Understanding Language Models Through the Lens of their Training Data

Nikhil Kandpal
University of Toronto & Vector Institute

What do I mean by "understanding through the lens of training data"

1. Understanding the relationship between an LLM's *capabilities* and the *quantity of relevant information* in its training set

What do I mean by “understanding through the lens of training data”

1. Understanding the relationship between an LLM’s *capabilities* and the *quantity of relevant information* in its training set

memorization
arithmetic
fact-learning
*zero-shot generalization**

What do I mean by "understanding through the lens of training data"

1. Understanding the relationship between an LLM's *capabilities* and the *quantity of relevant information* in its training set

2. Understanding the *counterfactual effect* of removing a single training example from the training data (i.e. training data attribution)

What do I mean by "understanding through the lens of training data"

1. Understanding the relationship between an LLM's *capabilities* and the *quantity of relevant information* in its training set

2. Understanding the *counterfactual effect* of removing a single training example from the training data (i.e. training data attribution)

AI created a song mimicking the work of Drake and The Weeknd. What does that mean for copyright law?

A Harvard Law expert explains why AI-generated art doesn't qualify for copyright protection — but how it nonetheless will 'materially affect' the music industry

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

COPYRIGHT —

Stable Diffusion copyright lawsuits could be a legal earthquake for AI

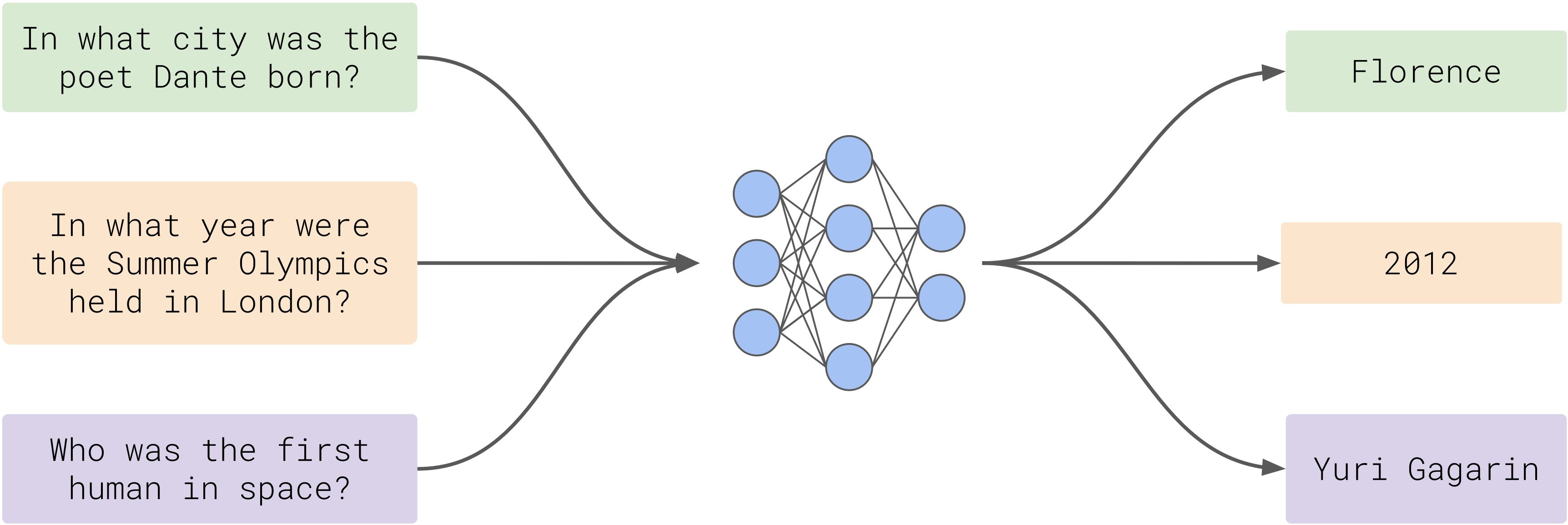
Experts say generative AI is in uncharted legal waters.

What do I mean by “understanding through the lens of training data”

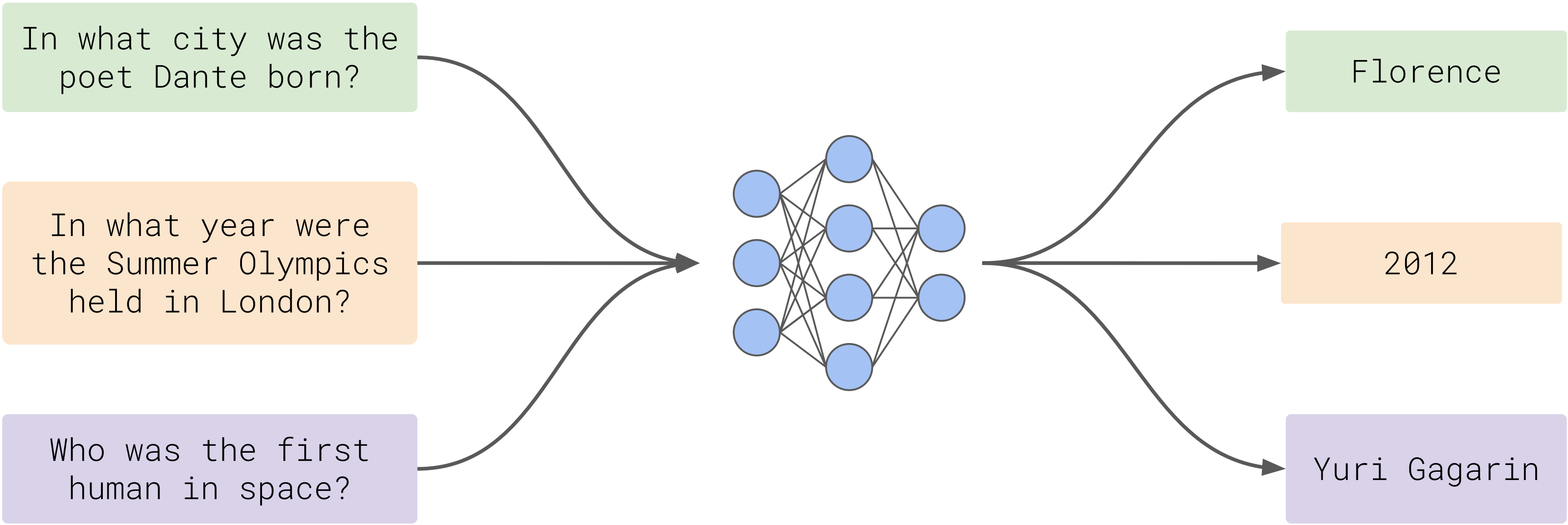
1. Understanding the relationship between an LLM’s *capabilities* and the *quantity of relevant information* in its training set

memorization
arithmetic
fact-learning
*zero-shot generalization**

Pre-trained language models capture a wide range of knowledge

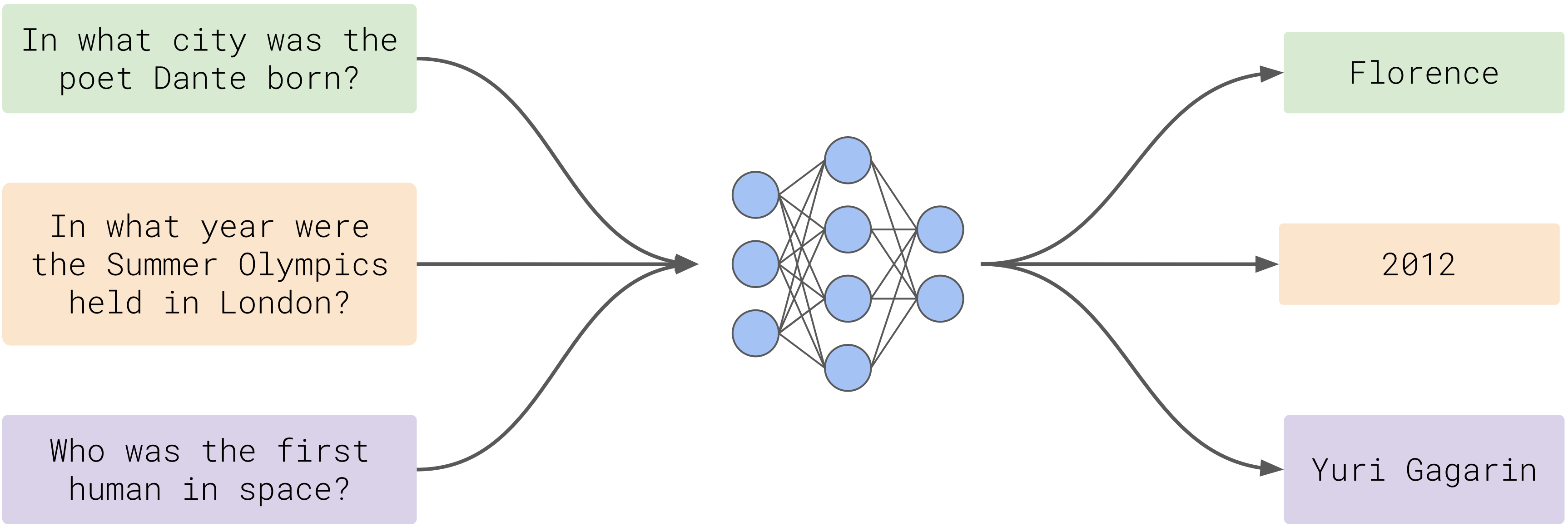


Pre-trained language models capture a wide range of knowledge



Certain pieces of information are learned (or not learned) by language models

Pre-trained language models capture a wide range of knowledge



Certain pieces of information are learned (or not learned) by language models

(How) is this related to the quantity of relevant training data?

How does the number of times a fact is seen during pre-training impact whether a language model learns that fact?

How does the number of times a fact is seen during pre-training impact whether a language model learns that fact?

Simple Experiment:

1. Identify a set of facts

How does the number of times a fact is seen during pre-training impact whether a language model learns that fact?

Simple Experiment:

1. Identify a set of facts
2. Count how many times each fact occurs in a pre-training dataset

How does the number of times a fact is seen during pre-training impact whether a language model learns that fact?

Simple Experiment:

1. Identify a set of facts
2. Count how many times each fact occurs in a pre-training dataset
3. Evaluate an LM's ability to recall each fact

Identifying a set of facts

Identifying a set of facts

Repurpose existing factoid QA datasets (think TriviaQA, Natural Questions, etc.):

Identifying a set of facts

Repurpose existing factoid QA datasets (think TriviaQA, Natural Questions, etc.):

Question - Answer Pair

(`In what city was the poet Dante born?` , `Florence`)

Identifying a set of facts

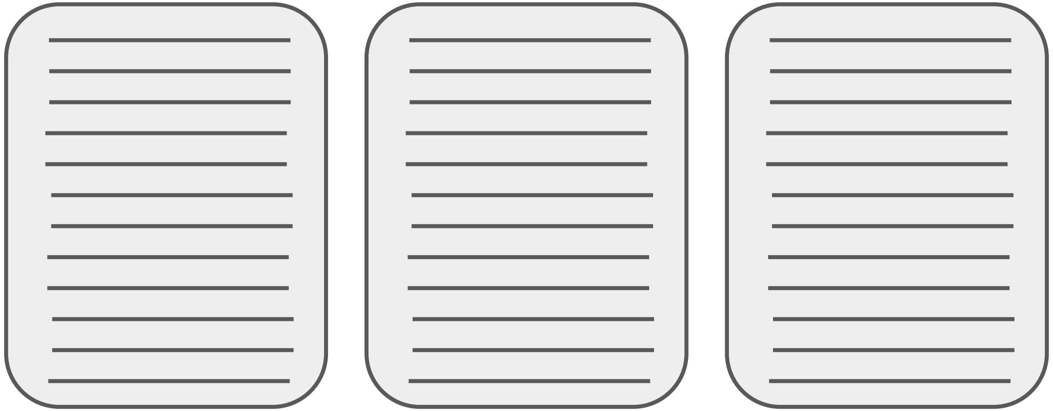
Repurpose existing factoid QA datasets (think TriviaQA, Natural Questions, etc.):

Question - Answer Pair

(`In what city was the poet Dante born?` , `Florence`) ⟷ `The poet Dante was born in the city of` `Florence`

Fact

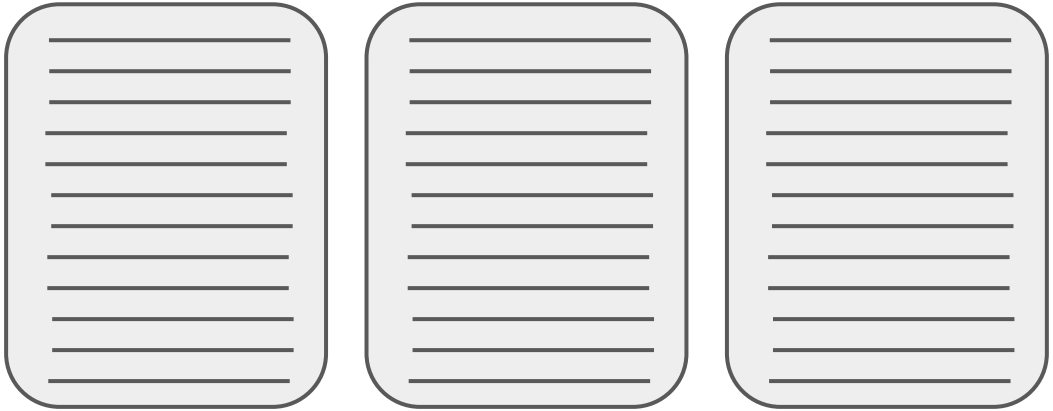
Counting fact instances in pre-training datasets



(In what city was the poet Dante born? , Florence)

Counting fact instances in pre-training datasets

1. Entity link Q-A pair



(In what city was the poet Dante born? , Florence)

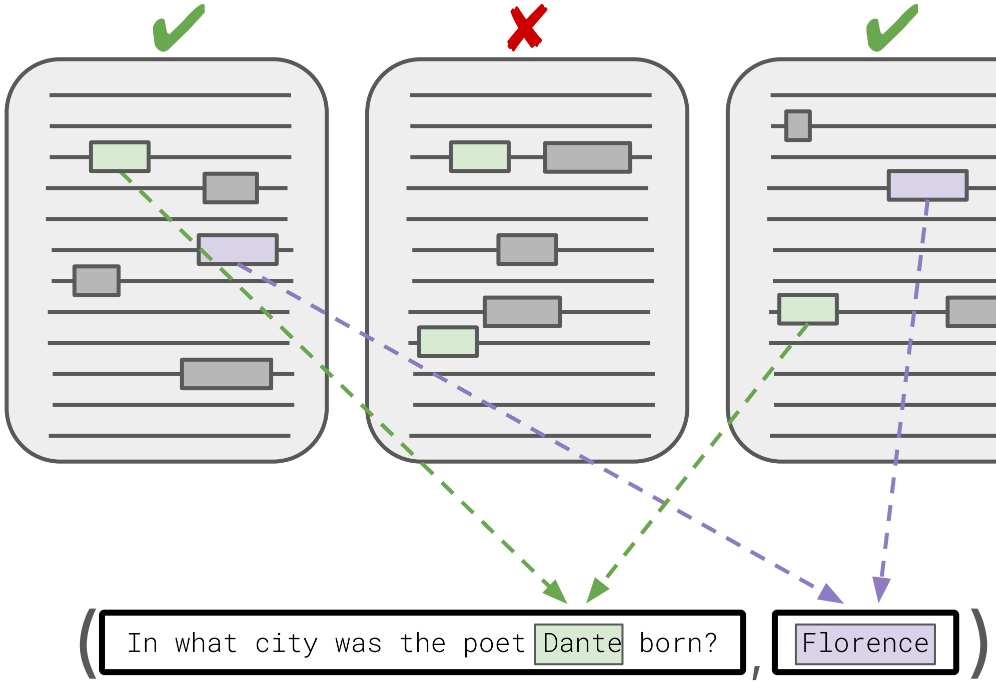
Counting fact instances in pre-training datasets

1. Entity link Q-A pair
2. Entity link training documents

(In what city was the poet Dante born? , Florence)

Counting fact instances in pre-training datasets

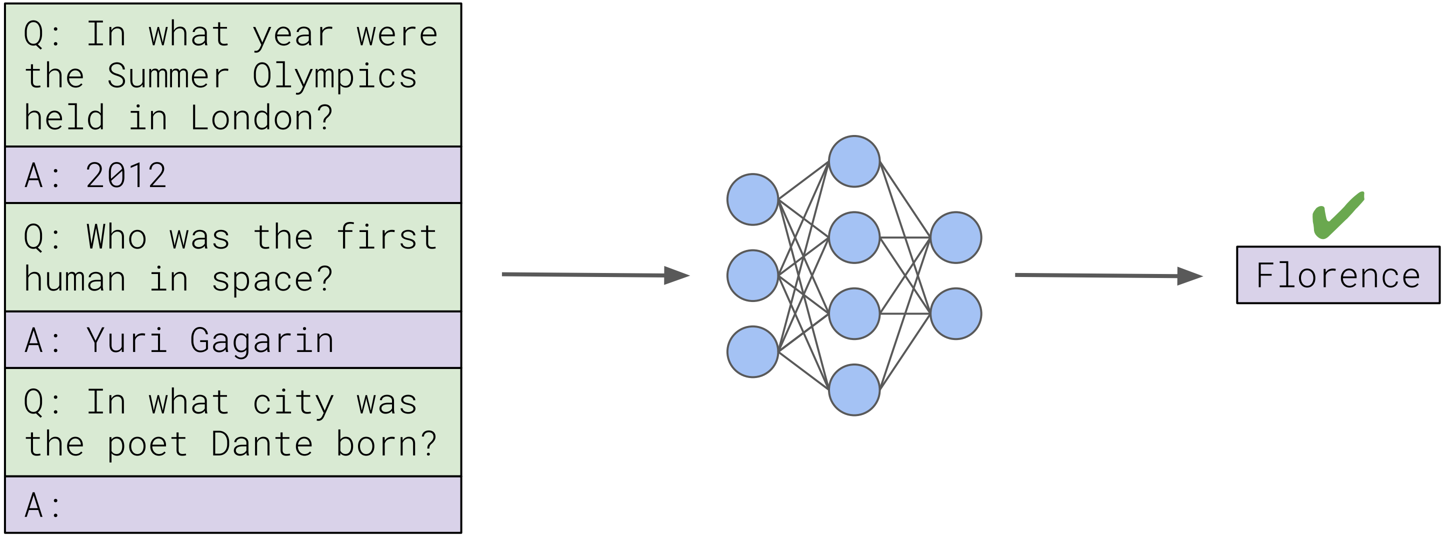
1. Entity link Q-A pair
2. Entity link training documents
3. Count documents containing both Q and A entities



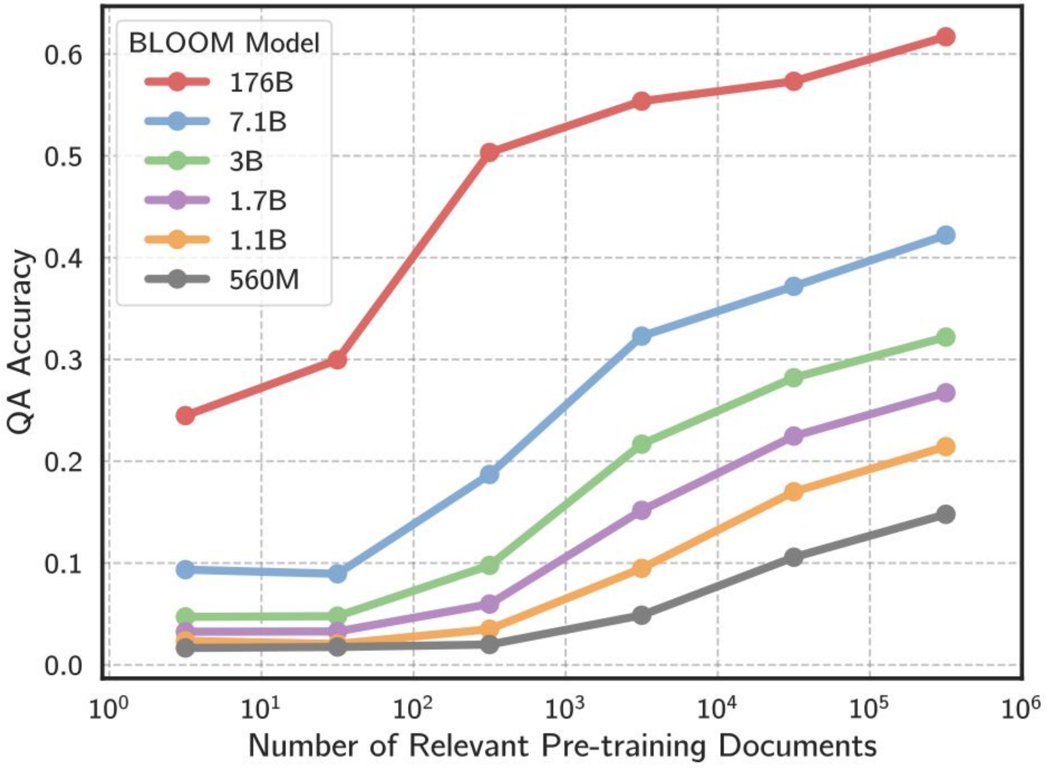
Evaluating a Language Model's Fact Recall

Evaluating a Language Model's Fact Recall

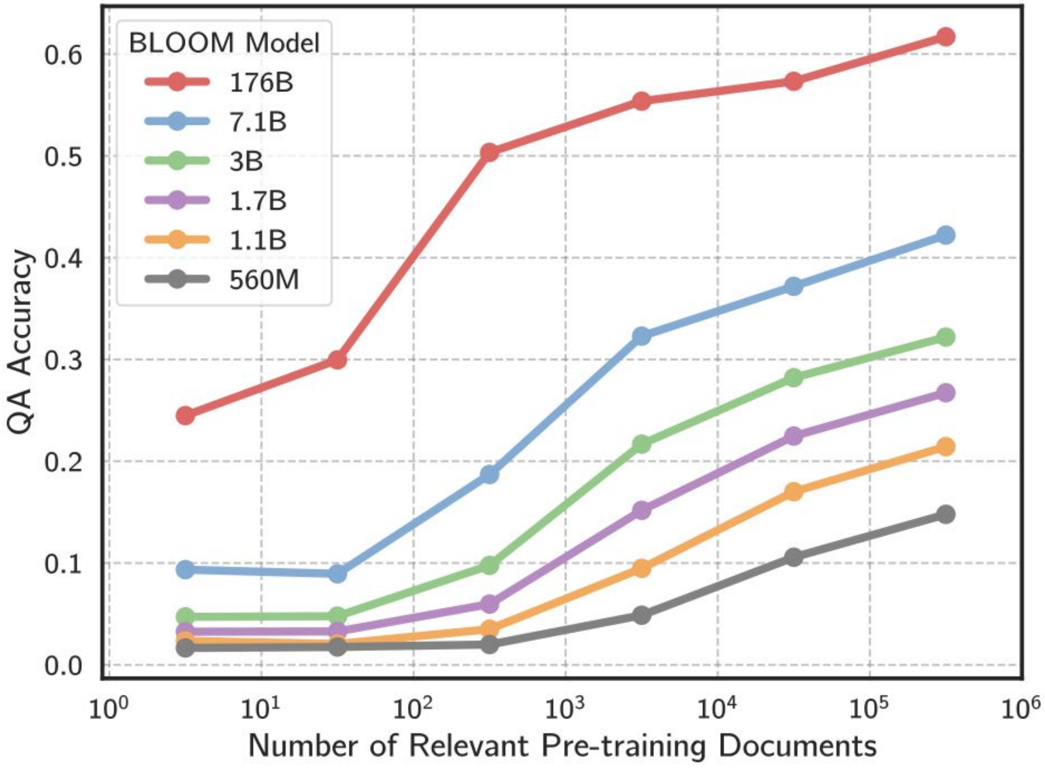
Few-Shot Question Answering



Language models struggle to capture long-tail facts

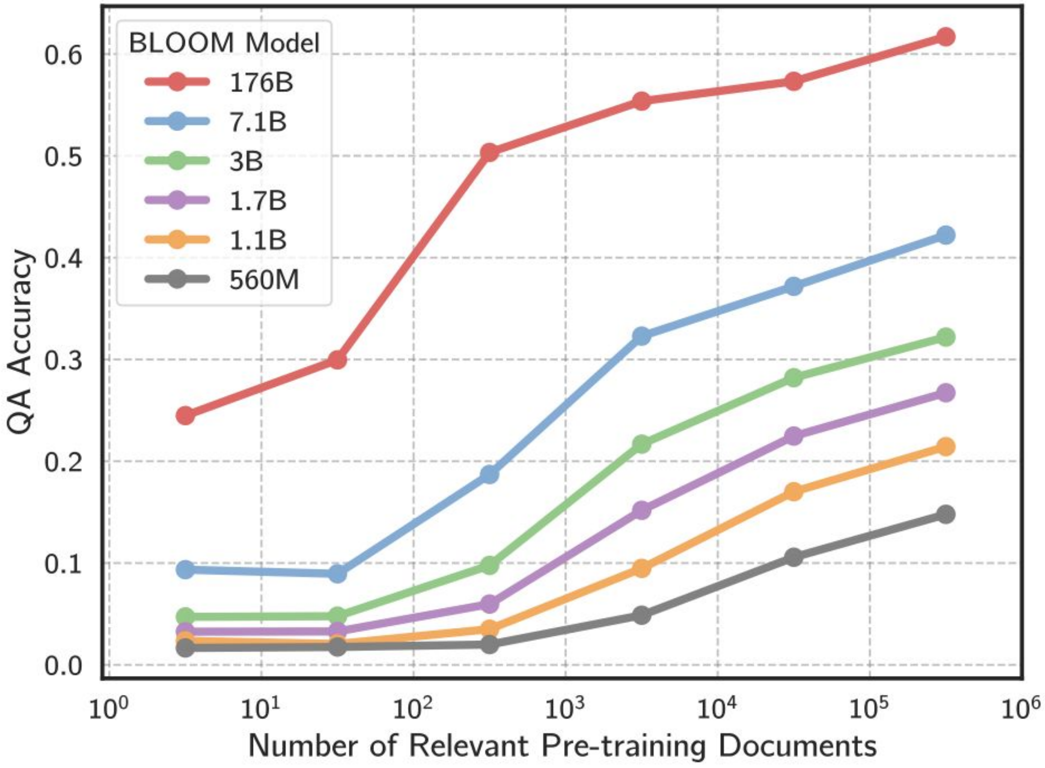


Language models struggle to capture long-tail facts



Observation #1
Larger models are more effective at capturing facts that are both rare and common in the training data

Language models struggle to capture long-tail facts

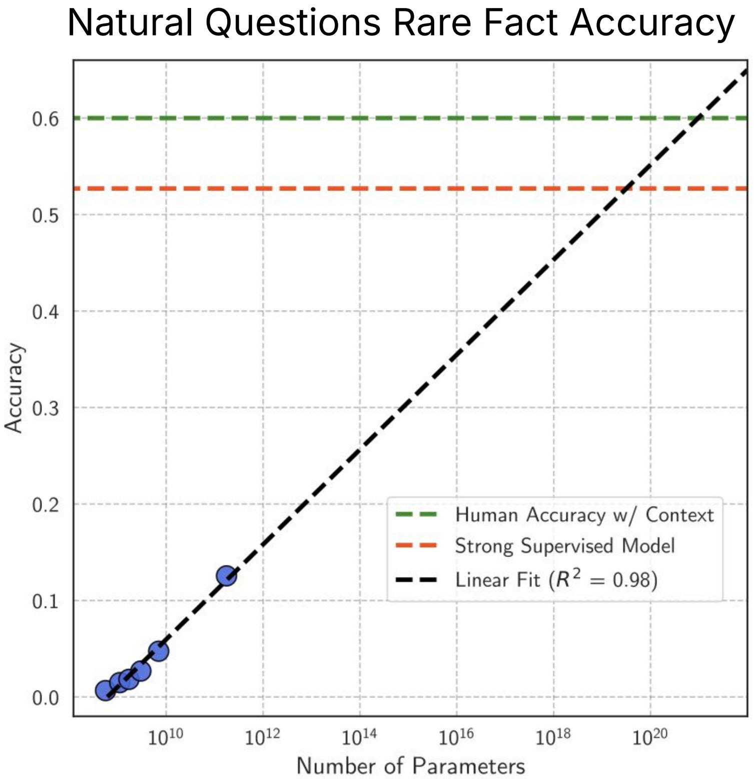


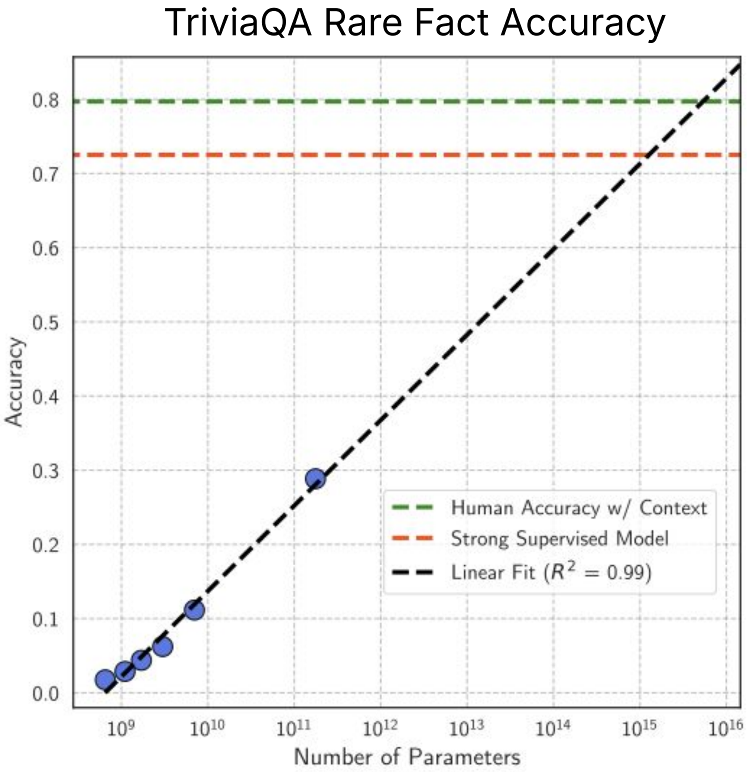
Observation #1
Larger models are more effective at capturing facts that are both rare and common in the training data

Observation #2
Models of all sizes require a fact to be present many times in the training data to reliably learn that fact

Scaling model size has diminishing returns for learning long-tail knowledge

Scaling model size has diminishing returns for learning long-tail knowledge



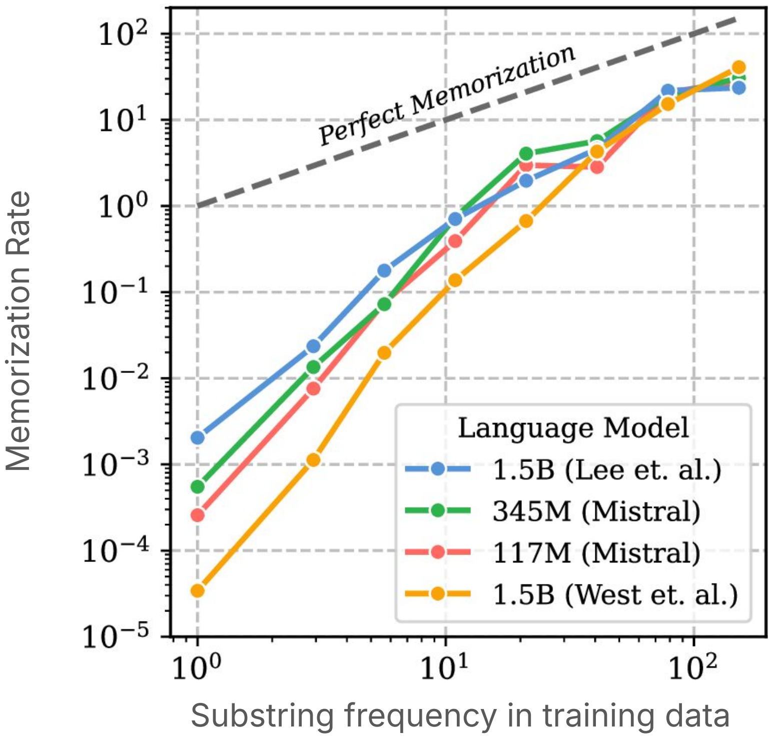


What other capabilities have been characterized in this way?

Quantity of relevant data influences language model capabilities

Example 1: LMs tend to memorize text that appears more in training data

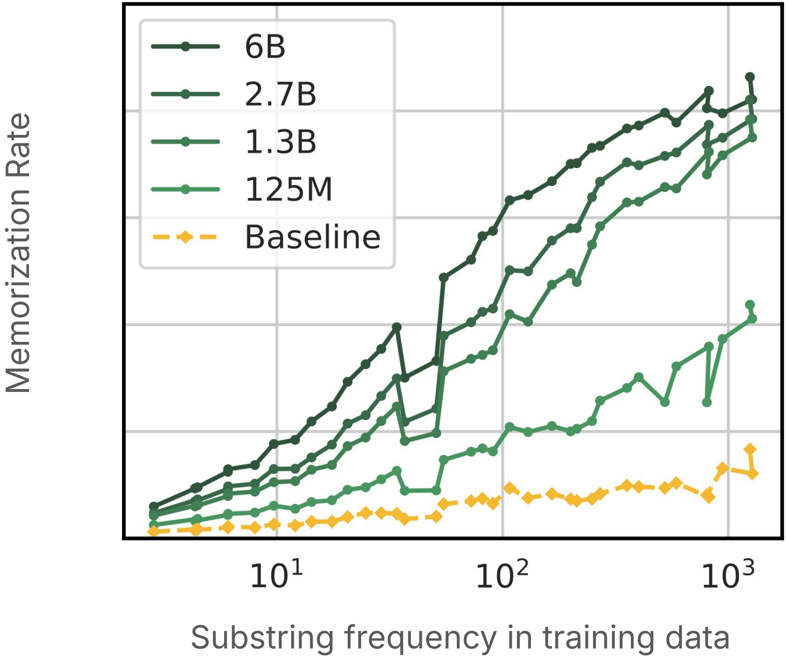
Kandpal et. al. (2022a)



Quantity of relevant data influences language model capabilities

Example 1: LMs tend to memorize text that appears more in training data

Kandpal et. al. (2022a), Carlini et. al. (2022)



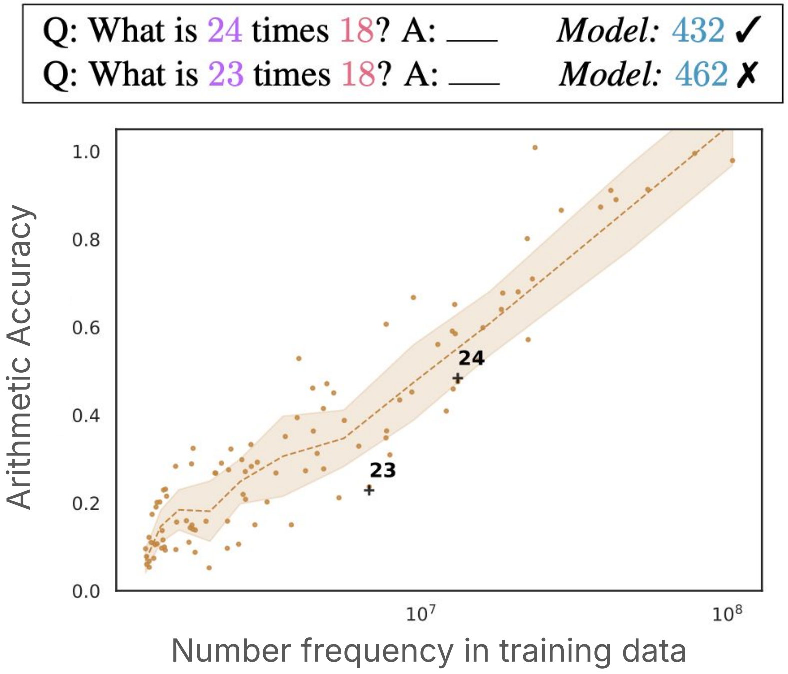
Quantity of relevant data influences language model capabilities

Example 1: LMs tend to memorize text that appears more in training data

Kandpal et. al. (2022a), Carlini et. al. (2022)

Example 2: LMs excel at arithmetic on numbers that appear more in the training data

Razeghi et. al. (2022)



Quantity of relevant data influences language model capabilities

Example 1: LMs tend to memorize text that appears more in training data

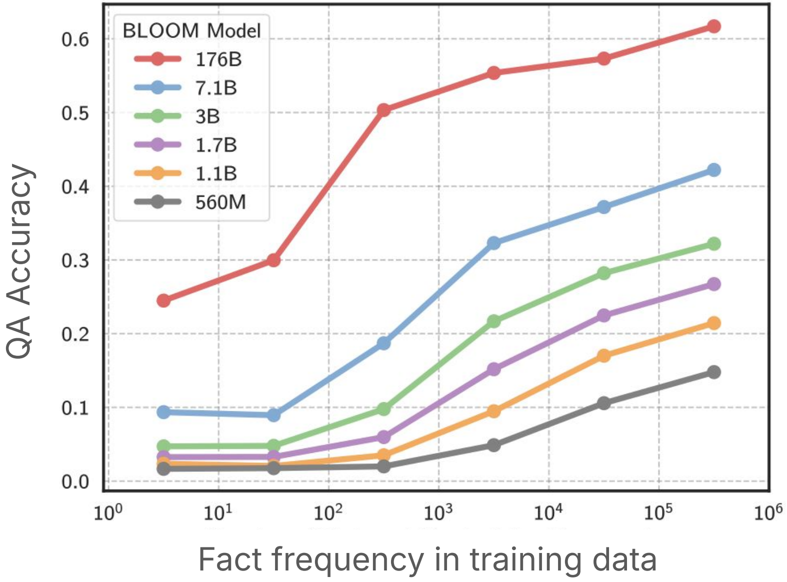
Kandpal et. al. (2022a), Carlini et. al. (2022)

Example 2: LMs excel at arithmetic on numbers that appear more in the training data

Razeghi et. al. (2022)

Example 3: LMs learn facts that appear more in the training data

Kandpal et. al. (2022b)



Quantity of relevant data influences language model capabilities

Example 1: LMs tend to memorize text that appears more in training data

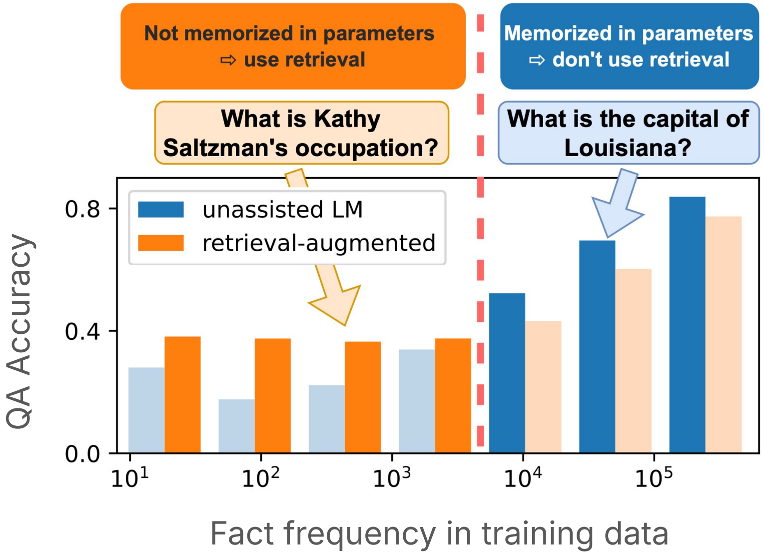
Kandpal et. al. (2022a), Carlini et. al. (2022)

Example 2: LMs excel at arithmetic on numbers that appear more in the training data

Razeghi et. al. (2022)

Example 3: LMs learn facts that appear more in the training data

Kandpal et. al. (2022b), Mallen et. al. (2022)



Quantity of relevant data influences language model capabilities

Example 1: LMs tend to memorize text that appears more in training data

Kandpal et. al. (2022a), Carlini et. al. (2022)

Example 2: LMs excel at arithmetic on numbers that appear more in the training data

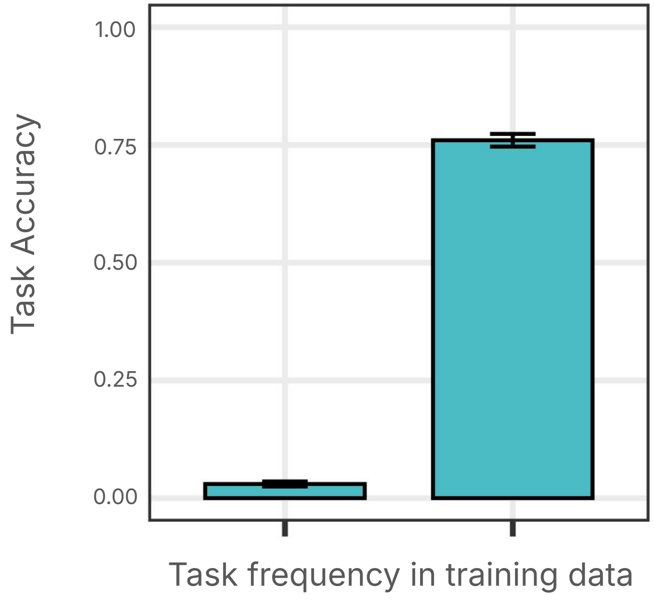
Razeghi et. al. (2022)

Example 3: LMs learn facts that appear more in the training data

Kandpal et. al. (2022b), Mallen et. al. (2022)

Example 4: LMs can perform variants of a task when that variant appears more in the training data

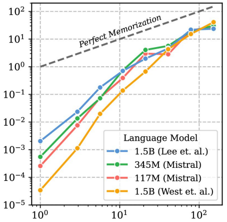
Mccoy et. al. (2023)



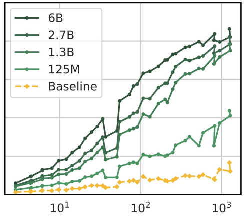
Quantity of relevant data influences language model capabilities

Memorization

Kandpal et. al. (2022a)

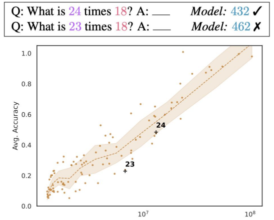


Carlini et. al. (2022)



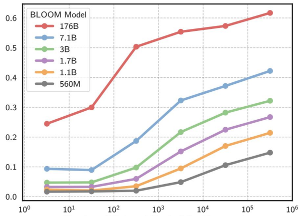
Arithmetic

Razeghi et. al. (2022)

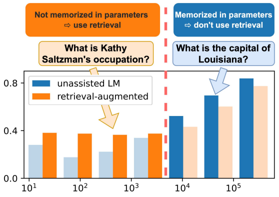


Fact Learning

Kandpal et. al. (2022b)

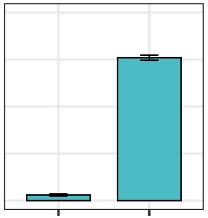


Mallen et. al. (2022)



Task Learning

Mccoy et. al. (2023)

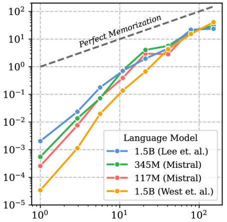


Quantity of relevant data influences language model capabilities

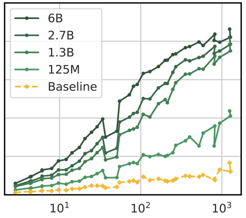
Higher-level (more “interesting”) behaviors

Memorization

Kandpal et. al. (2022a)

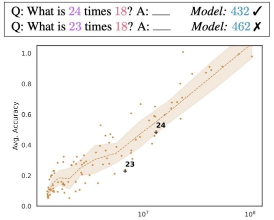


Carlini et. al. (2022)



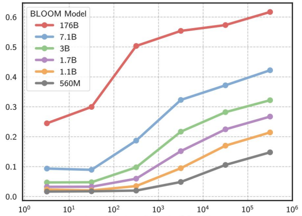
Arithmetic

Razeghi et. al. (2022)

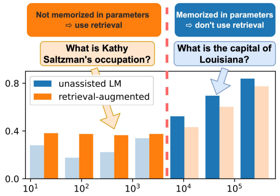


Fact Learning

Kandpal et. al. (2022b)

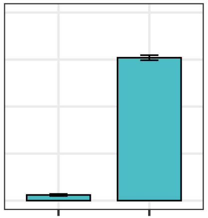


Mallen et. al. (2022)



Task Learning

Mccoy et. al. (2023)



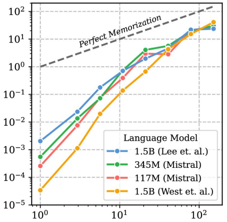
Quantity of relevant data influences language model capabilities

Higher-level (more "interesting") behaviors

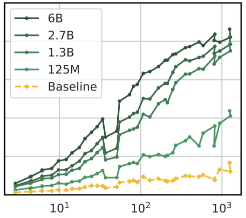
but also more difficult to study

Memorization

Kandpal et. al. (2022a)

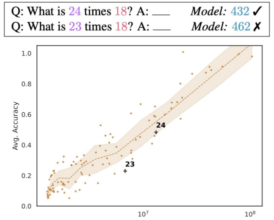


Carlini et. al. (2022)



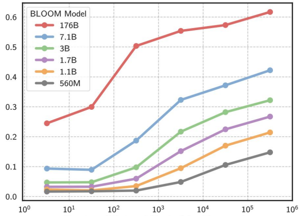
Arithmetic

Razeghi et. al. (2022)

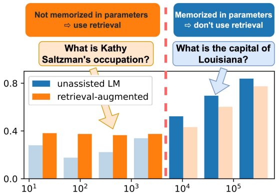


Fact Learning

Kandpal et. al. (2022b)

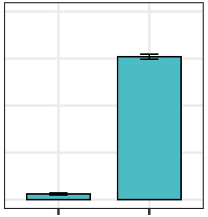


Mallen et. al. (2022)



Task Learning

Mccoy et. al. (2023)



Quantity of relevant data influences language model capabilities

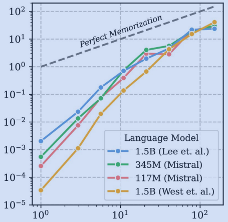
Higher-level (more “interesting”) behaviors

but also more difficult to study

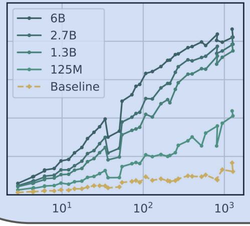
Exactly Compute Training Frequency

Memorization

Kandpal et. al. (2022a)

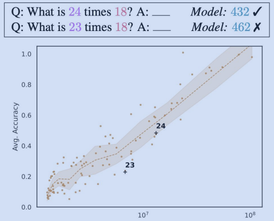


Carlini et. al. (2022)



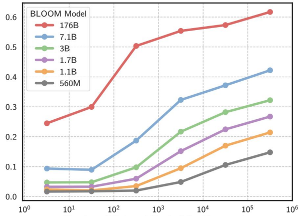
Arithmetic

Razeghi et. al. (2022)

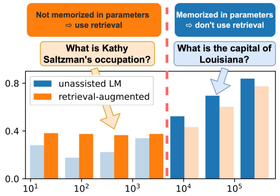


Fact Learning

Kandpal et. al. (2022b)

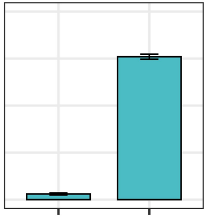


Mallen et. al. (2022)



Task Learning

Mccoy et. al. (2023)



Quantity of relevant data influences language model capabilities

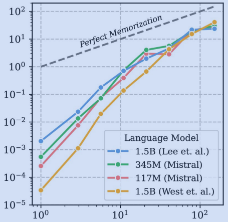
Higher-level (more “interesting”) behaviors

but also more difficult to study

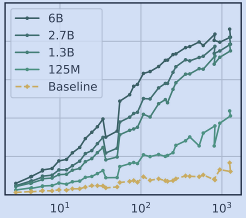
Exactly Compute Training Frequency

Memorization

Kandpal et. al. (2022a)

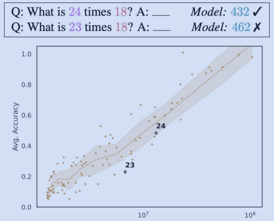


Carlini et. al. (2022)



Arithmetic

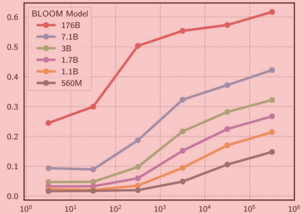
Razeghi et. al. (2022)



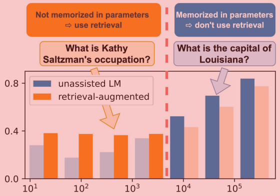
Approximate Training Frequency

Fact Learning

Kandpal et. al. (2022b)

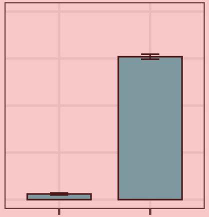


Mallen et. al. (2022)



Task Learning

Mccoy et. al. (2023)



Understanding the counterfactual effect of individual training examples

Simulating the removal of a training example without re-training from scratch is difficult

Understanding the counterfactual effect of individual training examples

Simulating the removal of a training example without re-training from scratch is difficult

What about an attribution method like Influence Functions [Koh & Liang 2017]?

Understanding the counterfactual effect of individual training examples

Simulating the removal of a training example without re-training from scratch is difficult

What about an attribution method like Influence Functions [Koh & Liang 2017]?

These only accurately simulate leave-one-out retraining when...

- Models are trained with a strongly-convex objective
- Models are trained to convergence
- Training is deterministic

Understanding the counterfactual effect of individual training examples

Simulating the removal of a training example without re-training from scratch is difficult

What about an attribution method like Influence Functions [Koh & Liang 2017]?

These only accurately simulate leave-one-out retraining when...

- Models are trained with a strongly-convex objective
- Models are trained to convergence
- Training is deterministic

Instead let's focus on methods that allow *exact(!)* and *scalable(!)* attribution under more realistic assumptions

Understanding the counterfactual effect of individual training examples

Assume: Our dataset is heterogeneous, containing some data that we *must do* attribution for and some *that do not need* attribution

Understanding the counterfactual effect of individual training examples

Assume: Our dataset is heterogeneous, containing some data that we *must do* attribution for and some *that do not need* attribution

Coming Soon: The Common Pile

~2 trillion tokens of permissively licensed and public domain text

UNIVERSITY OF TORONTO

EleutherAI

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Massachusetts Institute of Technology

Understanding the counterfactual effect of individual training examples

Assume: Our dataset is heterogeneous, containing some data that we *must do* attribution for and some *that do not need* attribution

Approach:

1. Pre-train an LLM on data that does not require attribution
2. Incorporate the remaining data into the LLM in a “simple” way that allows for exact and efficient attribution

Understanding the counterfactual effect of individual training examples

Assume: Our dataset is heterogeneous, containing some data that we *must do* attribution for and some *that do not need to do* attribution for

Approach:

1. Pre-train an LLM on data that does not require attribution
2. Incorporate the remaining data into the LLM in a “simple” way that allows for exact and efficient attribution

Semi-Parametric Language Models

Retrieval-Augmented Generation (RAG)

kNN-LM: One Flavor of a Semi-Parametric Language Model

Test Context x	Target
Obama's birthplace is	?

Figure from Khandelwal et. al. 2020

kNN-LM: One Flavor of a Semi-Parametric Language Model

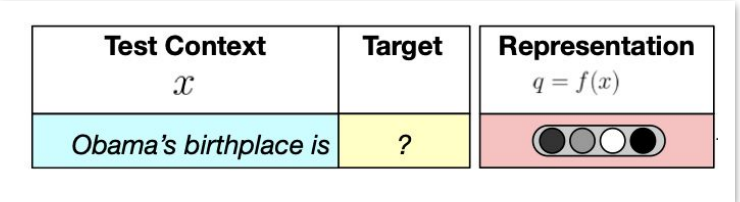


Figure from Khandelwal et. al. 2020

kNN-LM: One Flavor of a Semi-Parametric Language Model

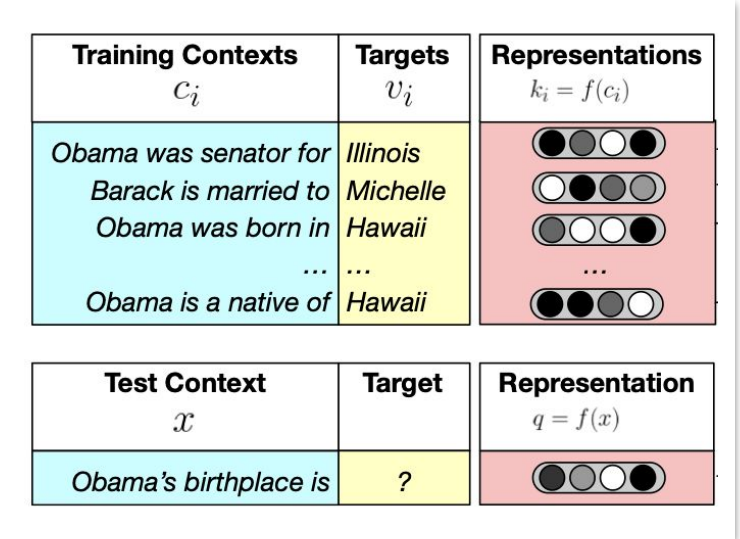


Figure from Khandelwal et. al. 2020

kNN-LM: One Flavor of a Semi-Parametric Language Model

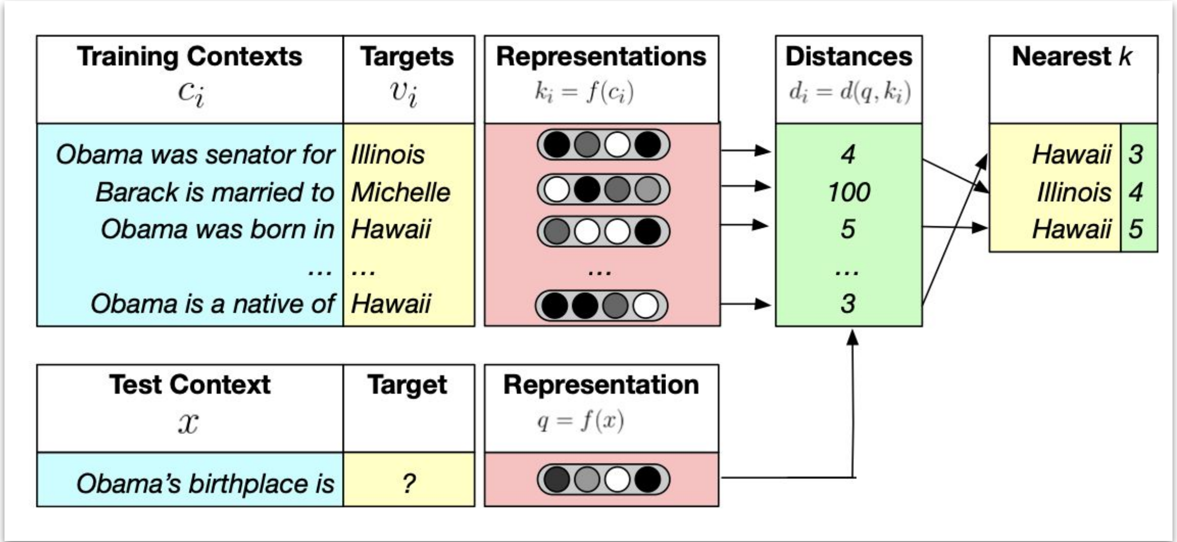


Figure from Khandelwal et. al. 2020

kNN-LM: One Flavor of a Semi-Parametric Language Model

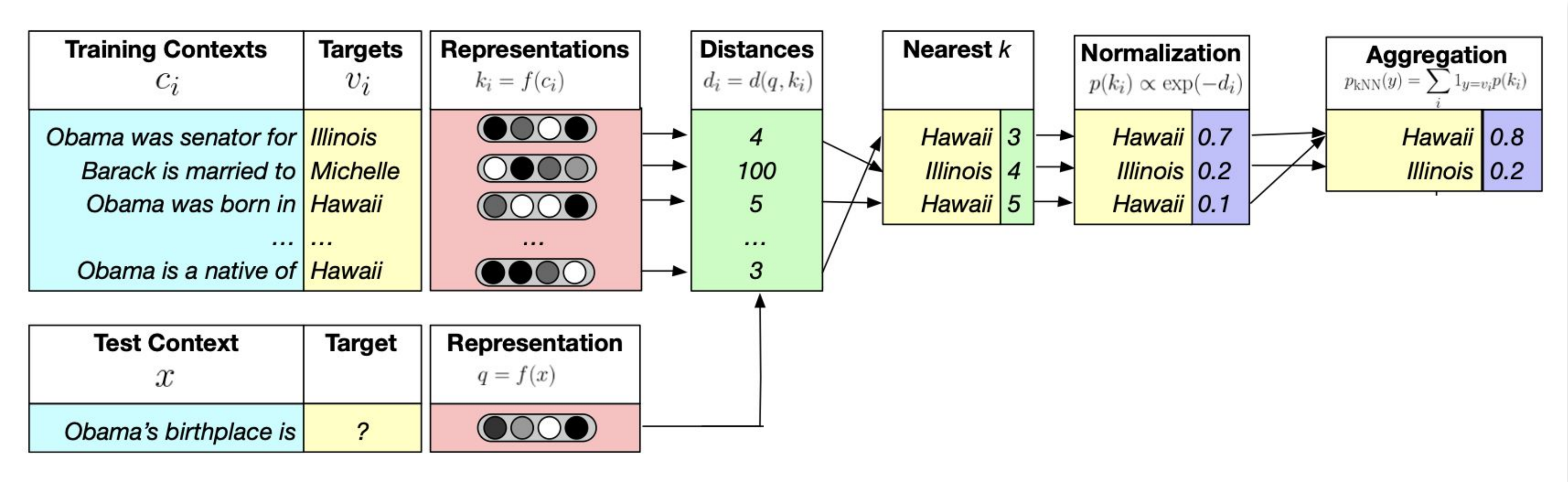


Figure from Khandelwal et. al. 2020

kNN-LM: One Flavor of a Semi-Parametric Language Model

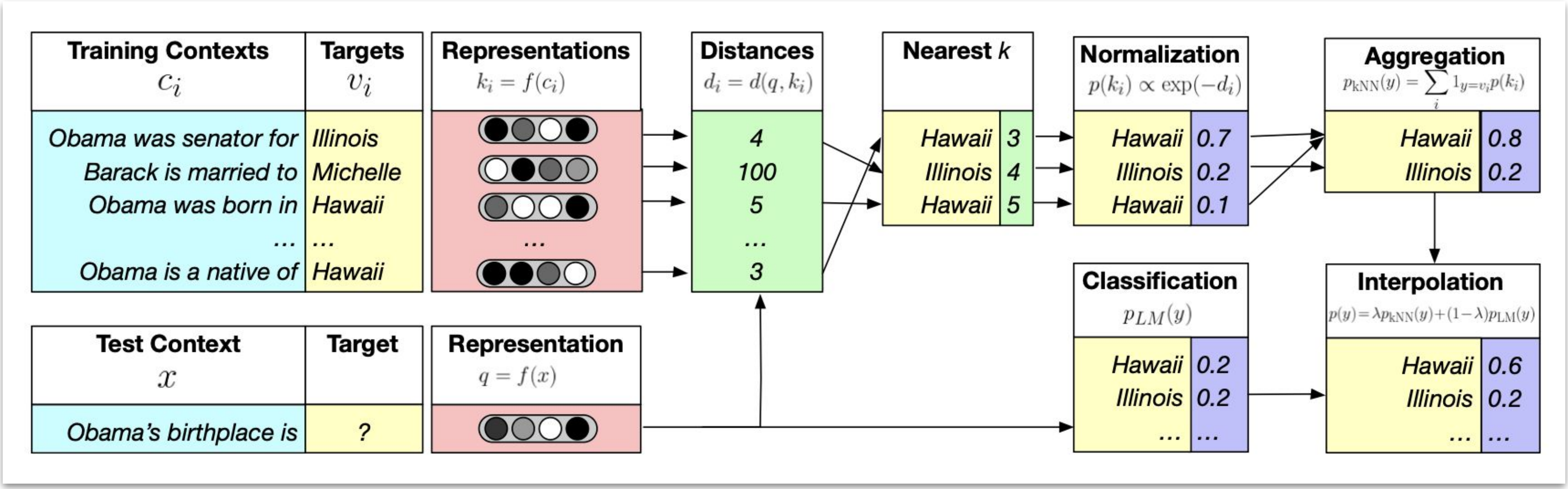


Figure from Khandelwal et. al. 2020

kNN-LM: One Flavor of a Semi-Parametric Language Model

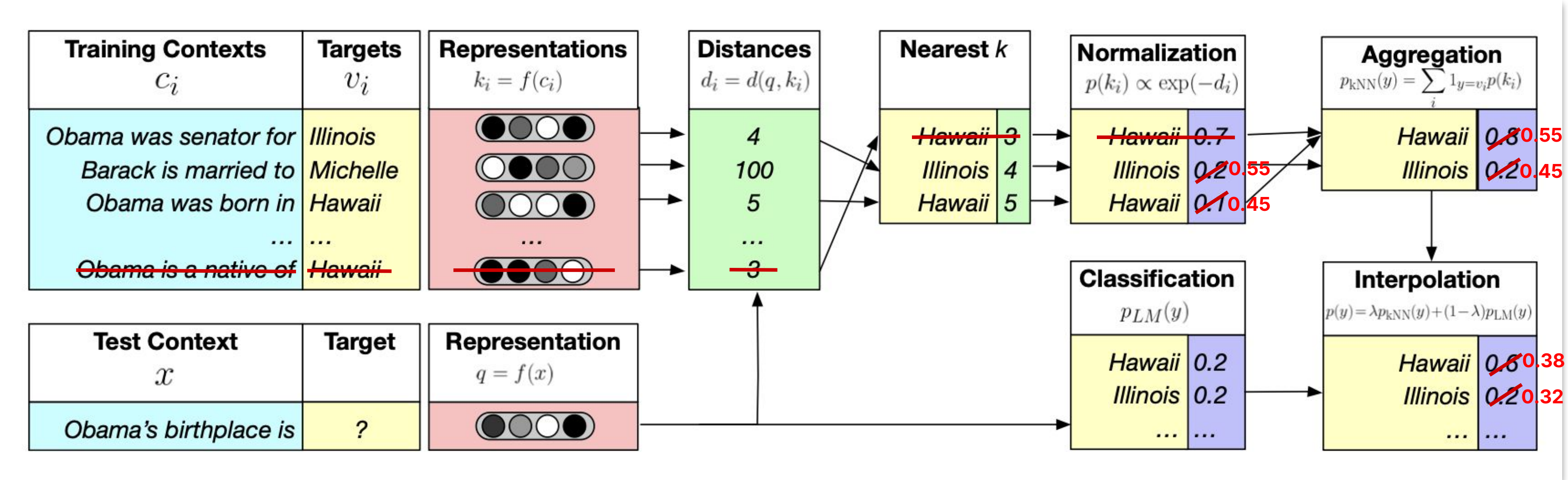
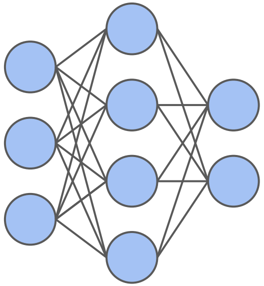


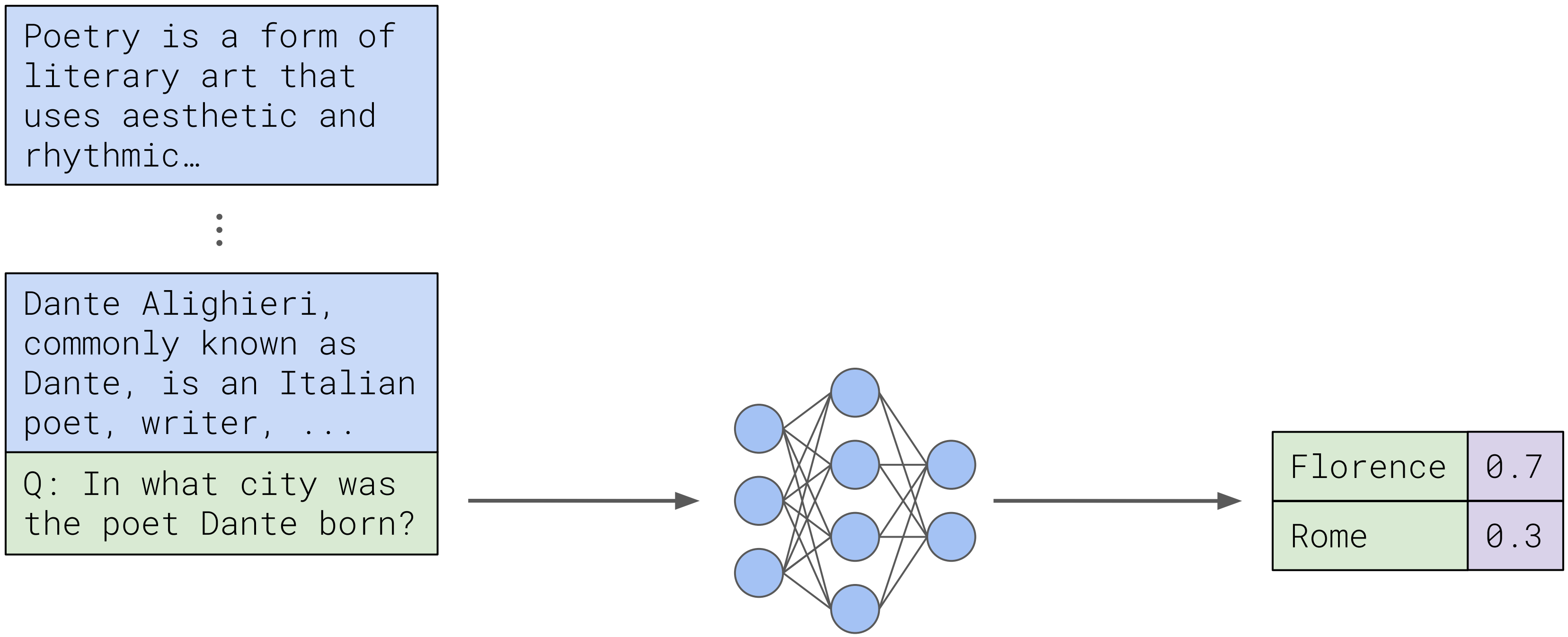
Figure from Khandelwal et. al. 2020

Retrieval Augmented Generation (RAG)

Q: In what city was the poet Dante born?



Retrieval Augmented Generation (RAG)



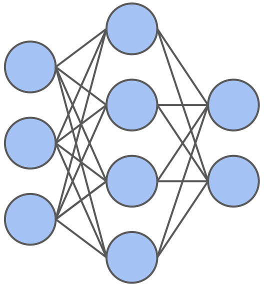
Retrieval Augmented Generation (RAG)

Poetry is a form of literary art that uses aesthetic and rhythmic…

⋮

~~Dante Alighieri, commonly known as Dante, is an Italian poet, writer, ...~~

Q: In what city was the poet Dante born?



Florence	~~0.7~~	0.4
Rome	~~0.3~~	0.6

An interesting research question on incentive-alignment

- If training data contributors were paid proportionally to the counterfactual value of their data, what kind of data are they incentivized to produce?
  - High-attribution → high-quality data?
  - High-attribution adversarial examples?

References

Capabilities and Relevant Training Data

1. Nikhil Kandpal, Eric Wallace, Colin Raffel. *Deduplicating Training Data Mitigates Privacy Risks in Language Models*. 2022a.
2. Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, Chiyuan Zhang. *Quantifying Memorization Across Neural Language Models*. 2022.
3. Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, Sameer Singh. *Impact of Pretraining Term Frequencies on Few-Shot Reasoning*. 2022.
4. Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, Colin Raffel. *Large Language Models Struggle to Learn Long-Tail Knowledge*. 2022b.
5. 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*. 2022.
6. R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, Thomas L. Griffiths. *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve*. 2023.

Influence Functions

7. Pang Wei Koh, Percy Liang. Understanding Black-box Predictions via Influence Functions. 2017.
8. Samyadeep Basu, Philip Pope, Soheil Feizi. Influence Functions in Deep Learning are Fragile. 2020.
9. Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, Roger Grosse. If Influence Functions are the Answer, Then What is the Question? 2022.

Semi-Parametric Language Models

10. Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis. Generalization Through Memorization: Nearest Neighbor Language Models. 2020.
11. Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer.SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore

Retrieval Augmented Generation

12. 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. 2020.
13. Benjamin Cohen-Wang, Harhsay Shah, Kristian Georgiev, Aleksander Madry. ContextCite: Attributing Model Generation to Context. 2024.