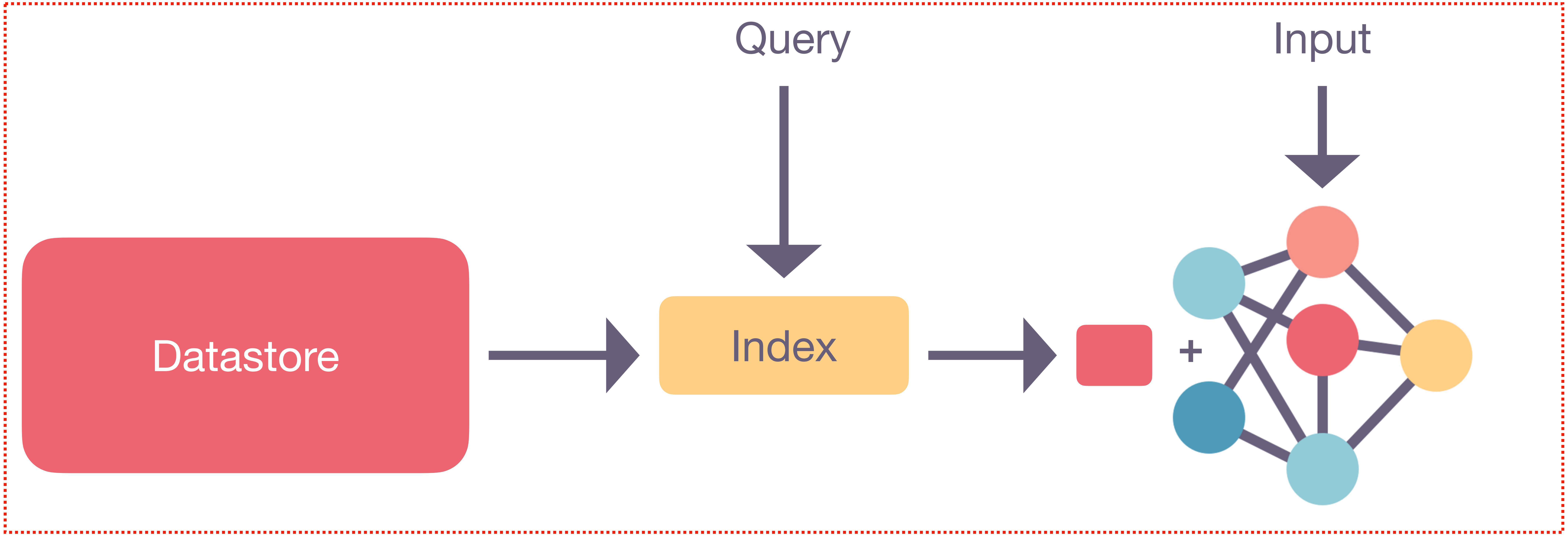
Jacqueline He

CSE599J

Feb. 16, 2023

Retrieval-based LMs

- Any **parametric** language model that queries from an external **non-parametric** datastore during inference time

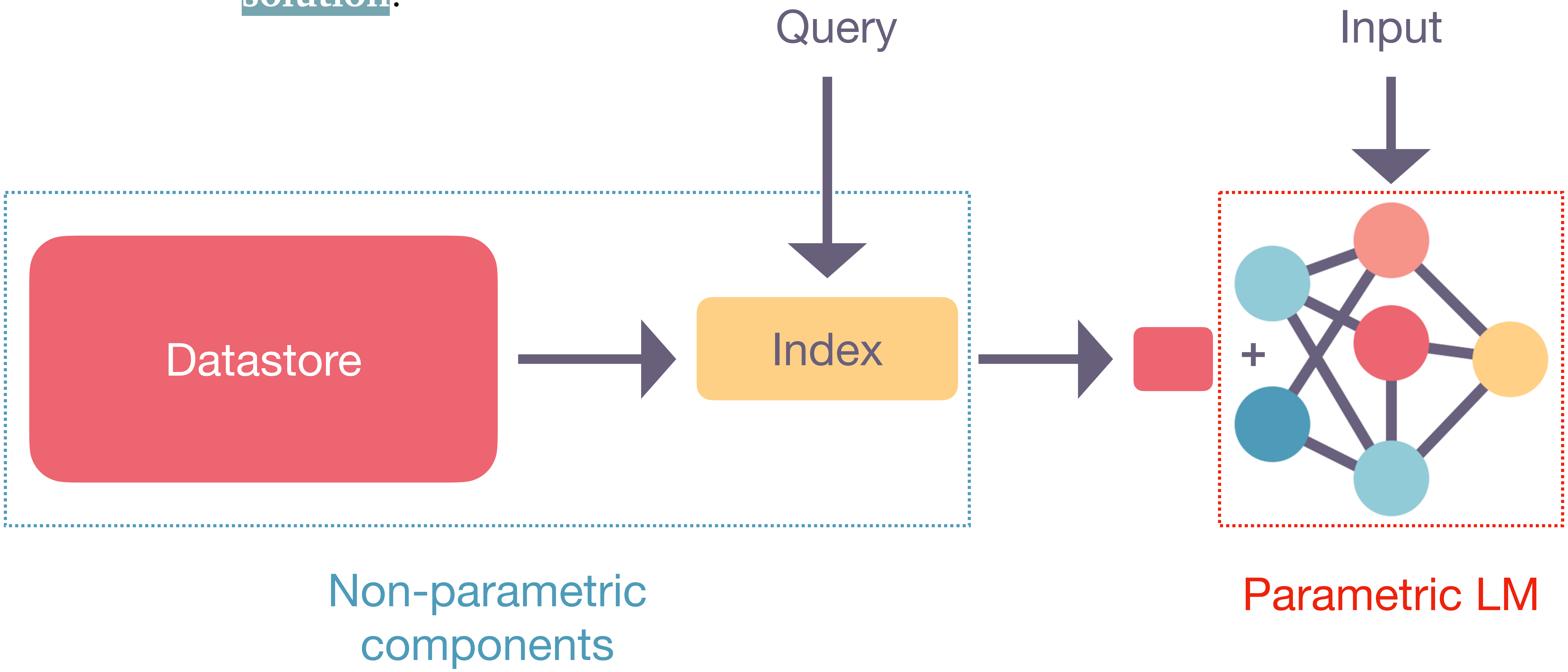


There is no unified retrieval-based LM architecture: NN-based LMs, retrieve-in-context LMs, etc.

But the underlying groundwork is pretty much the same!

Retrieval-based LMs

- Retrieval-based language models confer many advantages over parametric language models…but they are not a silver bullet solution!



When *Not* to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories

Alex Mallen[*◇] **Akari Asai**[*◇] **Victor Zhong**[◇] **Rajarshi Das**[◇]
Hannaneh Hajishirzi[◇♡] **Daniel Khashabi**[♠]
[◇]University of Washington [♠]Johns Hopkins University
[♡]Allen Institute for AI
{atmallen,akari,vzhong,rajarshi,hannaneh}@cs.washington.edu
danielk@jhu.edu

Parametric LMs encode factual knowledge, but not equally

You
Who is Donna Tartt?

ChatGPT
Donna Tartt is an American author known for her novels, particularly her debut, "The Secret History," which was published in 1992. Born on December 23, 1963, in Greenwood, Mississippi, Tartt gained widespread acclaim with "The Secret History," a psychological thriller set in an elite college and centered around a group of students studying ancient Greek. The novel explores themes of obsession, morality, and the consequences of intellectual pursuits.

Following the success of her first novel, Donna Tartt published "The Little Friend" in 2002, a Southern Gothic mystery set in Mississippi. In 2013, she released "The Goldfinch," which won the Pulitzer Prize for Fiction in 2014. "The Goldfinch" follows the life of a young boy who survives a traumatic event and becomes entangled in the world of art theft.

Donna Tartt is known for her meticulous writing style and her ability to create richly detailed and immersive narratives. Her novels often delve into complex psychological and philosophical themes.

GPT 3.5, accessed Feb. 2024

Donna Tartt

43 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Donna Louise Tartt (born December 23, 1963)[2] is an American novelist and essayist. Her novels are *The Secret History* (1992), *The Little Friend* (2002), and *The Goldfinch* (2013), which has been adapted into a 2019 film of the same name[3] She was included in *Time* magazine's 2014 "100 Most Influential People" list.[4]

Early life [edit]

Tartt was born in Greenwood, Mississippi, in the Mississippi Delta, the elder of two daughters. She was raised in the nearby town of Grenada. Her father, Don Tartt, was a rockabilly musician, turned freeway "service station owner-cum-local politician", while her mother, Taylor, was a secretary.[5][6][7] Her parents were avid readers, and her mother would read while driving.[8]

> I know a ton of poetry by heart, When I was a little kid, first thing I memorized were really long poems by A. A. Milne ... I also know all these things that I was made to learn. I'm sort of this horrible repository of doggerel verse.[5]

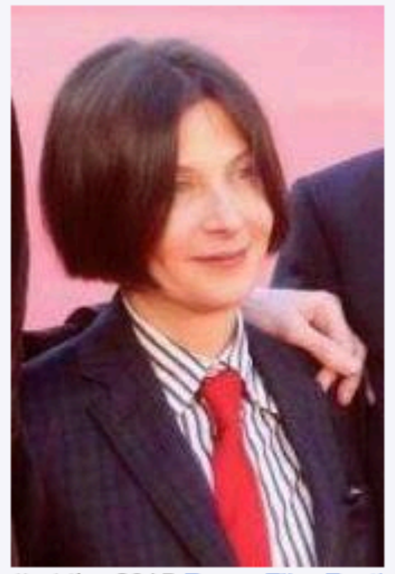
In 1968, aged five, Tartt wrote her first poem.[9] In 1976, aged thirteen, she was published for the first time when a sonnet was included in the *Mississippi Review*.[5][10] In high school, she was a freshman cheerleader for the basketball team and worked in the public library.[6][11][12]

In 1981, Tartt enrolled in the University of Mississippi where her writing caught the attention of Willie Morris while she was a freshman. Finding her in the Holiday Inn bar one evening, Morris said to her, "My name is Willie Morris, and I think you're a genius."[9][13][14][15][16]

Following a recommendation from Morris, Barry Hannah, then an Ole Miss writer-in-residence, admitted the eighteen-year-old Tartt into his graduate course on the short story. "She was deeply literary", said Hannah. "Just a rare genius, really. A literary star."[17]

In 1982, following the suggestion of Morris and others, she transferred to Bennington College. At Bennington, Tartt studied classics with Claude Fredericks, and also met Bret Easton Ellis, Jonathan Lethem, and Jill Eisenstadt.[18][2] Tartt graduated in 1986.[19]

Donna Tartt



Tartt at the 2015 Rome Film Festival

Born	December 23, 1963 (age 60) Greenwood, Mississippi, U.S.
Occupation	Fiction writer
Alma mater	Bennington College
Period	1992–present
Literary movement	Neo-romanticism
Notable works	*The Secret History* (1992) *The Little Friend* (2002) *The Goldfinch* (2013)
Notable awards	WH Smith Literary Award 2003 *The Little Friend* Pulitzer Prize for Fiction 2014 *The Goldfinch* Andrew Carnegie Medal for Excellence in Fiction 2014 *The Goldfinch*

"Donna Tartt", Wikipedia, accessed Feb. 2024

- Language models **memorize** factual knowledge…

Parametric LMs encode factual knowledge, but not equally

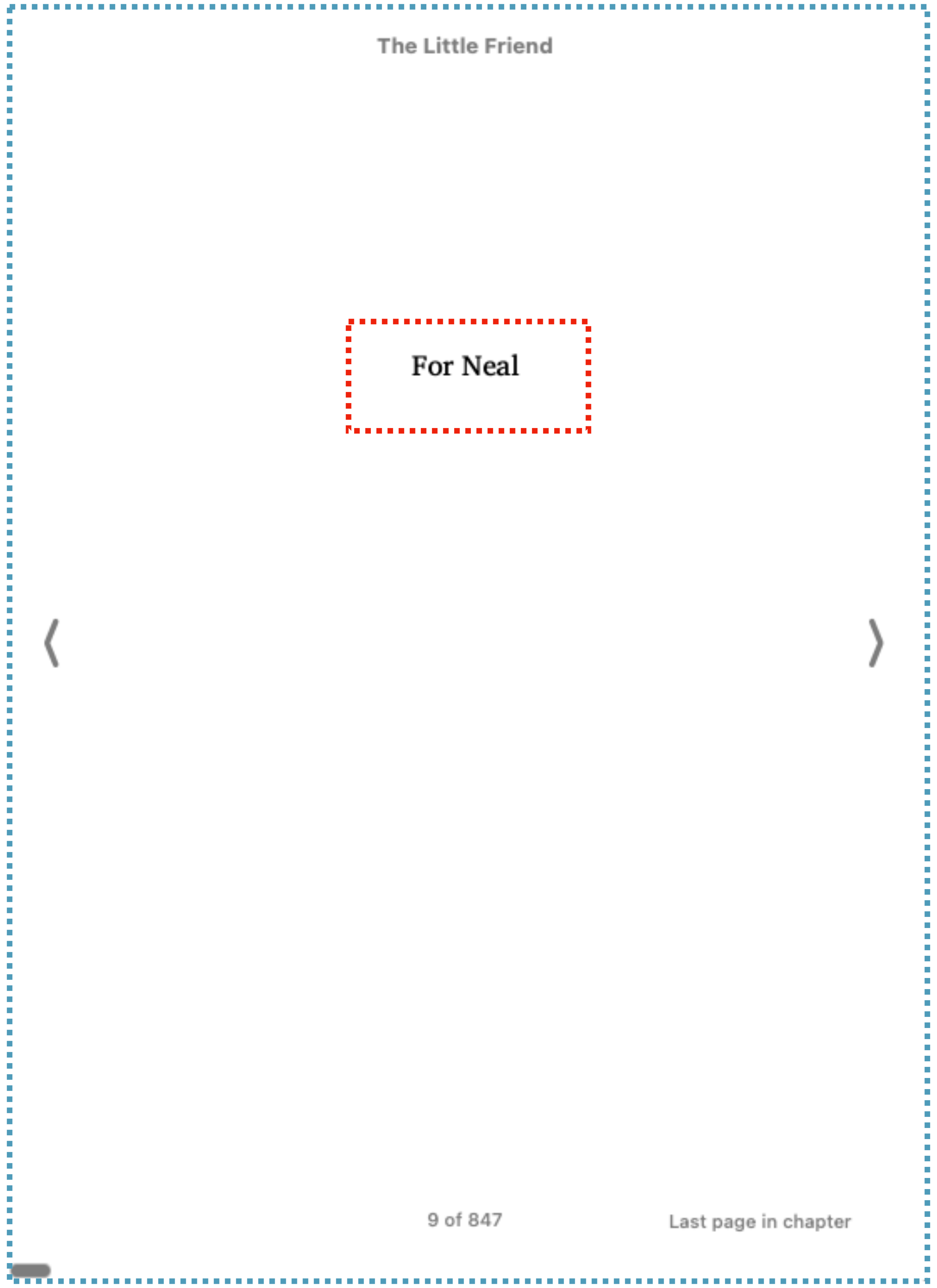
You
Who did Donna Tartt dedicate The Little Friend to?

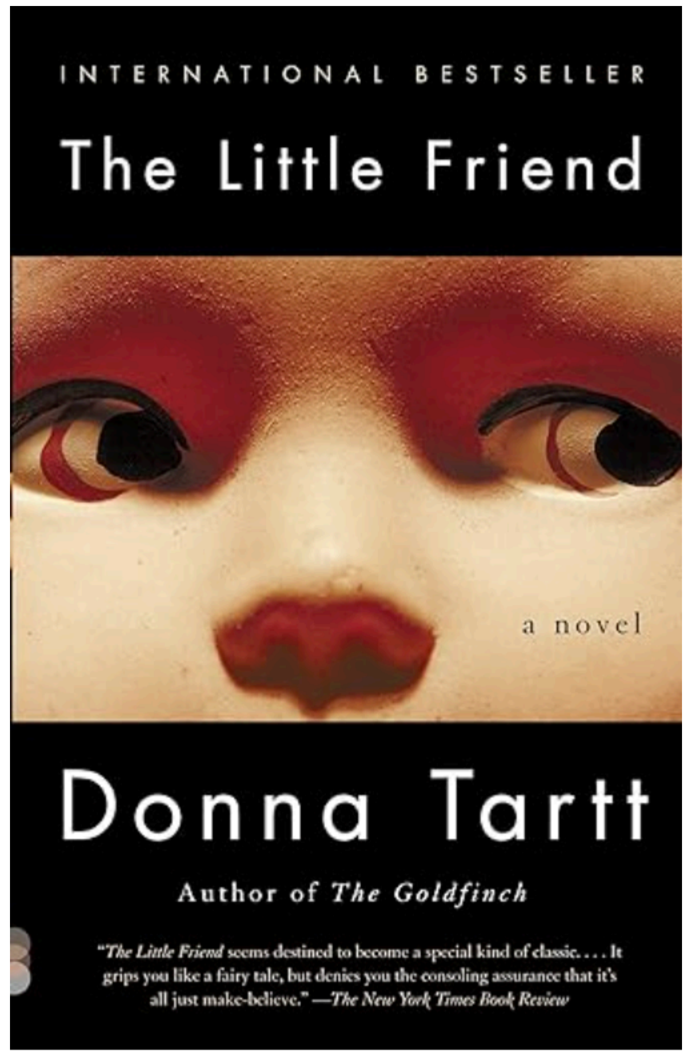
ChatGPT
"The Little Friend" by Donna Tartt is dedicated to her father, Don Tartt. The dedication reads: "For my father, Don Tartt, with love." Donna Tartt is known for being private about her personal life, but she expressed her appreciation and love for her father through this dedication in her second novel, published in 2002.

GPT 3.5, accessed Feb. 2024

Nope! TLF is dedicated to a "Neal"

To be fair, this question is more obscure than the previous one…





"The Little Friend"—Donna Tartt

- Language models **memorize** factual knowledge…**but not perfectly!!!**

Parametric LMs encode factual knowledge, but not equally

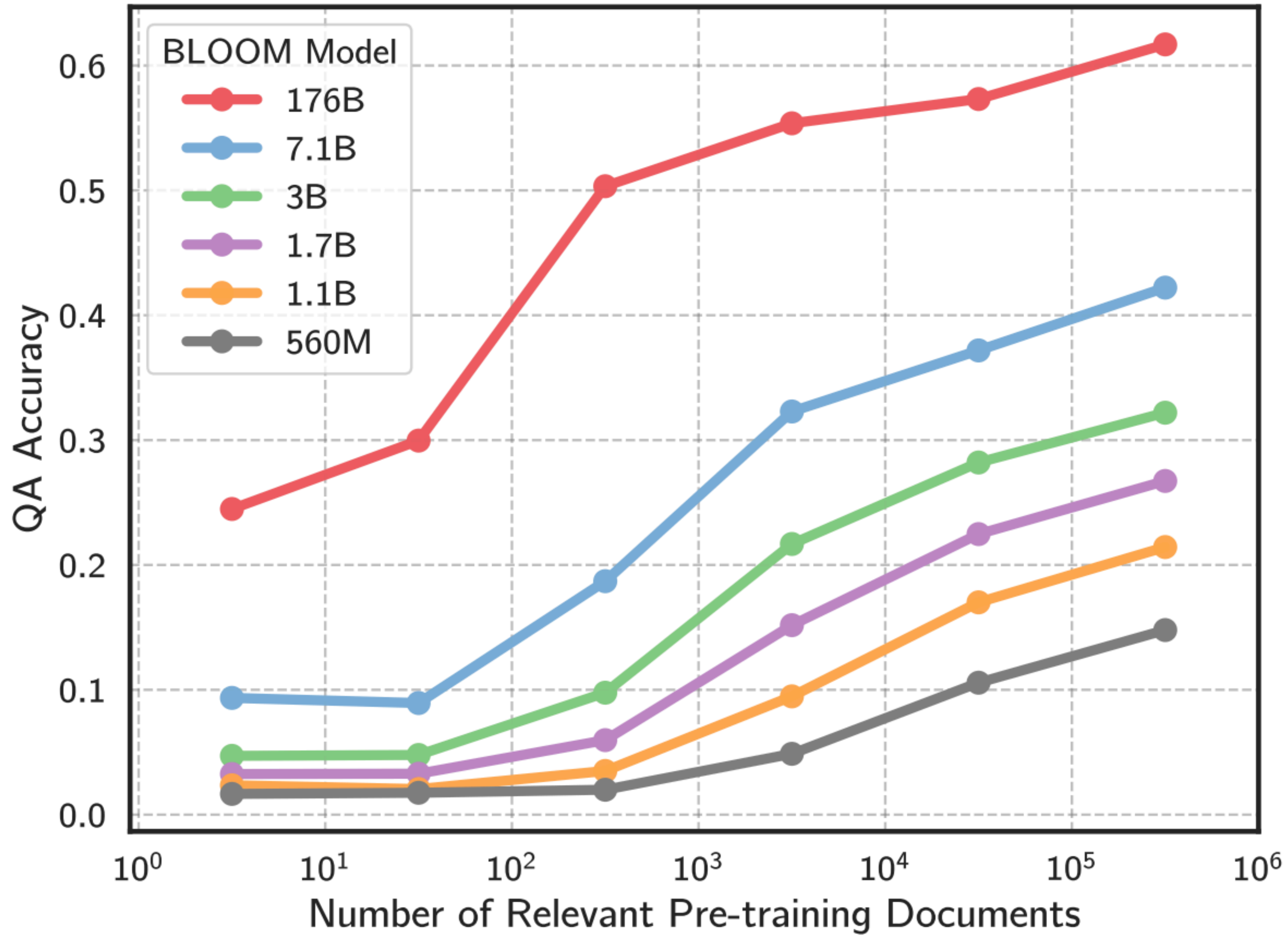


Figure 1 from Kandpal et al. (2022)

Large language models struggle with long-tail knowledge; scaling to achieve good accuracy on the long tail is infeasible (BLOOM → 1 quintillion params.!!!)

Kandpal et al., 2022. Large Language Models Struggle to Learn Long-Tail Knowledge

Parametric LMs encode factual knowledge, but not equally

Relationship	Template
occupation	What is [subj] 's occupation?
place of birth	In what city was [subj] born?
genre	What genre is [subj]?
father	Who is the father of [subj] ?
country	In what country is [subj] ?
producer	Who was the producer of [subj] ?
director	Who was the director of [subj] ?
capital of	What is [subj] the capital of?
screenwriter	Who was the screenwriter for [subj] ?
composer	Who was the composer of [subj] ?
color	What color is [subj] ?
religion	What is the religion of [subj] ?
sport	What sport does [subj] play?
author	Who is the author of [subj] ?
mother	Who is the mother of [subj] ?
capital	What is the capital of [subj] ?

Table 2 from Mallen et al. (2023)

Pop = monthly Wikipedia page views

Pop(Donna Tartt) < Pop(J. K. Rowling)

$\mathrm{Acc}_{\mathrm{LM}}$(Donna Tartt, occupation, Writer)

< $\mathrm{Acc}_{\mathrm{LM}}$(J. K. Rowling, occupation, Writer)

- **PopQA**: Dataset of 14k questions about long-tail entities
  - **Query:** Knowledge triple (*S*, *R*, *O*)
  - **Answers**: Set of entities *E* s.t. knowledge triple (*S*, *R*, *E*) exists in the knowledge graph

Mallen et al., 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

Parametric LMs encode factual knowledge, but not equally

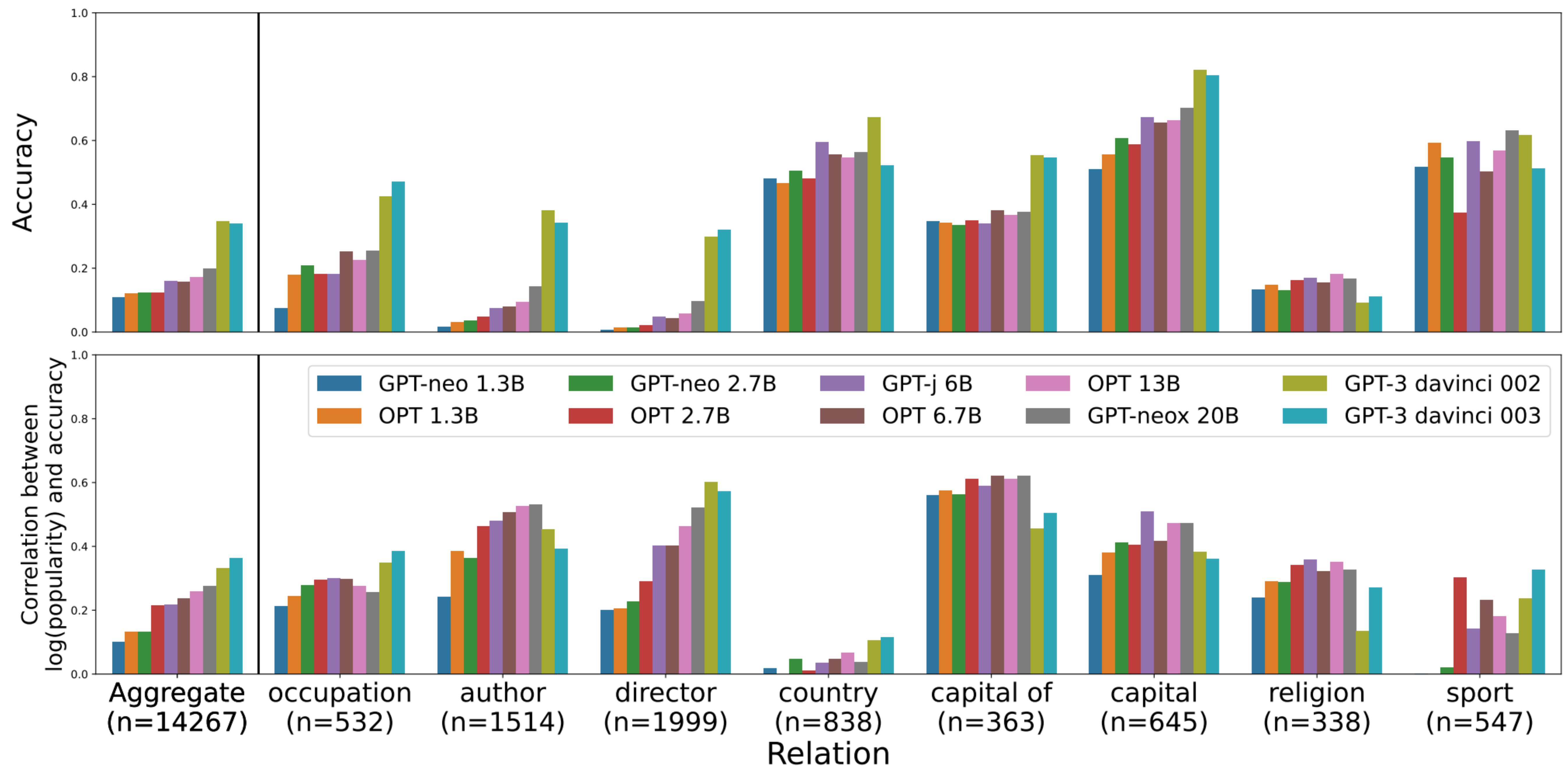


Figure 4 from Mallen et al. (2023)

- Entity popularity and relationship type are **strong predictors** of memorization ability
- LM parametric knowledge fails to extend to long-tail distributions

Mallen et al., 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

Parametric LMs encode factual knowledge, but not equally

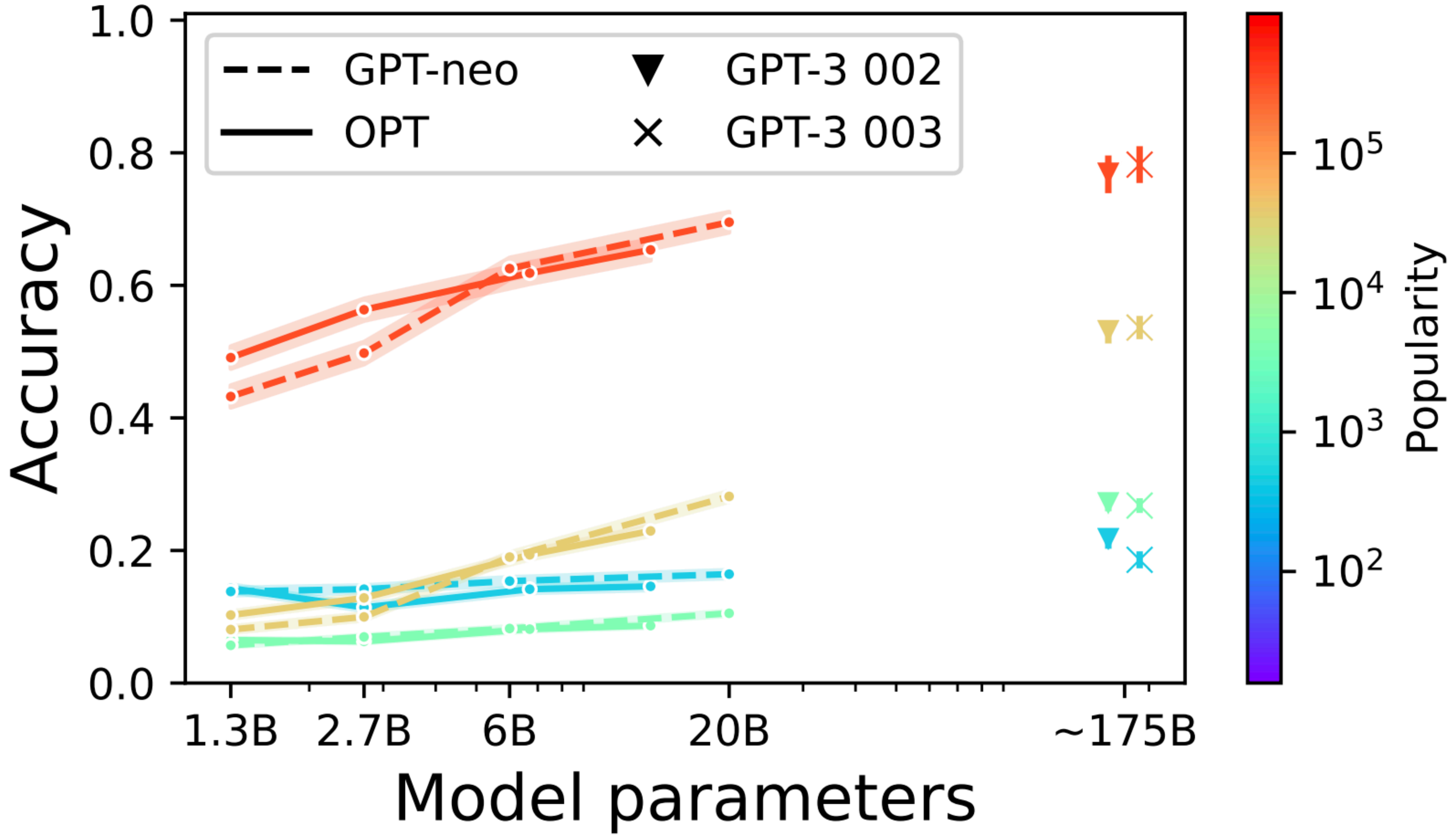


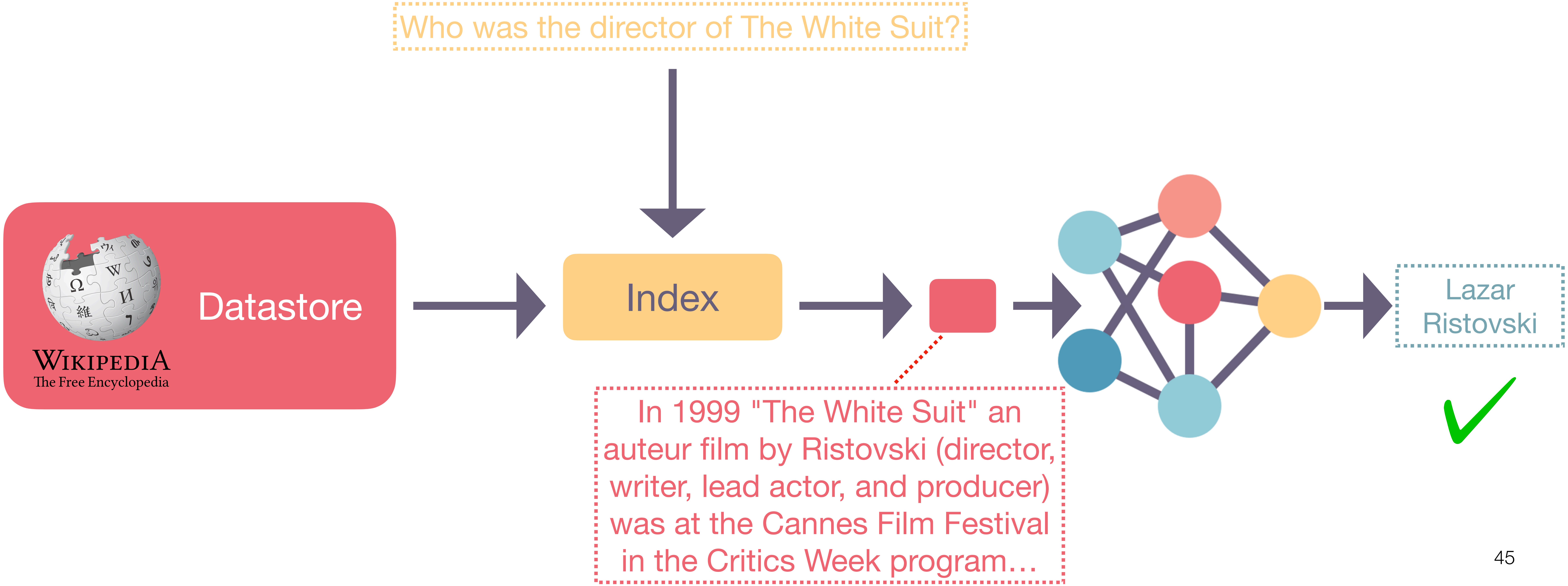
Figure 6 from Mallen et al. (2023)

Scaling does **not** improve memorization of long-tail knowledge!

Mallen et al., 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

Use non-parametric memorization for long-tail knowledge

Key idea: Just as you would consult an **encyclopedia** to look up obscure knowledge, an LM can query an **external datastore** for long-tail information!



Use non-parametric memorization for long-tail knowledge

WIKIPEDIA
The Free Encyclopedia

External datastore

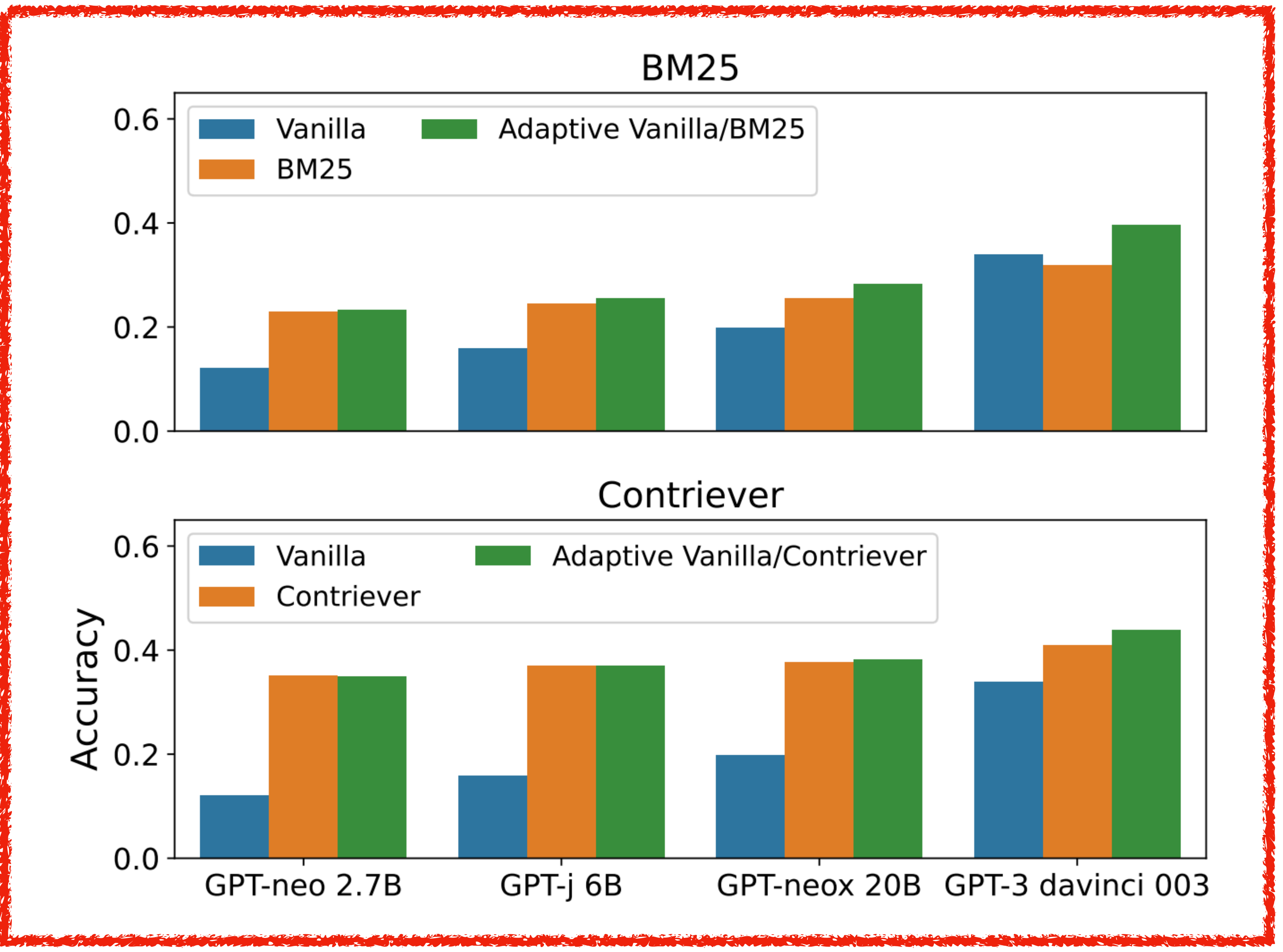


Figure 9 from Mallen et al. (2023)

Better performance on QA with low-popularity entities

- Not only does retrieval help average QA performance, the improvements are significant for **less popular facts.**

Mallen et al., 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

What's the catch? Retrieved context does not always help

	Contriever-augmented LM	
	succeeded	failed
LM succeeded	0.83 (24%)	0.14 (10%)
LM failed	0.88 (17%)	0.11 (49%)

Table 1 from Mallen et al. (2023)

- For 10% of questions, retrieval augmentation can mislead the LM and induce it to answer incorrectly (without retrieval, it would answer correctly) ❌
- Constant retrieval means higher costs and inference-time latency! ❌
- **Workaround**: Adaptive retrieval based on popularity of query ✅

Mallen et al., 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

What's the catch? Retrieved context does not always help

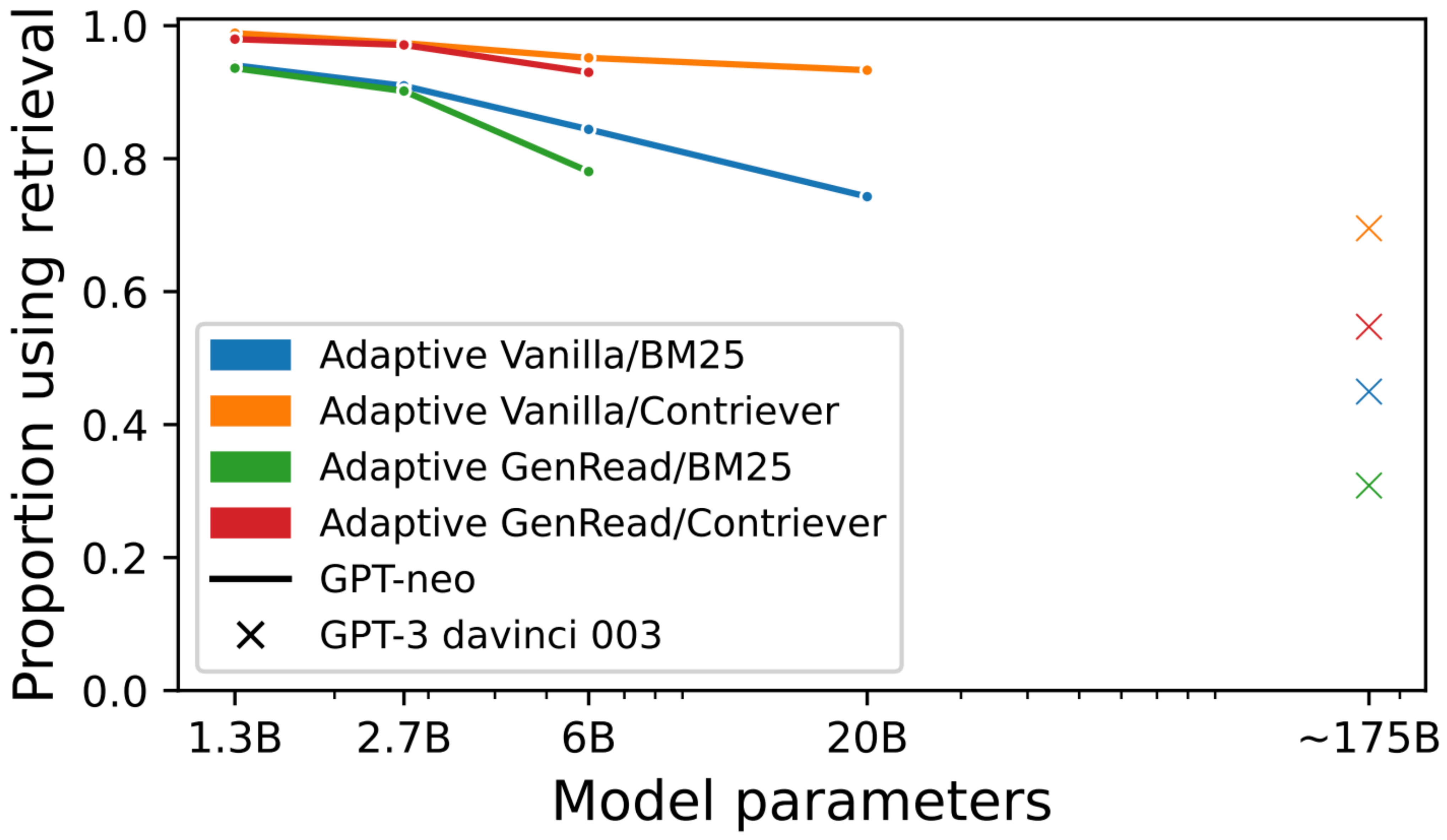


Figure 10 from Mallen et al. (2023)

- Small models (which memorize less) reap the benefits of retrieval more than large models
- Relative to vanilla (constant) retrieval, adaptive retrieval helps **larger** models more than **smaller** models
- Is **entity popularity** the **best** proxy for deciding when to retrieve?

Mallen et al., 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

What are other **downsides** of parametric LMs?
How can retrieval-based LMs **close the gap?**

Easy Knowledge Updates

- **Not all information in the pre-trained LM is desirable!**
  - **Out-of-date information** (e.g., “Ben Bernanke is the chair of the US Federal Reserve…”)
  - **Personally identifiable information (PII)** (e.g, “My Club Penguin password is xxxx…”)
  - **Copyrighted or restricted data** (e.g., “The snow in the mountains was melting and Bunny had been dead for several weeks before we came to understand the gravity of our situation…”)
  - **Domain adaptation** (e.g., “These hipster glasses look so cheugy…2/5 stars 🥸🤬”)

★ Solution: Use retrieval!

Easy Knowledge Updates

- **Key idea:** Simply swap the index—no need for further re-training

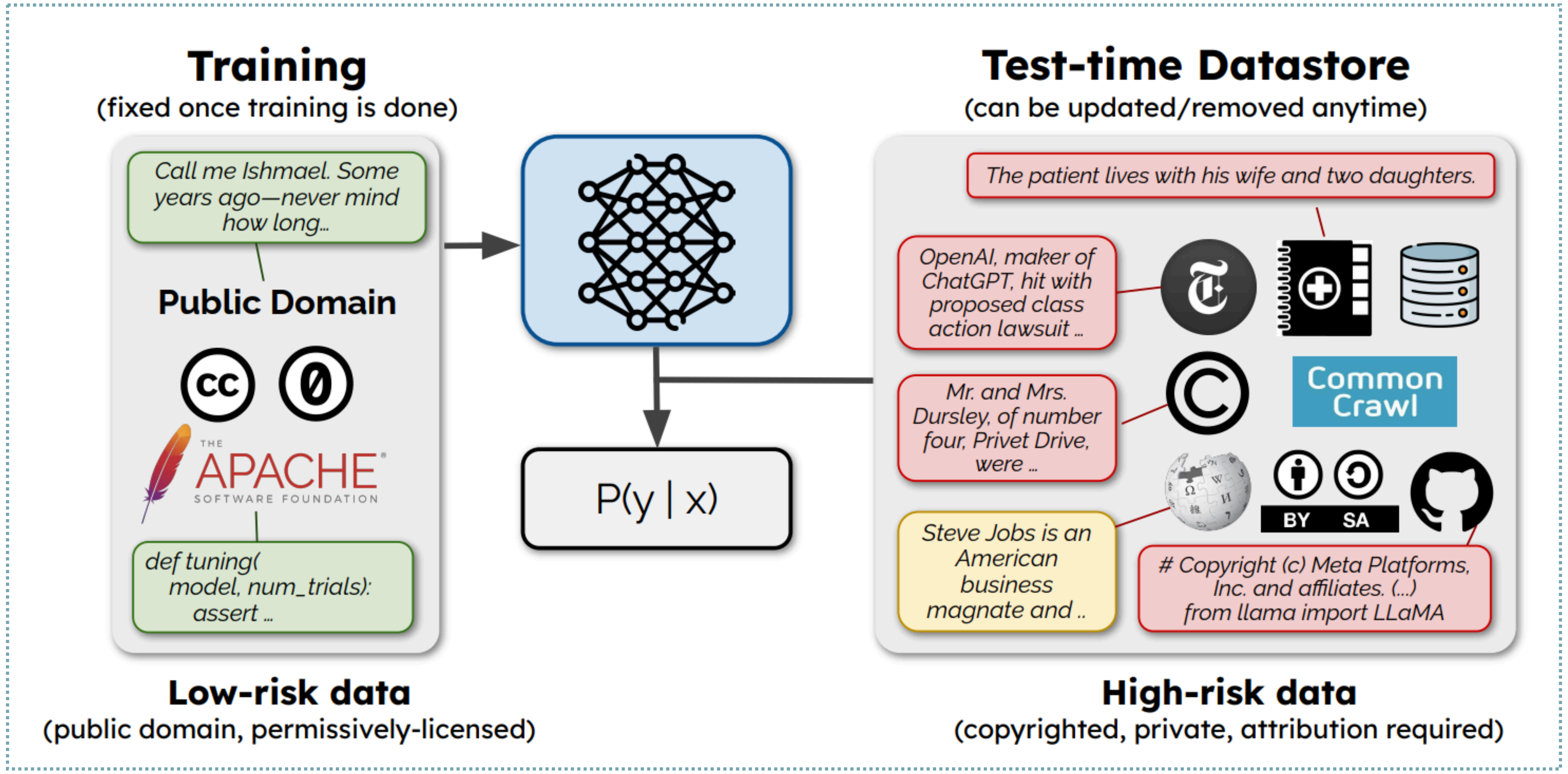


Figure 1 from Min et al. (2023)

Copyright / restricted data: Parametric LMs trained on permissively-licensed data can use a datastore with copy-righted / restricted data, which can be easily swapped out

Min et al., 2023. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore

Easy Knowledge Updates

- **Key idea:** Simply swap the index—no need for further re-training

Training Data	Datastore	Perplexity (↓)	
		Dev	Test
WIKI-3B	-	37.13	34.84
BOOKS	-	14.75	11.89
WIKI-3B	BOOKS	24.85	20.47

Table 4 from Khandelwal et al. (2020)

Domain adaptation: "Free" (i.e., no parameter updates) domain adaptation by creating a datastore for the target domain

Khandelwal et al., 2020.Generalization through Memorization: Nearest Neighbor Language Models

What are some **failure modes** in retrieval-based LMs?

1. Retrieval-based LMs struggle with reasoning

- Retrieval-based LMs show a competitive edge on knowledge-intensive tasks (e.g., ODQA), but improvements do not generalize to other tasks
- Many retrieval-based LMs struggle with **multi-step entailments or logical reasoning**: kNN-LM, REALM, DPR+FiD, Contriever + ATLAS/Flan-T5…
- Retrieval based on similarity metric—which is an imperfect proxy!!

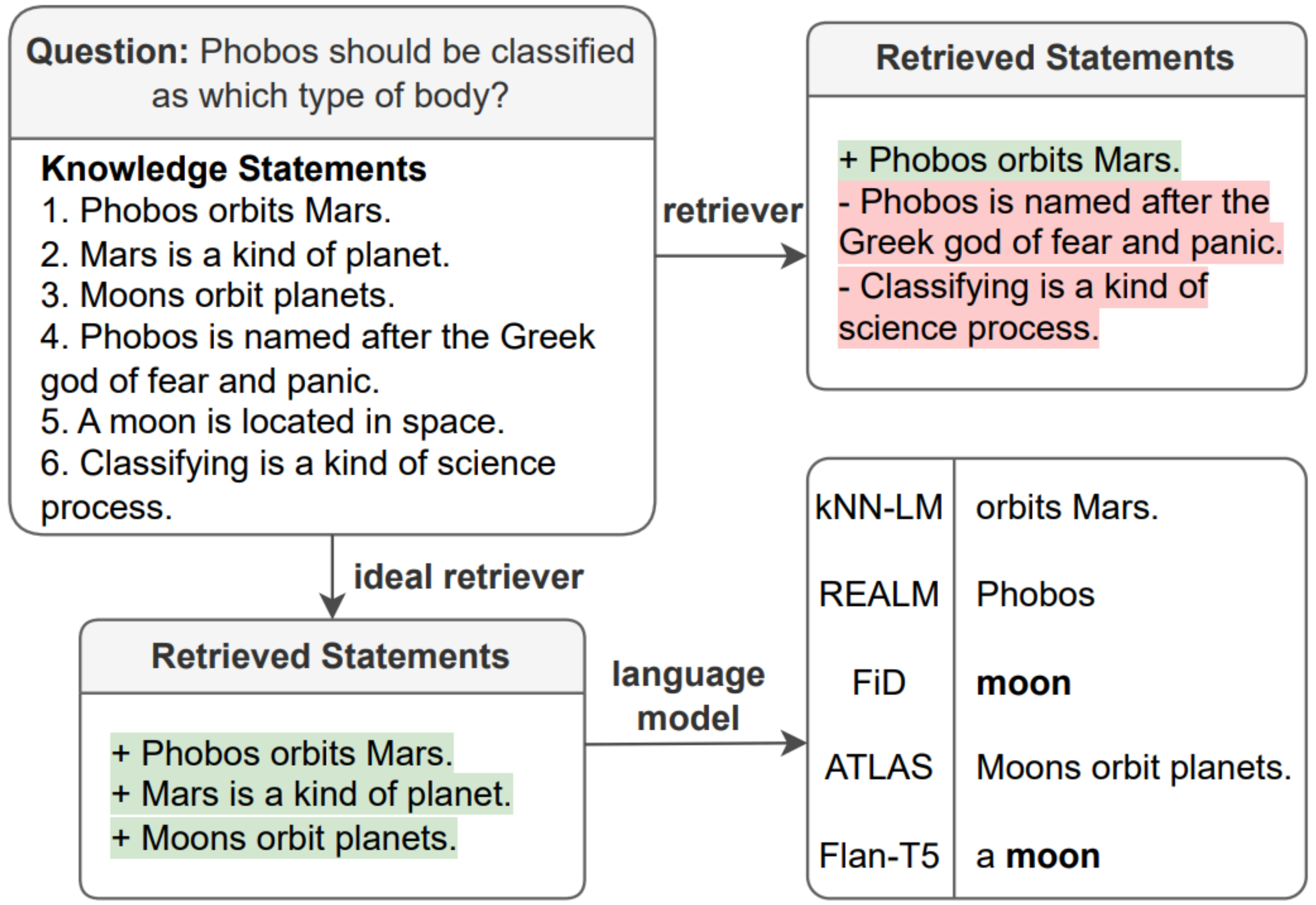


Figure 1 from BehnamGhader et al. (2023)

Both the retriever and the LM are distinct sources of failure

BehnamGhader et al., 2023. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model

2. Retrieval-based LMs are easily distracted by bad context

- Bad (e.g., random, low-quality) context hurts retrieval-based LM performance significantly, such that even a no-retrieval baseline performs better
- Amount of bad context retrieved is datastore-dependent; currently no good intuition as to what constitutes a desirable datastore (besides Wikipedia)

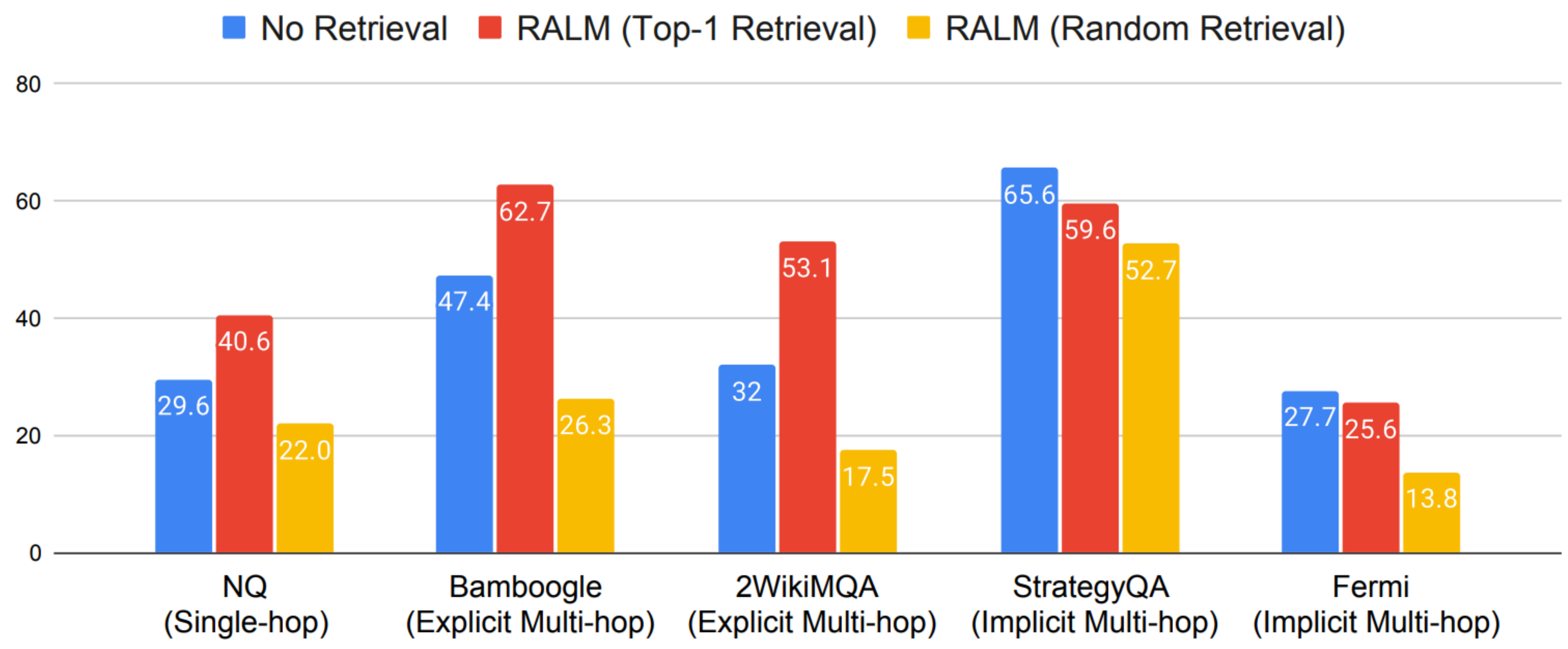


Figure 2 from Yoran et al. (2023)

Yoran et al., 2023. Making Retrieval-Augmented Language Models Robust to Irrelevant Context

References

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, Wen-tau Yih. 2024. Reliable, Adaptable, and Attributable Language Models with Retrieval.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 15492–15509, Singapore. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

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

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2023. Silo language models: Isolating legal risk in a nonparametric datastore. Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context.

Food for Thought! (Discussion Questions)

Izacard et al., 2022

1. Why didn't the four training objectives for the retriever result in a notable improvement in end-to-end performance compared to pretext training?
2. What additional desirable properties of retrievers were not focused on for optimization in this study?
3. Concerning temporal sensitivity, how could we enhance optimization for queries with ambiguous target time periods?

Mallen et al., 2023

1. Using retrieval helps with domain adaptation. What are the pros and cons of using retrieval-based language modeling versus domain-adaptive pre-training?
2. In Mallen et al., 2023, adaptive retrieval works based on whether the query falls under a pre-determined popularity threshold. What are the limitations of this heuristic; is there a better proxy to decide when to retrieve?
3. A recurrent theme with retrieval-based LMs is that we can beat scaling trends simply by offloading knowledge from the model parameters to some non-parametric repository. Do smaller retrieval-based LMs beat larger parametric LMs on every task?
4. How would you design a retrieval-based LM that can better withstand irrelevant or misleading context? Does the source of failure lie in the base LM, or the retriever?
5. Besides QA, what are other knowledge-intensive tasks that retrieval-based LMs might have an edge on? What are tasks that retrieval-based LMs might struggle with?