

---

## About Me

- Hi, I'm Nikhil
- Research Interests
  - Relating training data to model behavior
  - Data privacy
  - ML security
- Non-Research Interests
  - Chess
  - Basketball





## Collaborators



Haikang Deng  
UNC Chapel Hill



Adam Roberts  
Google Brain



Eric Wallace  
UC Berkeley



Colin Raffel  
UNC Chapel Hill



# Large Language Models Struggle to Learn Long-Tail Information

Nikhil Kandpal



# **ML Models Work and Break Unexpectedly**





# ML Models Work and Break Unexpectedly


```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```


# ML Models Work and Break Unexpectedly




|   |                                |   |                         |
|---|--------------------------------|---|-------------------------|
| 1 | Translate English to French:   | ← | <i>task description</i> |
| 2 | sea otter => loutre de mer     | ← | <i>examples</i>         |
| 3 | peppermint => menthe poivrée   | ← |                         |
| 4 | plush girafe => girafe peluche | ← |                         |
| 5 | cheese => .....                | ← | <i>prompt</i>           |

**TECH** USA TODAY Tech   
@usatodaytech · [Follow](#) 

Watch a Tesla in "Smart Summon" mode run into a \$3.5 million private jet.



7:54 PM · Apr 25, 2022 

 109  Reply  Share

[Read 21 replies](#)



# What Can Machine Learning Models Do?

---

# What Can Machine Learning Models Do?

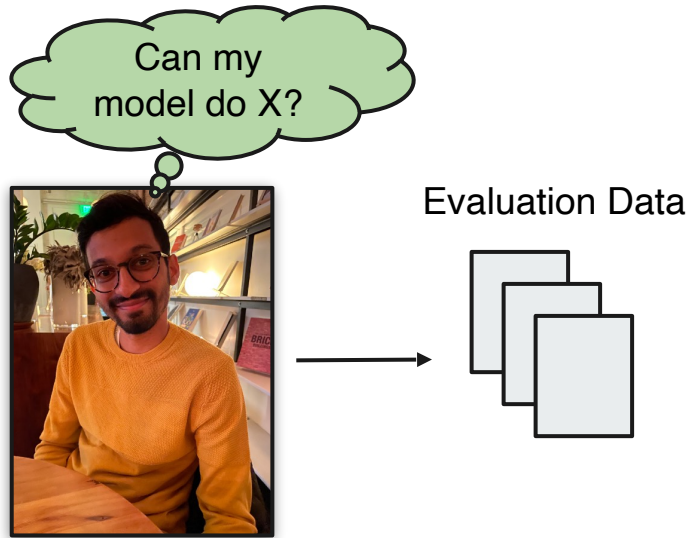
Can my  
model do X?





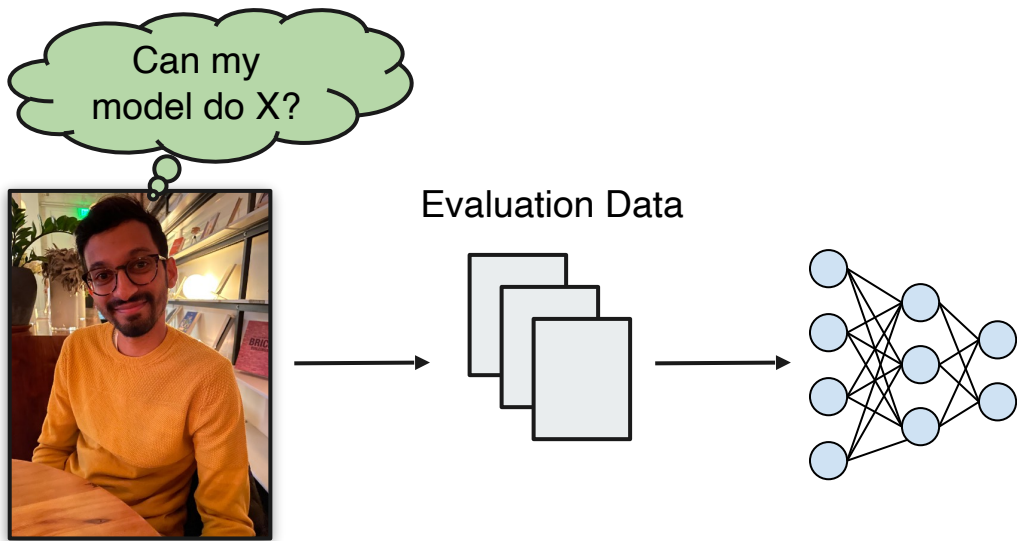


# What Can Machine Learning Models Do?

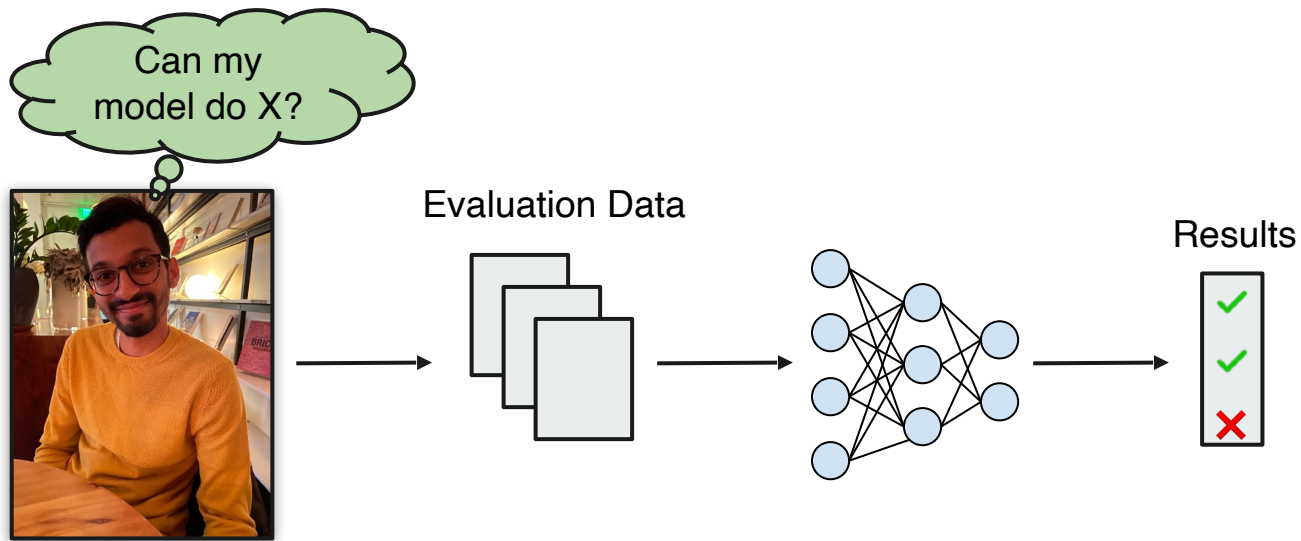




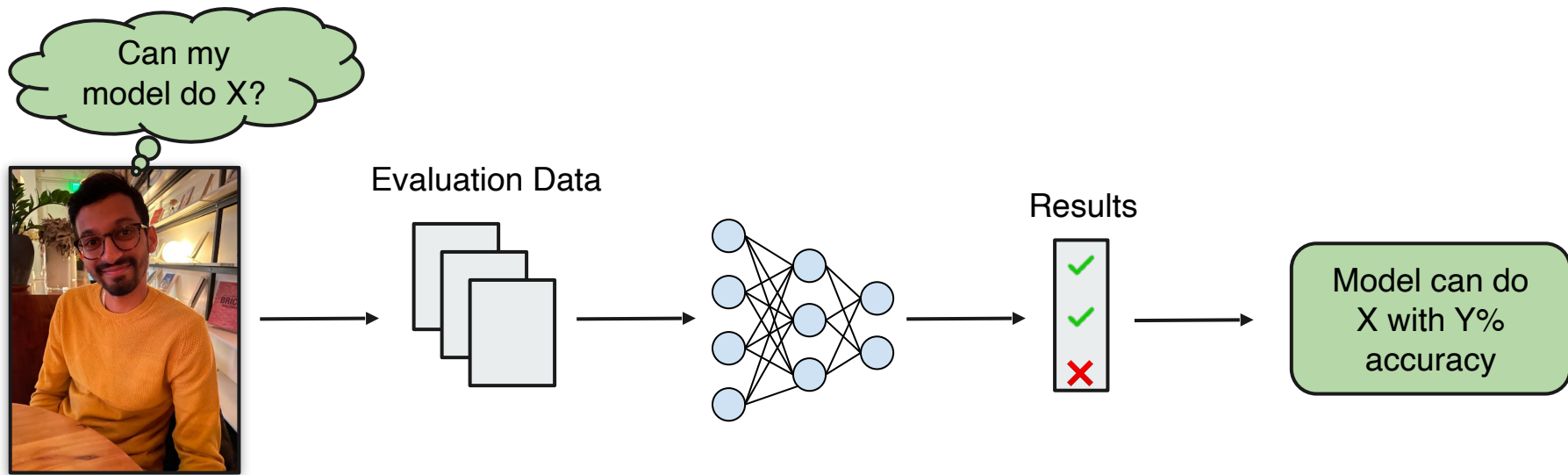
# What Can Machine Learning Models Do?



# What Can Machine Learning Models Do?



# What Can Machine Learning Models Do?





# **WHY Can Machine Learning Models Do Certain Things?**

---

# WHY Can Machine Learning Models Do Certain Things?

Why can my  
model do X?



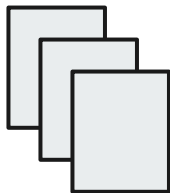
---

# WHY Can Machine Learning Models Do Certain Things?

Why can my  
model do X?

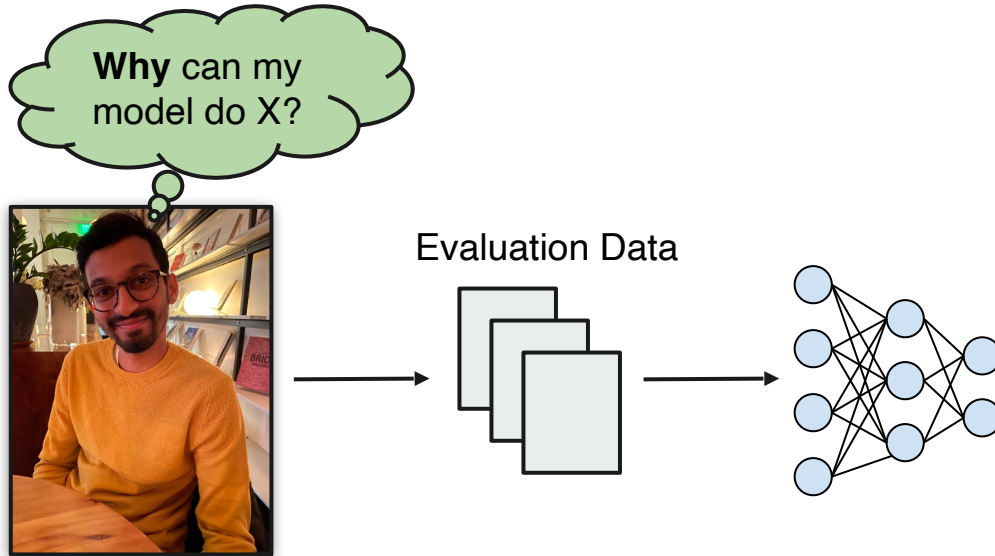


Evaluation Data



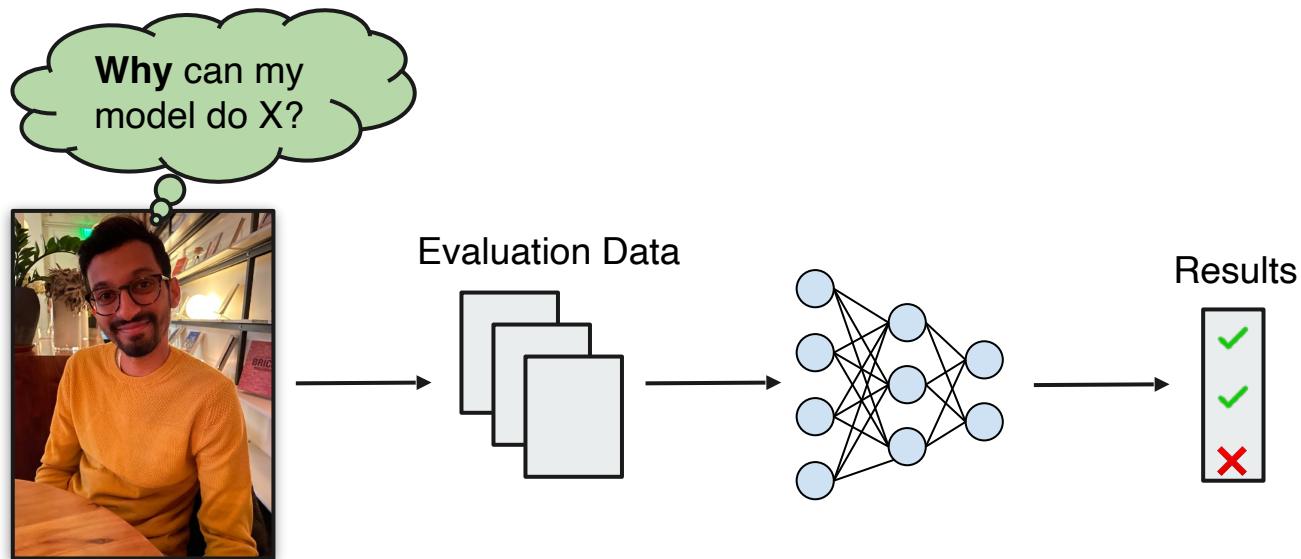


# WHY Can Machine Learning Models Do Certain Things?

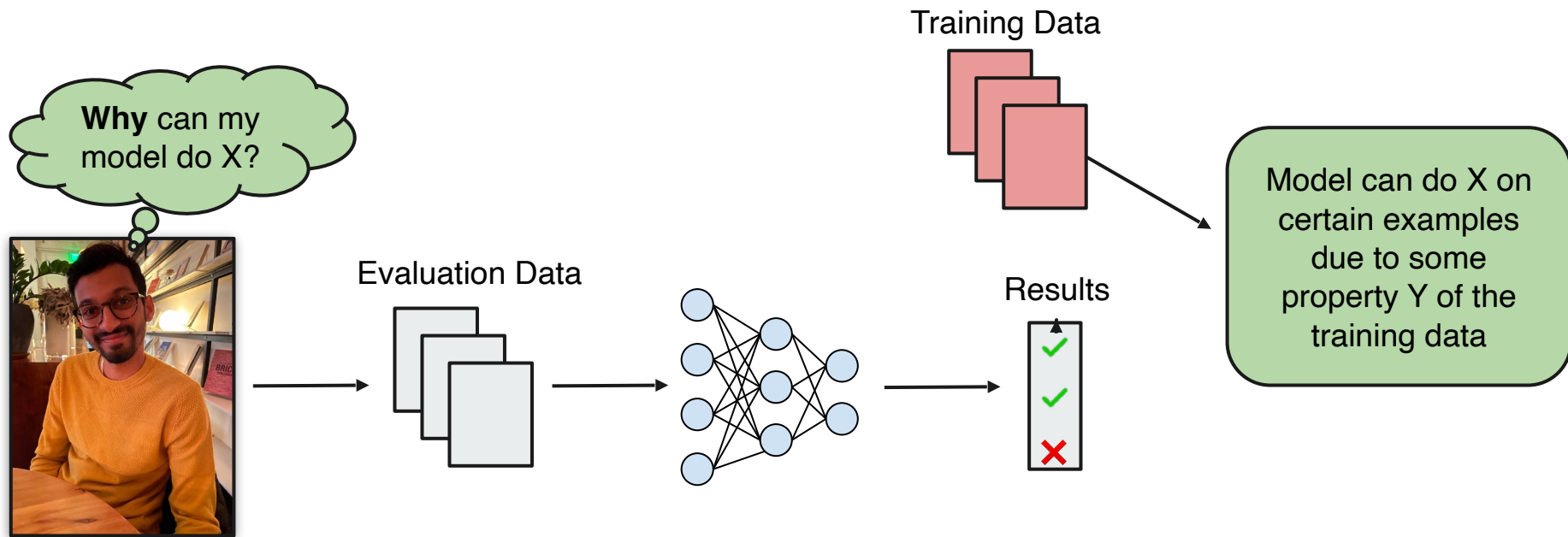


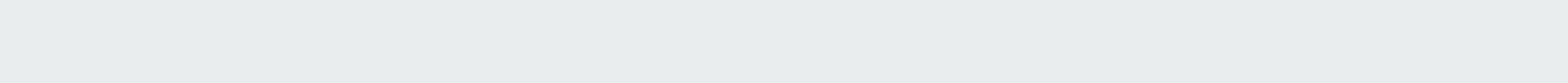


# WHY Can Machine Learning Models Do Certain Things?



# WHY Can Machine Learning Models Do Certain Things?







**Question:** With this methodology, can we use the training data to gain insight into why machine learning models learn certain behaviors?



**Question:** With this methodology, can we use the training data to gain insight into why machine learning models learn certain behaviors?

**Short Answer:** Yes, frequency statistics of the training data impact the information that a model learns, making rare, long-tail information difficult to capture



# Experimental Setting



# Experimental Setting

- Focus on large language models



## Experimental Setting

- Focus on large language models
- Analyze memorization and factoid knowledge learning





## Experimental Setting

- Focus on large language models
- Analyze memorization and factoid knowledge learning
- Study behavior of pre-trained models and pre-training datasets



## Outline for the rest of the talk

1. Background on Language Models
2. Eidetic Memorization in Language Models
3. Knowledge Learning in Language Models



# Language Models



# Language Models

- How are language models pre-trained?



# Language Models

- How are language models pre-trained?
  - Large-scale web text datasets



# Language Models

- How are language models pre-trained?
  - Large-scale web text datasets
  - Language modeling objective



# Language Models

- How are language models pre-trained?
  - Large-scale web text datasets
  - Language modeling objective



# Language Models

- How are language models pre-trained?
  - Large-scale web text datasets
  - Language modeling objective
  
- What are some things language models can do?





# Language Models

- How are language models pre-trained?
  - Large-scale web text datasets
  - Language modeling objective
  
- What are some things language models can do?
  - Unconditional generation



# Language Models

- How are language models pre-trained?
  - Large-scale web text datasets
  - Language modeling objective
  
- What are some things language models can do?
  - Unconditional generation
  - In-context learning



# Eidetic Memorization in Language Models



# Eidetic Memorization in Language Models

- Behavior: Eidetic Memorization



# Eidetic Memorization in Language Models

- Behavior: Eidetic Memorization
  - The ability to **perfectly recall and generate** text from the training data



# Eidetic Memorization in Language Models

- Behavior: Eidetic Memorization
  - The ability to **perfectly recall and generate** text from the training data



# Eidetic Memorization in Language Models

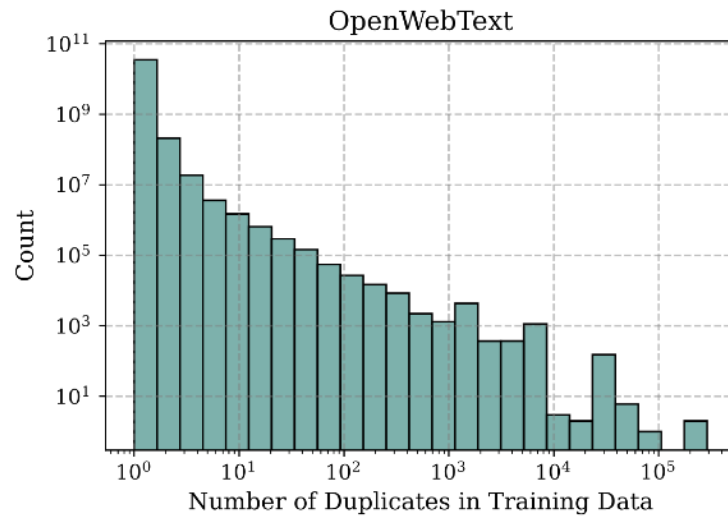
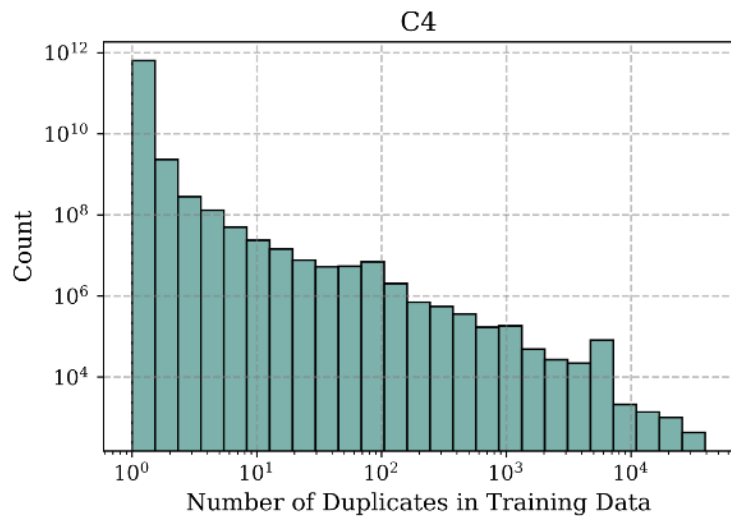
- Behavior: Eidetic Memorization
  - The ability to **perfectly recall and generate** text from the training data
- Question: How does the number of times a piece of text appears in the training data impact how often a language model generates that text?



# Duplicated Text in Pre-Training Datasets



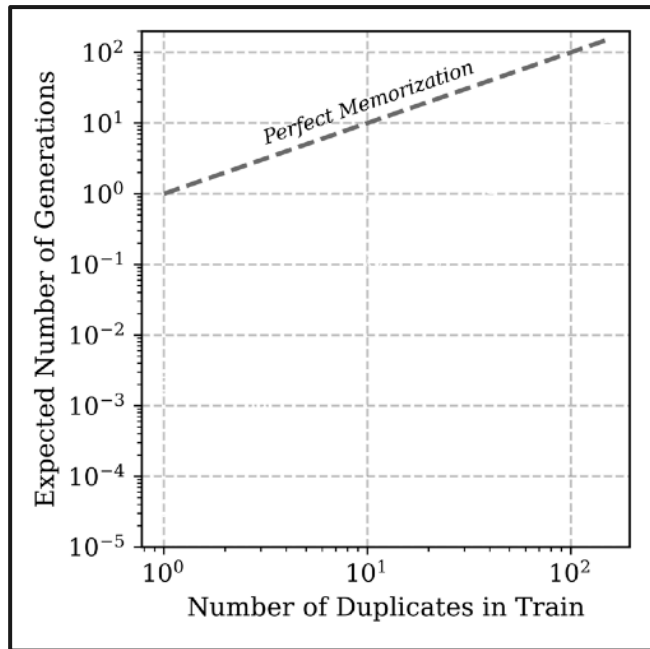
# Duplicated Text in Pre-Training Datasets



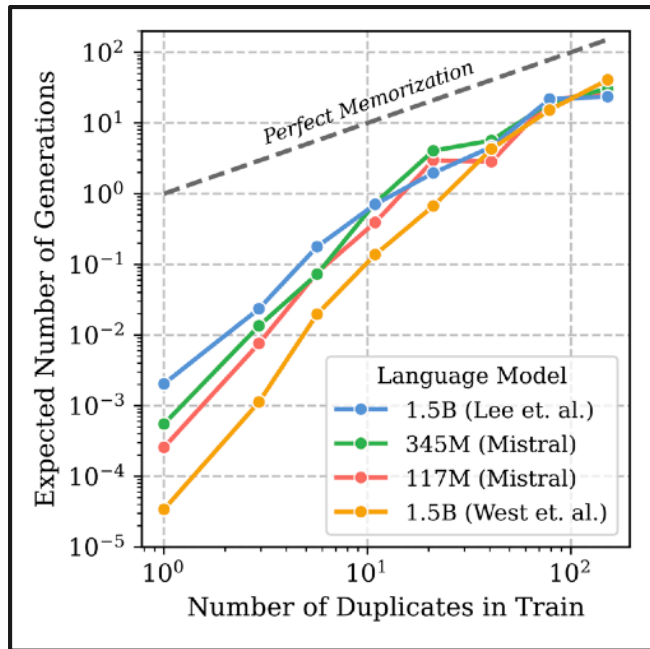


# Relationship Between Training Data and Generation

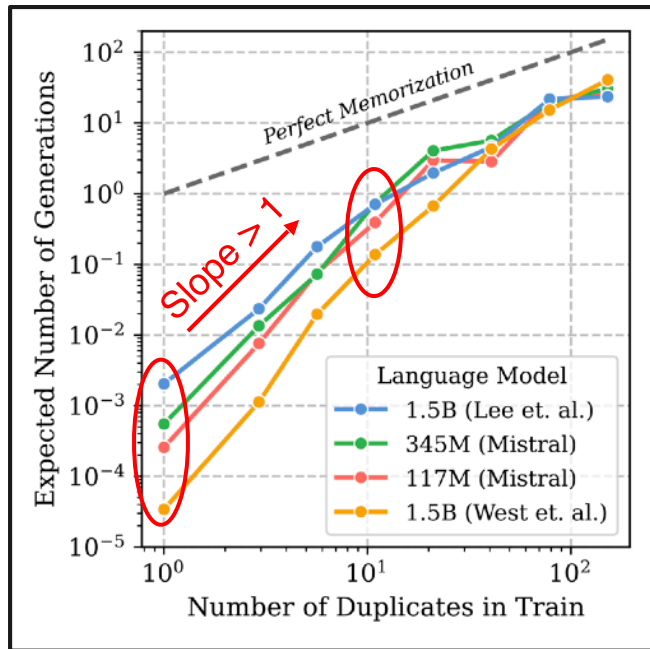
# Relationship Between Training Data and Generation



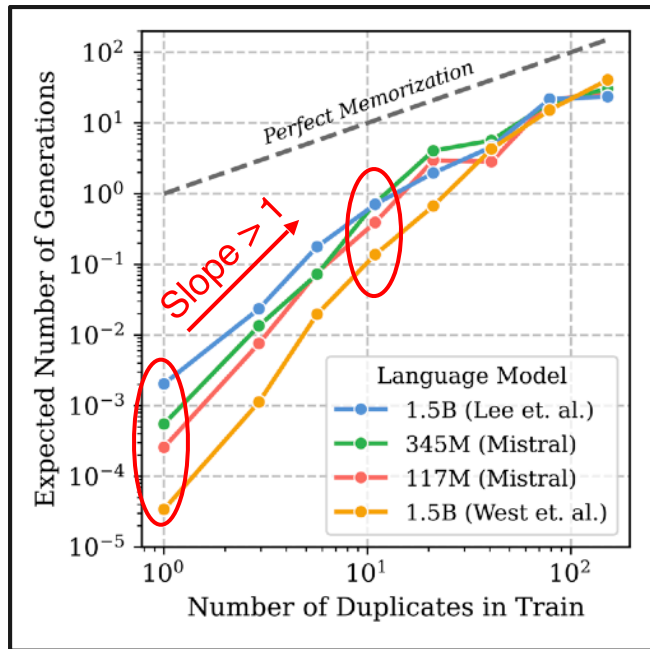
# Relationship Between Training Data and Generation



# Relationship Between Training Data and Generation



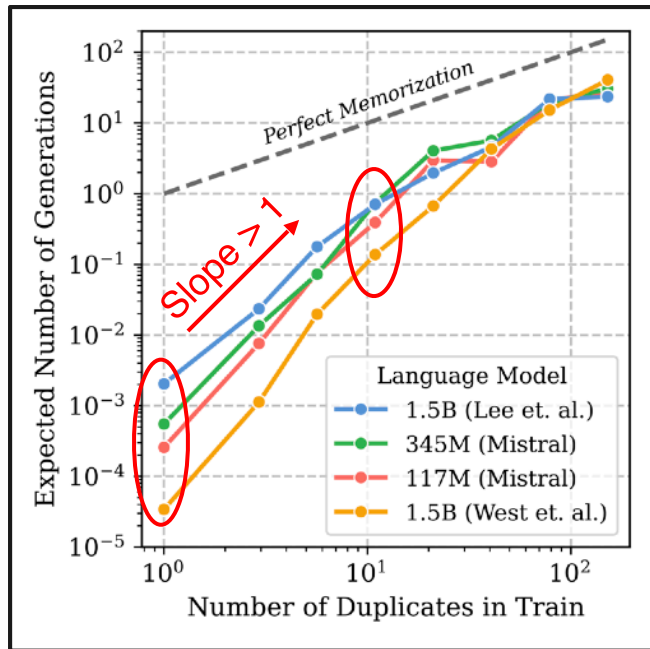
# Relationship Between Training Data and Generation



## Observation #1

Eidetic memorization rate is superlinearly related to the number of times a sequence appears in the training data

# Relationship Between Training Data and Generation

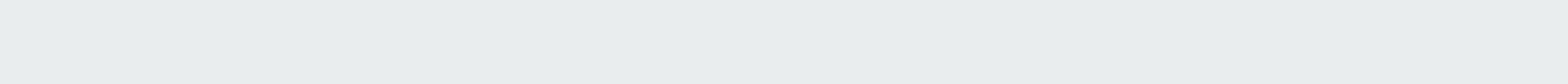


## Observation #1

Eidetic memorization rate is superlinearly related to the number of times a sequence appears in the training data

## Observation #2

Language Models are uncalibrated – generation frequency does not reflect training data frequency







**Conclusion:** Pre-training data text frequency is related to the rate at which language models generate that text **verbatim**



**Conclusion:** Pre-training data text frequency is related to the rate at which language models generate that text  
**verbatim**

**Question:** Is this true for more interesting behavior than verbatim regeneration (e.g., knowledge learning)?



# Knowledge Learning in Language Models



# Knowledge Learning in Language Models

- Behavior: Knowledge Learning



# Knowledge Learning in Language Models

- Behavior: Knowledge Learning
  - The ability to **correctly answer questions** about a piece of knowledge



# Knowledge Learning in Language Models

- Behavior: Knowledge Learning
  - The ability to **correctly answer questions** about a piece of knowledge



# Knowledge Learning in Language Models

- Behavior: Knowledge Learning
  - The ability to **correctly answer questions** about a piece of knowledge
- Question: How does the number of times a fact appears in the training data impact how well a language model learns that fact?



# Evaluating Whether Language Models Know Facts





# Evaluating Whether Language Models Know Facts

Fact: Dante Alighieri was born in Florence



# Evaluating Whether Language Models Know Facts

Fact: Dante Alighieri was born in Florence

Q: What is the capital of France?  
A: Paris

Q: At what temperature does water boil?  
A: 100C

Q: In what city was the poet Dante born?  
A:



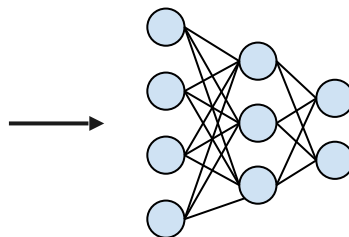
# Evaluating Whether Language Models Know Facts

Fact: Dante Alighieri was born in Florence

Q: What is the capital of France?  
A: Paris

Q: At what temperature does water boil?  
A: 100C

Q: In what city was the poet Dante born?  
A:





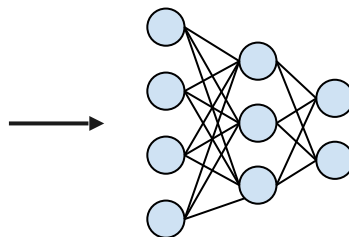
# Evaluating Whether Language Models Know Facts

Fact: Dante Alighieri was born in Florence

Q: What is the capital of France?  
A: Paris

Q: At what temperature does water boil?  
A: 100C

Q: In what city was the poet Dante born?  
A:



Florence

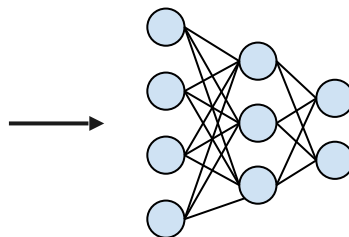
# Evaluating Whether Language Models Know Facts

Fact: Dante Alighieri was born in Florence

Q: What is the capital of France?  
A: Paris

Q: At what temperature does water boil?  
A: 100C

Q: In what city was the poet Dante born?  
A:



Florence



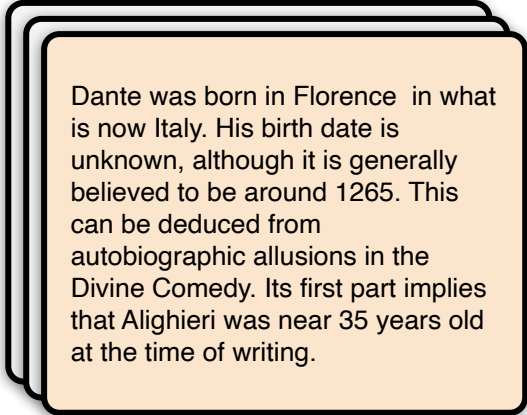
# Approximating Fact Frequency in Pre-Training Data



# Approximating Fact Frequency in Pre-Training Data

---

## Pre-training Documents



Dante was born in Florence in what is now Italy. His birth date is unknown, although it is generally believed to be around 1265. This can be deduced from autobiographic allusions in the Divine Comedy. Its first part implies that Alighieri was near 35 years old at the time of writing.

# Approximating Fact Frequency in Pre-Training Data

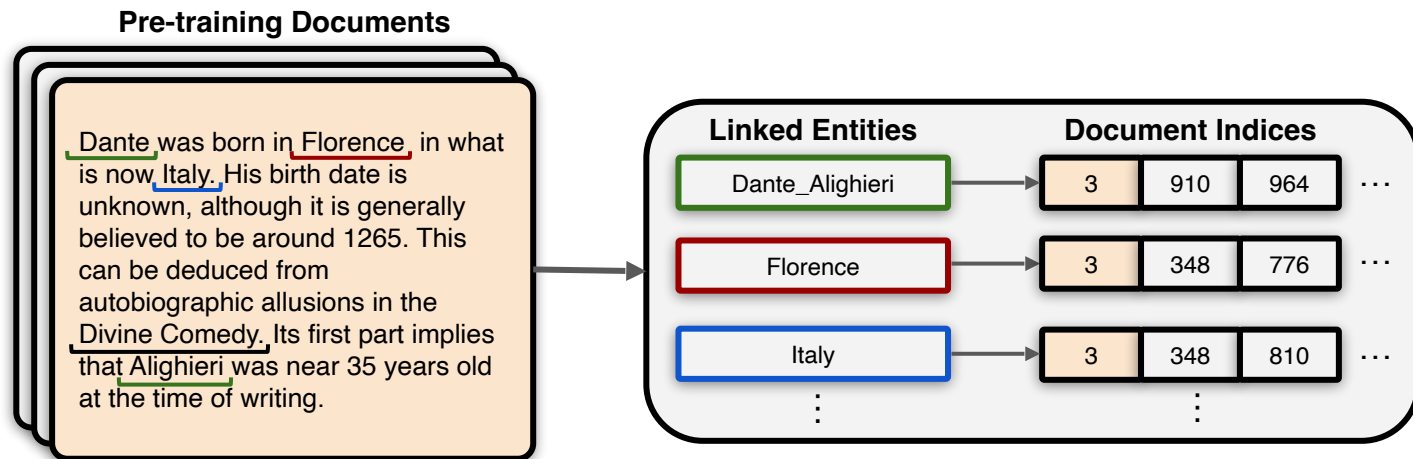
---

## Pre-training Documents

Dante was born in Florence, in what is now Italy. His birth date is unknown, although it is generally believed to be around 1265. This can be deduced from autobiographic allusions in the Divine Comedy. Its first part implies that Alighieri was near 35 years old at the time of writing.



# Approximating Fact Frequency in Pre-Training Data



# Approximating Fact Frequency in Pre-Training Data

## Pre-training Documents

Dante was born in Florence, in what is now Italy. His birth date is unknown, although it is generally believed to be around 1265. This can be deduced from autobiographic allusions in the Divine Comedy. Its first part implies that Alighieri was near 35 years old at the time of writing.

## Linked Entities

Dante\_Alighieri

Florence

Italy

⋮

## Document Indices

3 910 964 ...

3 348 776 ...

3 348 810 ...

⋮

## Question Answering Examples

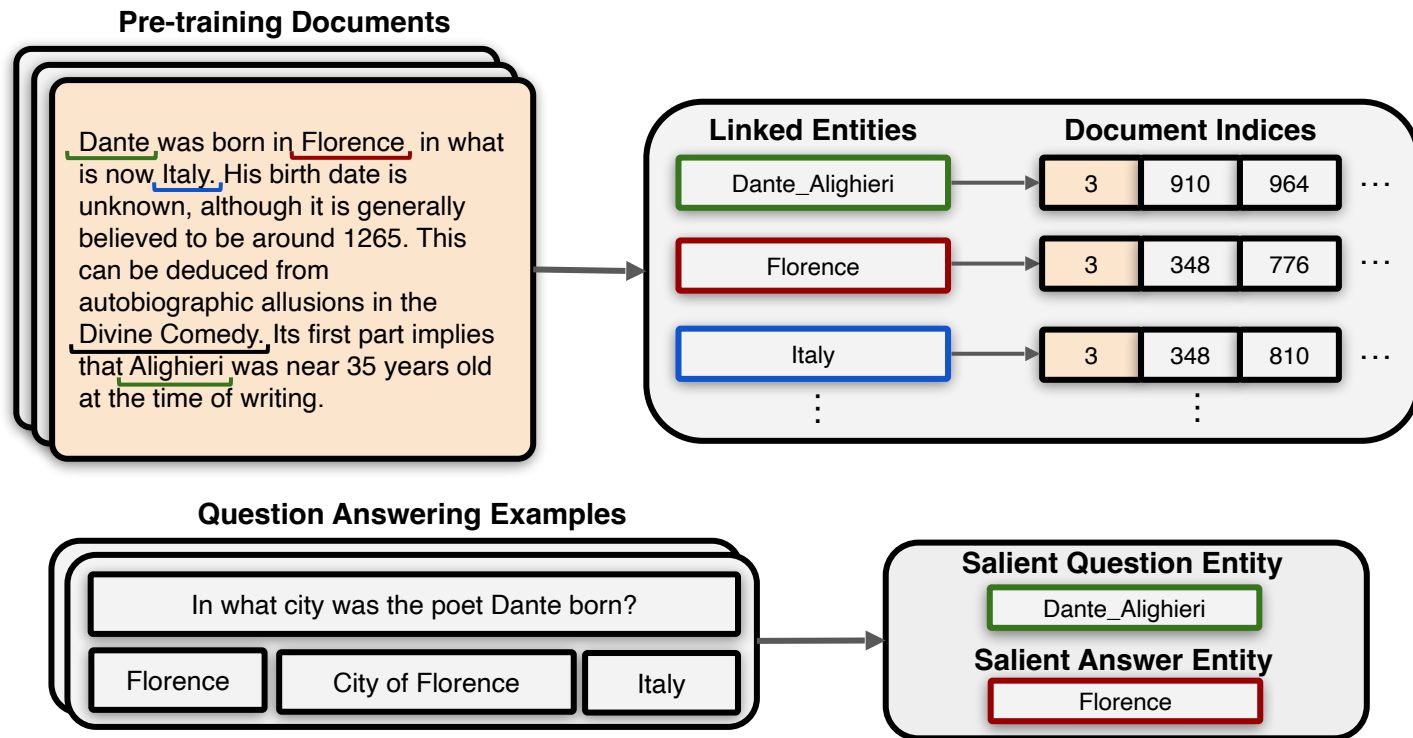
In what city was the poet Dante born?

Florence

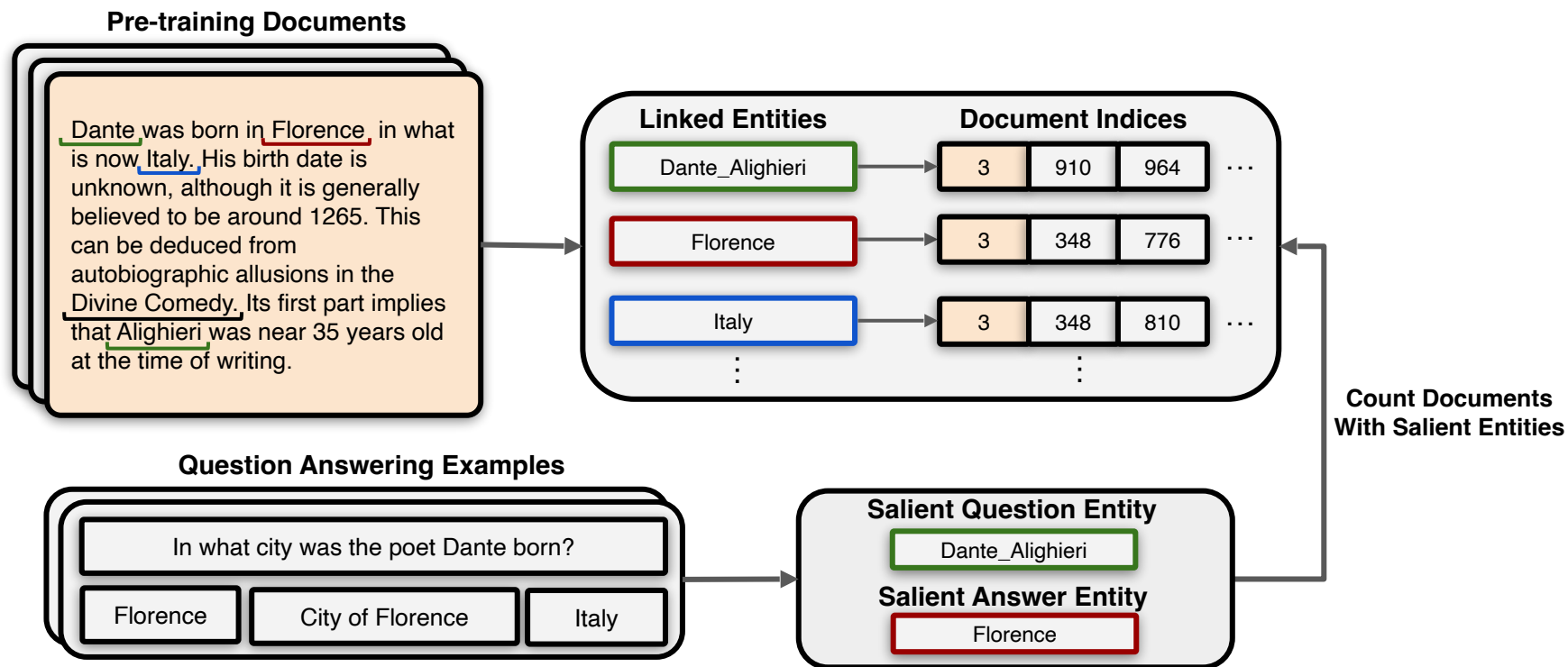
City of Florence

Italy

# Approximating Fact Frequency in Pre-Training Data



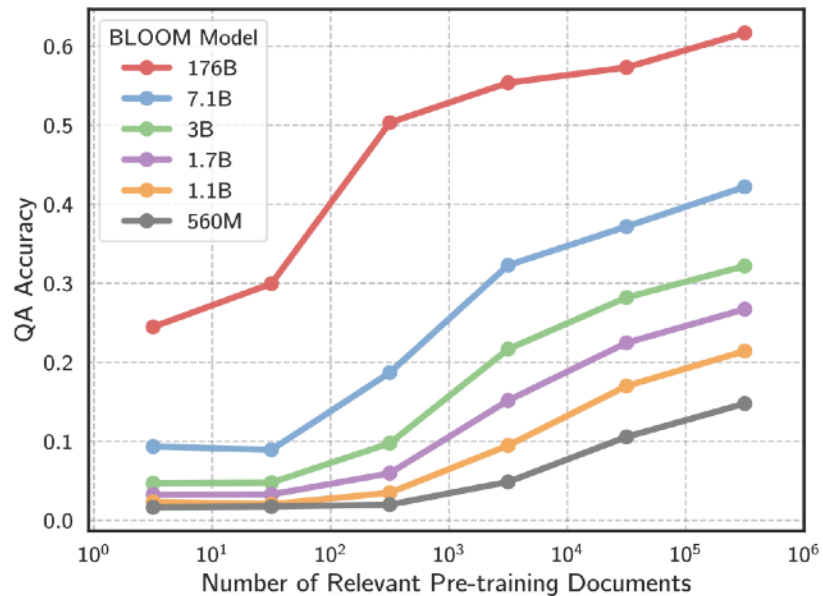
# Approximating Fact Frequency in Pre-Training Data



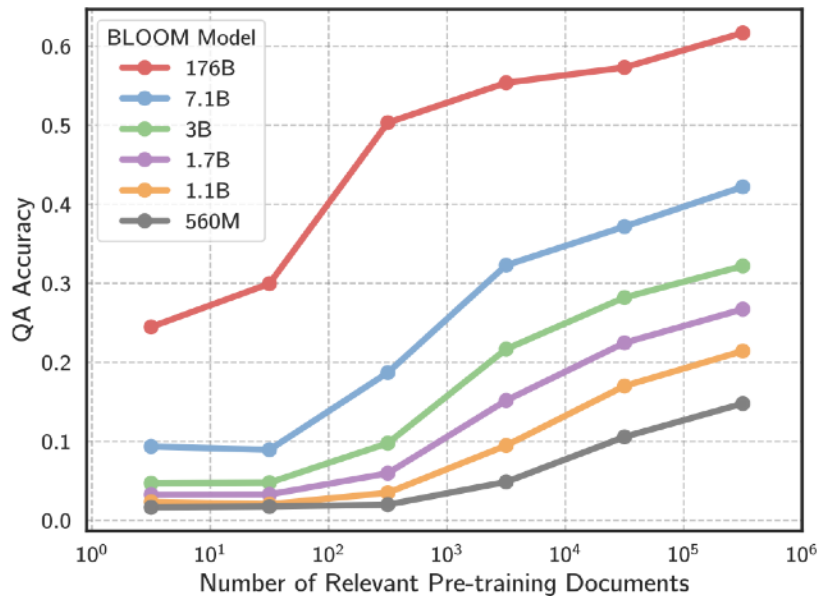


# Relationship Between Training Data and Fact Learning

# Relationship Between Training Data and Fact Learning



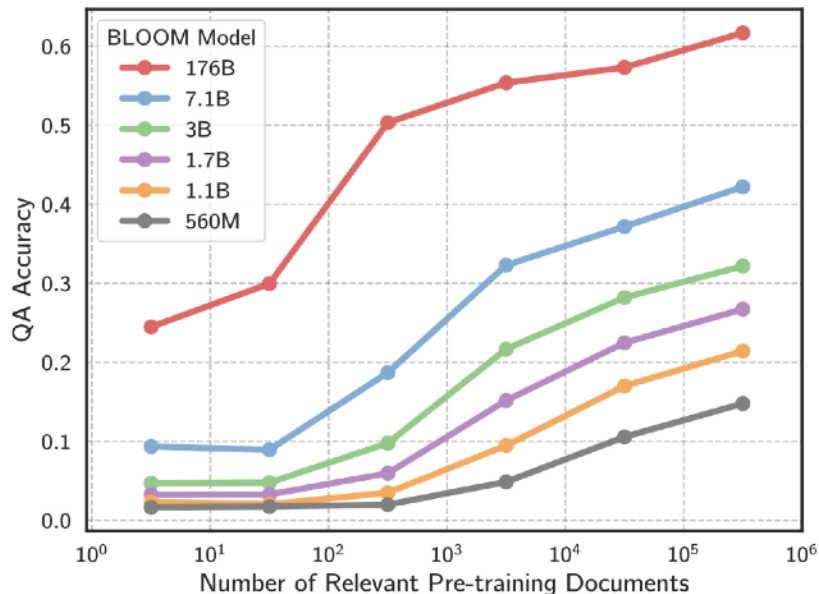
# Relationship Between Training Data and Fact Learning



## Observation #1

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

# Relationship Between Training Data and Fact Learning



## 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

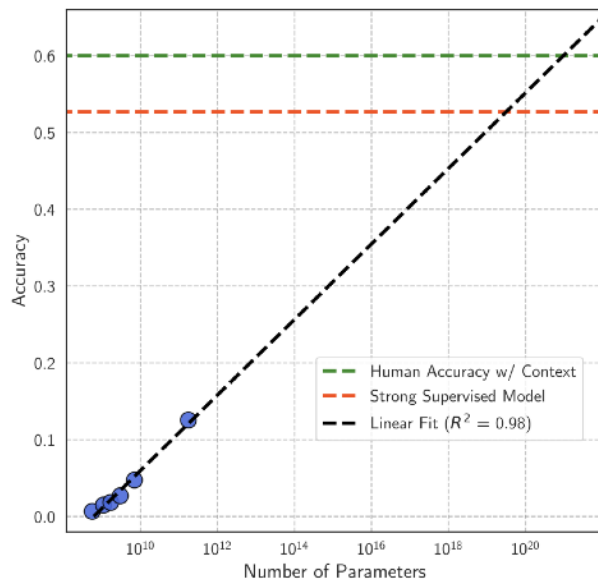




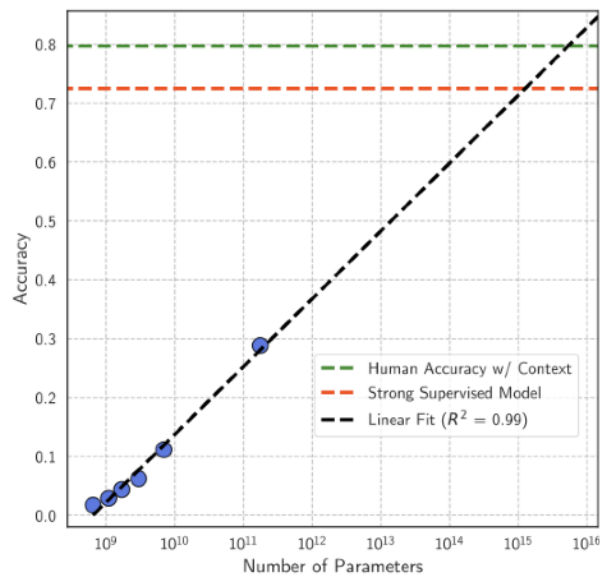
# How Large do Models Need to be to Learn Rare Facts?

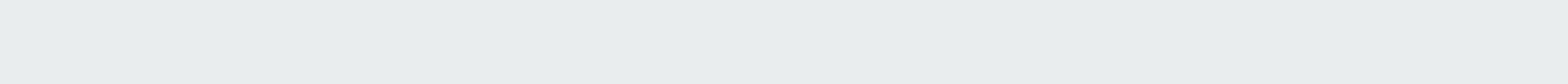
# How Large do Models Need to be to Learn Rare Facts?

## Natural Questions Rare Fact Accuracy



## TriviaQA Rare Fact Accuracy







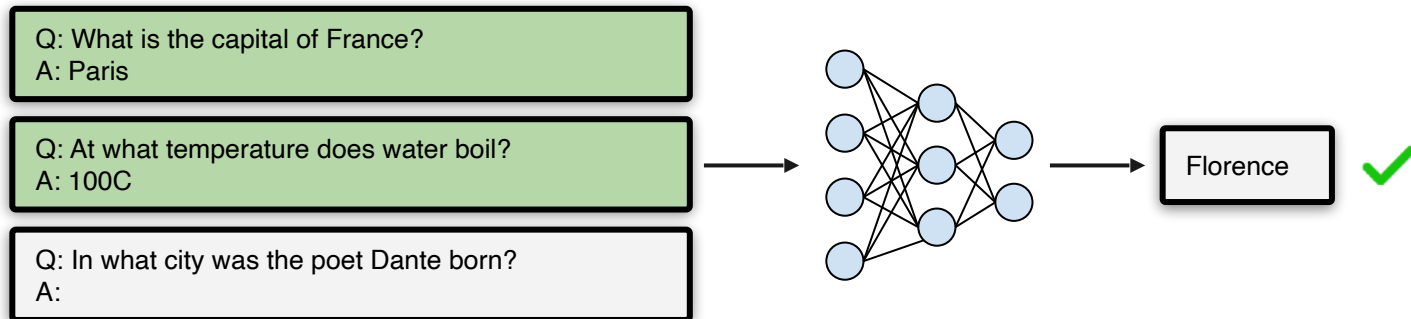
**Conclusion:** A language model's acquisition of a fact is heavily dependent on how many times it is trained on the fact



**Conclusion:** A language model's acquisition of a fact is heavily dependent on how many times it is trained on the fact

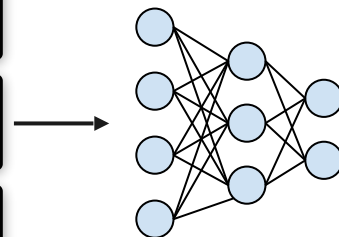
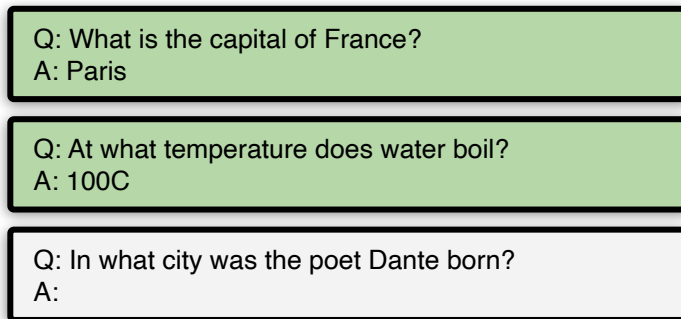
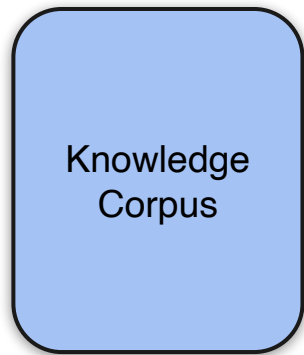
**Question:** Can we remove this dependence on training data fact frequency?

# Retrieval Augmentation



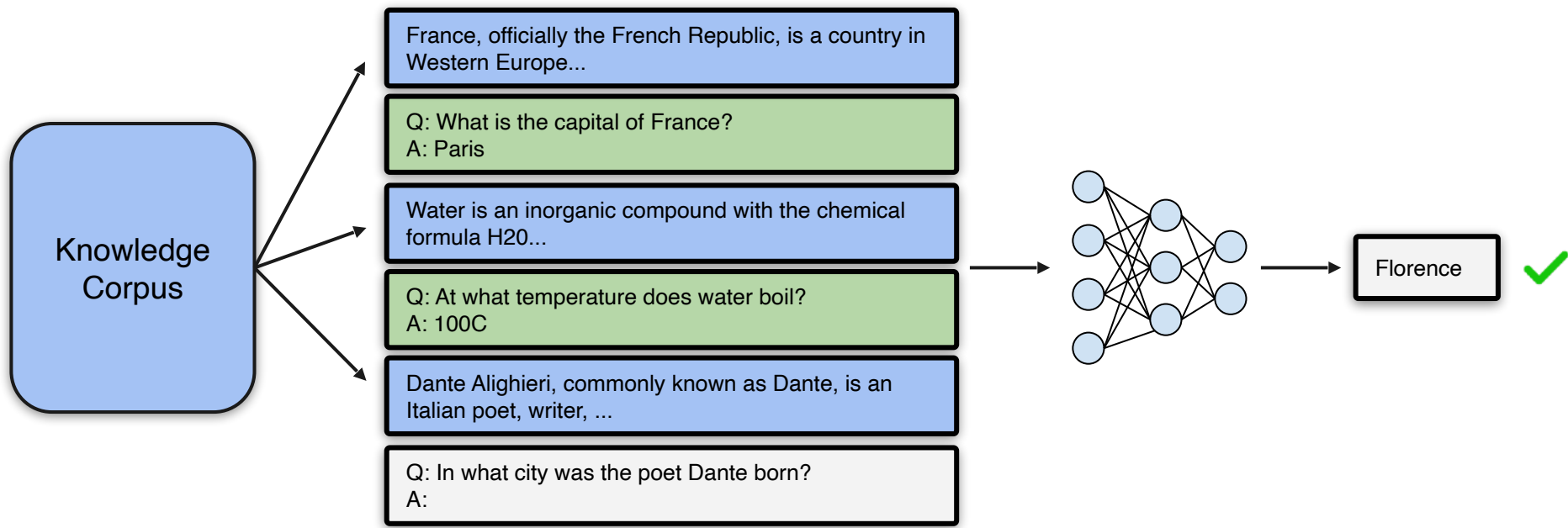


# Retrieval Augmentation





# Retrieval Augmentation



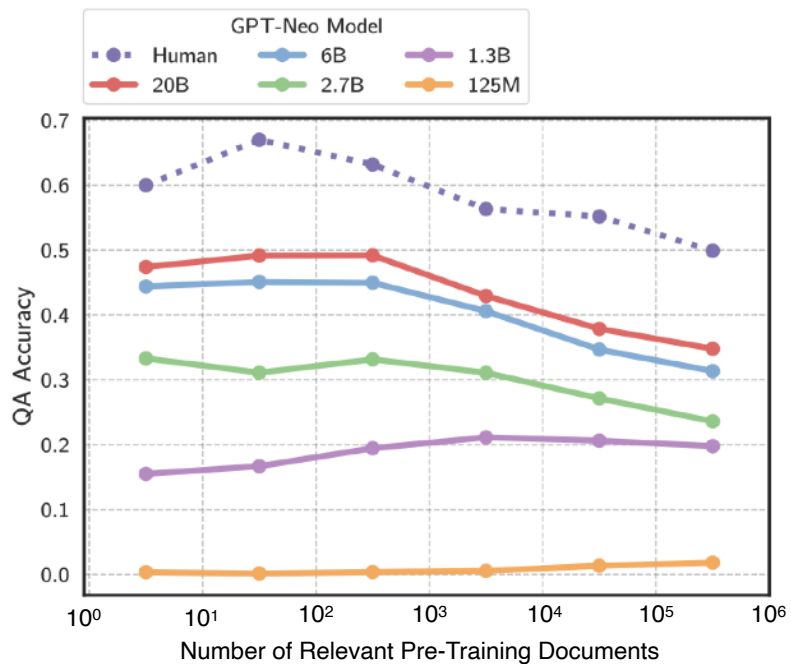




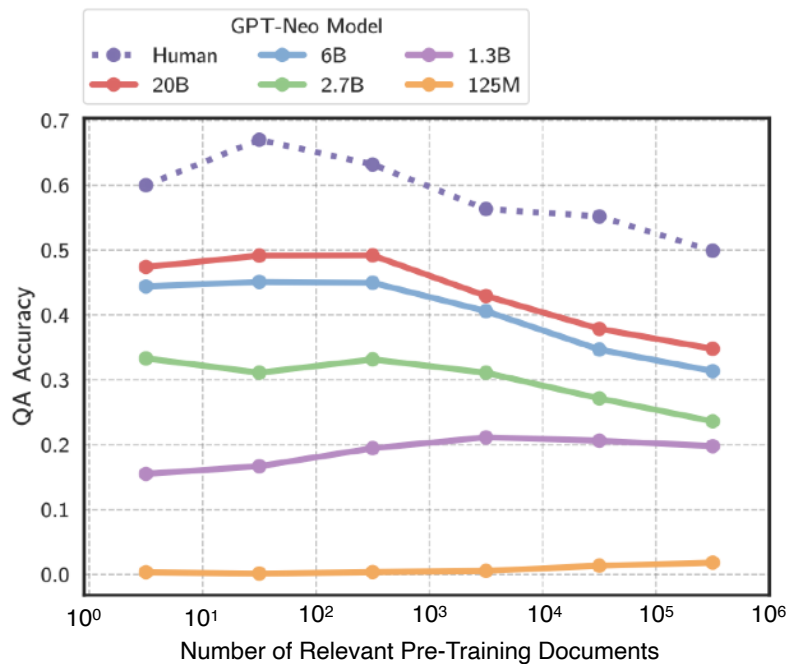
# Retrieval Augmentation

$10^0$      $10^1$      $10^2$      $10^3$      $10^4$      $10^5$      $10^6$   
Number of Relevant Pre-Training Documents

# Retrieval Augmentation



# Retrieval Augmentation



## Observation #1

Retrieval Augmented Language Models don't require facts to be in the training data to do knowledge-intensive tasks



# Takeaways



## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly



## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly
  - Eidetic memorization



## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly
  - Eidetic memorization
  - Fact learning



## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly
  - Eidetic memorization
  - Fact learning
- Model scaling may be inefficient





## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly
  - Eidetic memorization
  - Fact learning
- Model scaling may be inefficient
- Refactoring models to reduce the need for recall



## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly
  - Eidetic memorization
  - Fact learning
- Model scaling may be inefficient
- Refactoring models to reduce the need for recall



## Takeaways

- Recall-based behaviors require the information being recalled to be in the training data repeatedly
    - Eidetic memorization
    - Fact learning
  - Model scaling may be inefficient
  - Refactoring models to reduce the need for recall
- 
- General Point: All model behaviors stem from the training data



# Any Questions?

Email: [nkandpa2@cs.unc.edu](mailto:nkandpa2@cs.unc.edu)