

# What do language models learn from language modeling?

Colin Raffel

## Pre-training

The cabs \_\_\_ the same rates as those \_\_\_ by horse-drawn cabs and were \_\_\_ quite popular, \_\_\_ the Prince of Wales (the \_\_\_ King Edward VII) travelled in \_\_\_. The cabs quickly \_\_\_ known as "hummingbirds" for \_\_\_ noise made by their motors and their distinctive black and \_\_\_ livery. Passengers \_\_\_ \_\_\_ the interior fittings were \_\_\_ when compared to \_\_\_ cabs but there \_\_\_ some complaints \_\_\_ the \_\_\_ lighting made them too \_\_\_ to those outside \_\_\_.

charged, used, initially, even,  
future, became, the, yellow,  
reported, that, luxurious,  
horse-drawn, were that,  
internal, conspicuous, cab

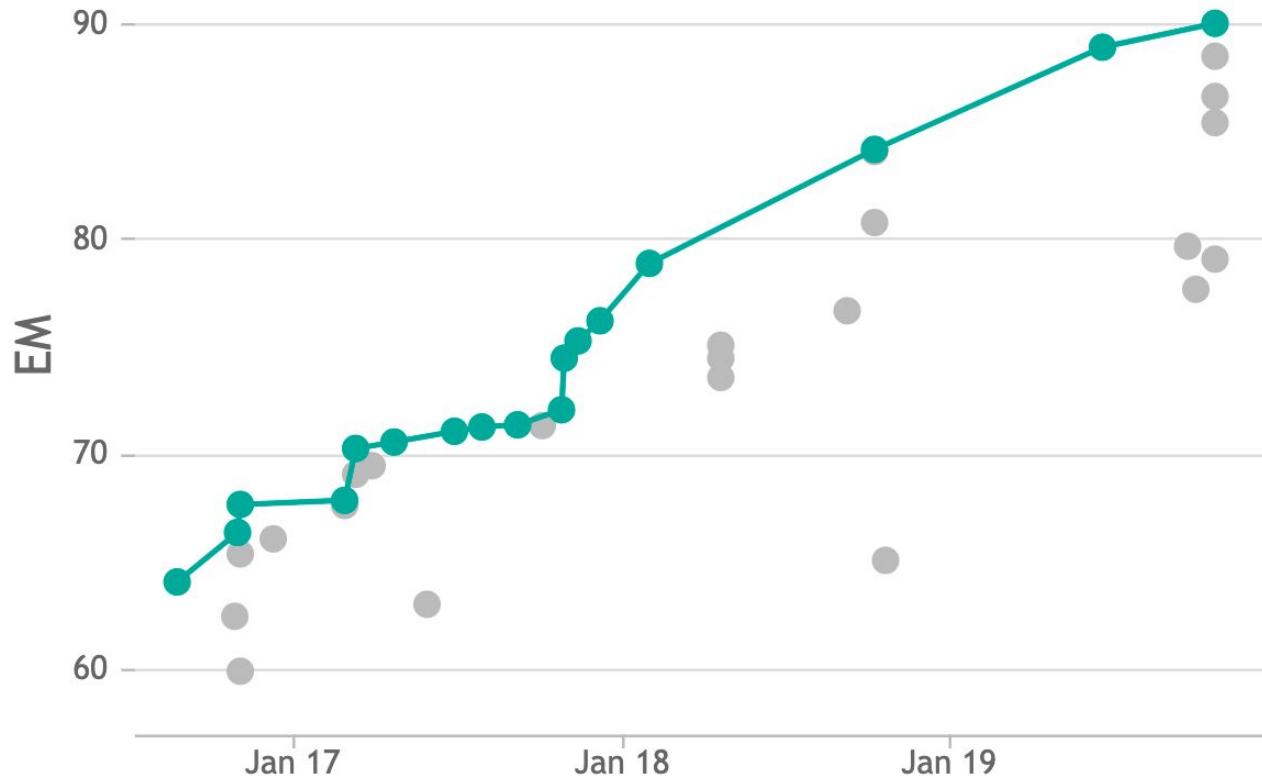


## Adaptation

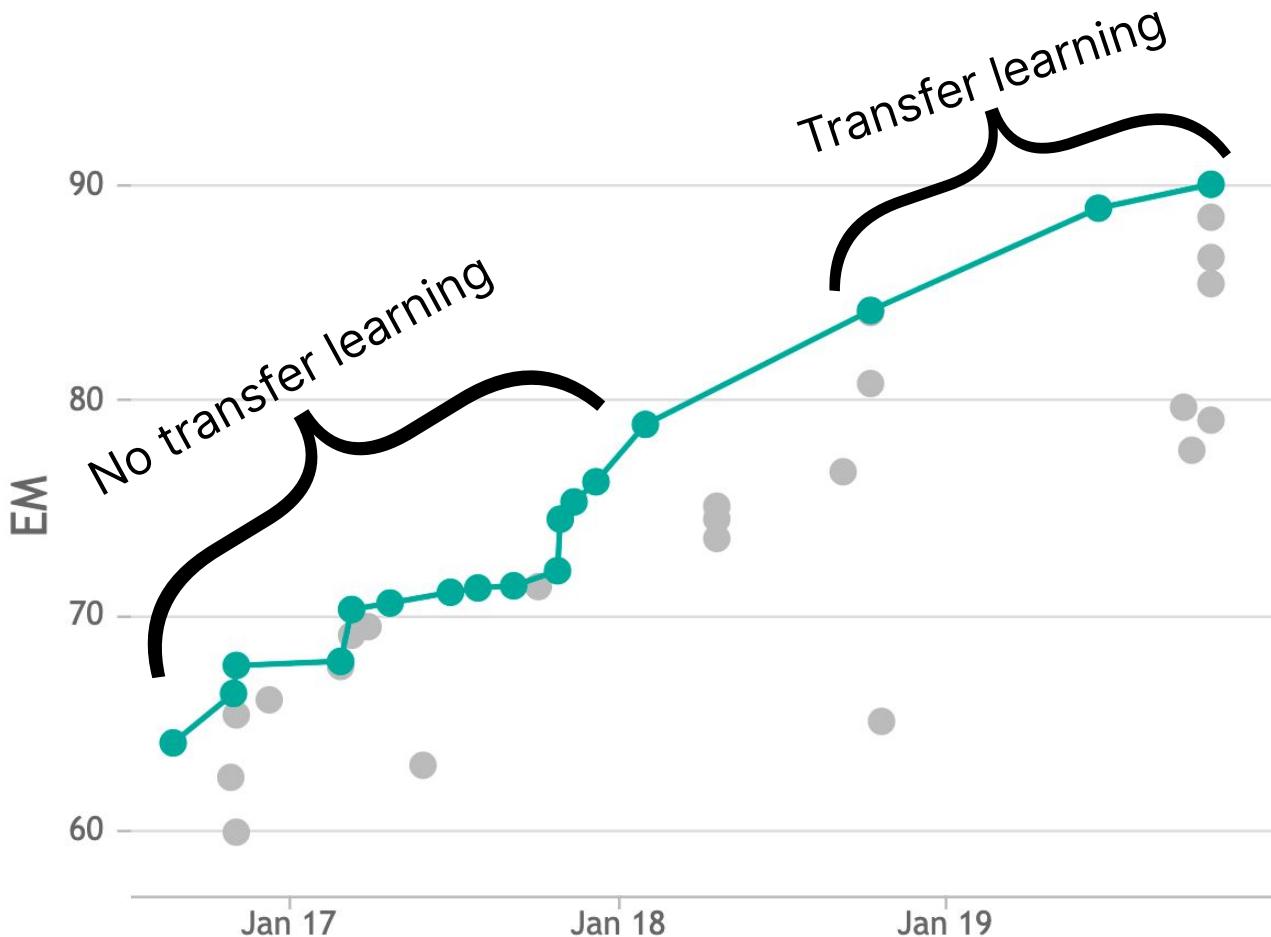
This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!

negative

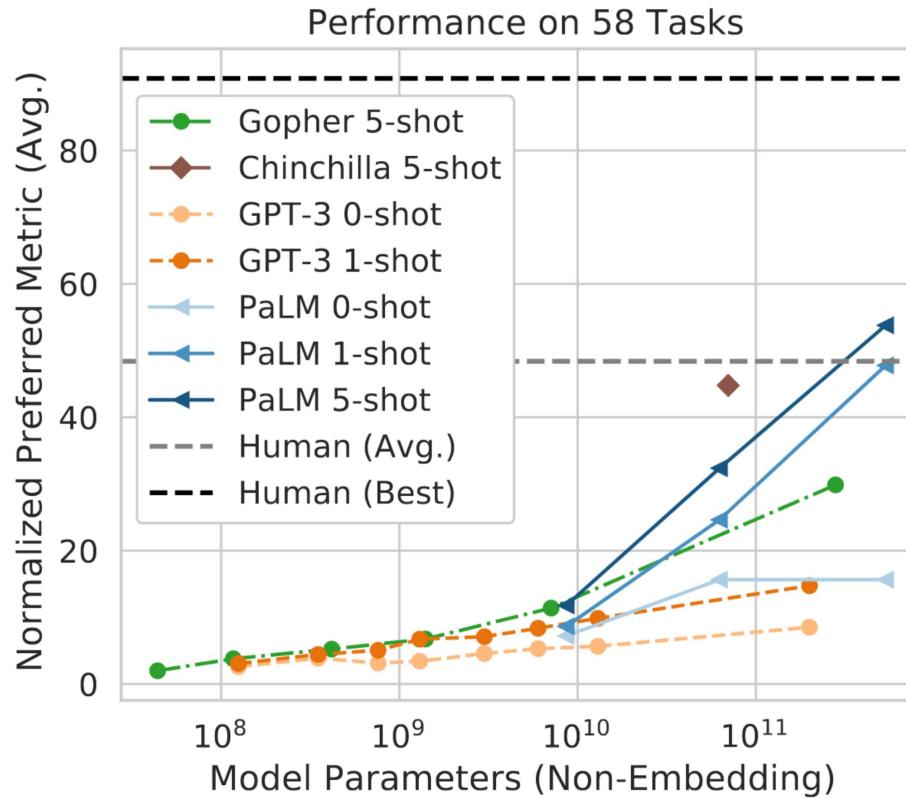
# SQuAD Exact Match score (validation set)



from <https://paperswithcode.com/sota/question-answering-on-squad11-dev>



from <https://paperswithcode.com/sota/question-answering-on-squad11-dev>



from “PaLM: Scaling Language Modeling with Pathways” by Chowdhery et al.

# *What do language models learn from language modeling?*

Language modeling seems to teach models

- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks

# *What do language models learn from language modeling?*

Language modeling seems to teach models

- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks

the dog and cat ate pot pie



$$p(\text{ate}|\text{cat}) = \text{softmax}(Vw_{\text{cat}})_{\text{ate}}$$

*Skip-gram word vector model*

- 0. frog
- 1. frogs
- 2. toad
- 3. litoria
- 4. leptodactylidae
- 5. rana
- 6. lizard
- 7. eleutherodactylus



3. litoria



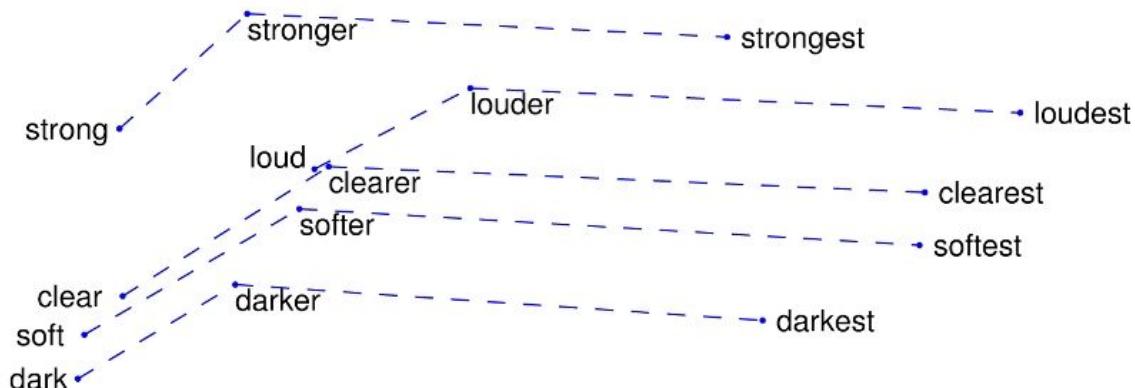
4. leptodactylidae

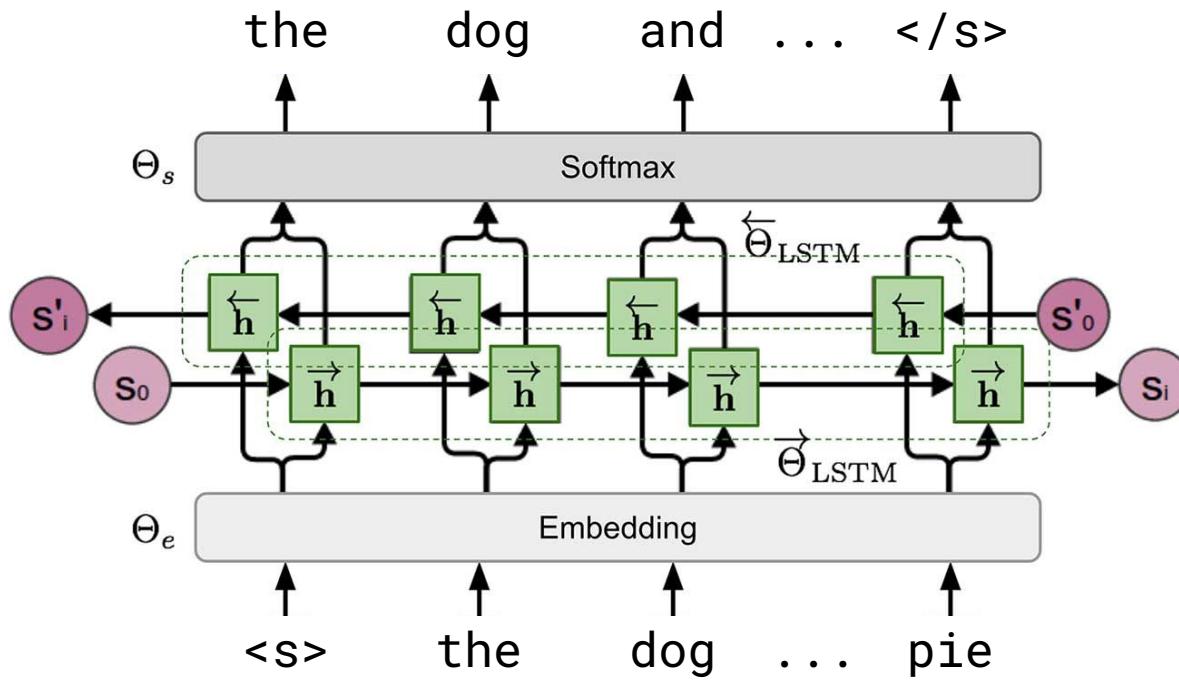


5. rana



7. eleutherodactylus

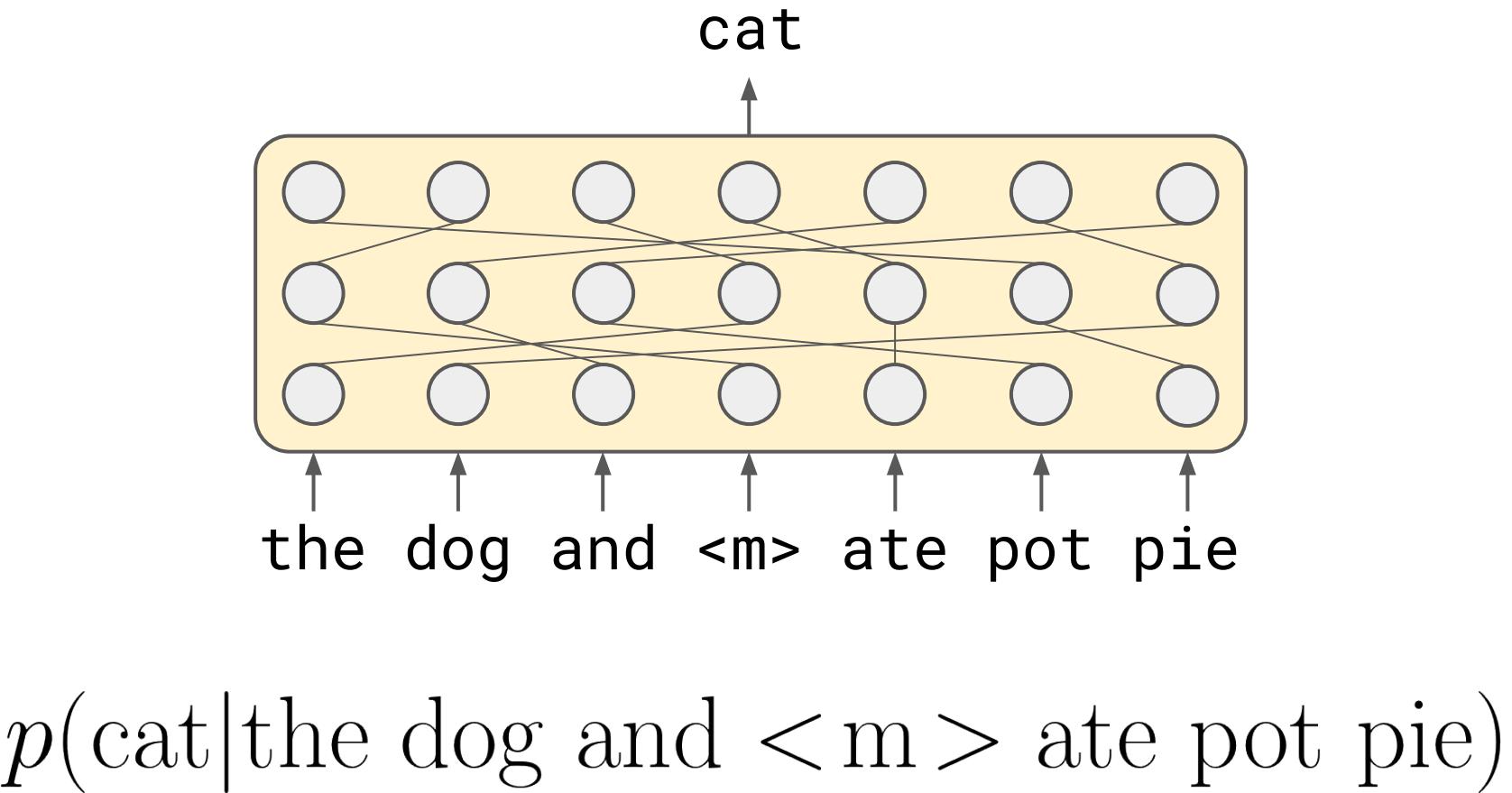




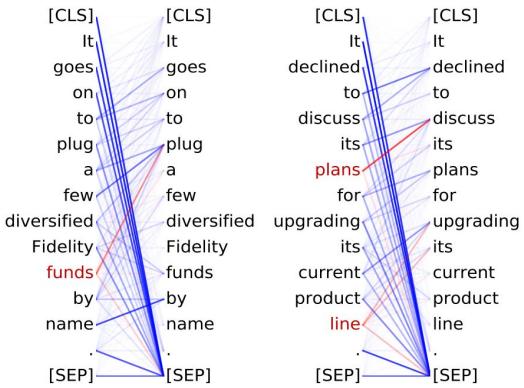
$$p(\text{cat} | \text{the dog and}, \vec{\Theta}_{LSTM}) + p(\text{cat} | \text{ate pot pie}, \overleftarrow{\Theta}_{LSTM})$$

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular <u>play</u> on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent <u>play</u> .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{... } they were actors who had been handed fat roles in a successful <u>play</u> , and had talent enough to fill the roles competently , with nice understatement .

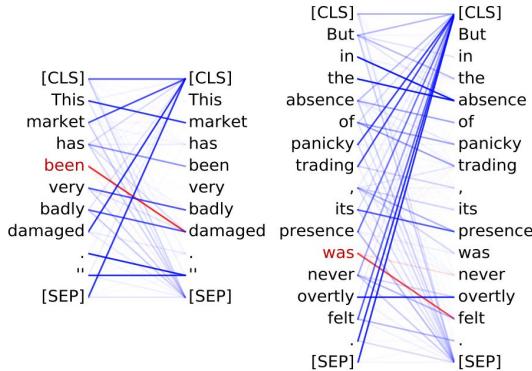
*from “Deep contextualized word representations” by Peters et al.*



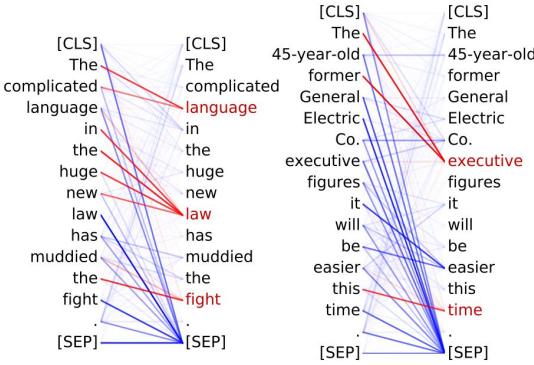
- **Direct objects** attend to their verbs
- 86.8% accuracy at the **dobj** relation



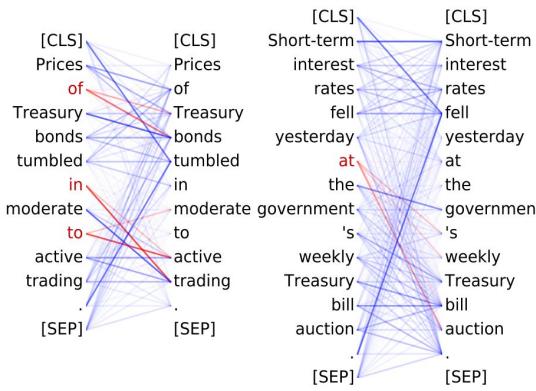
- **Passive auxiliary verbs** attend to the verb they modify
- 82.5% accuracy at the **auxpass** relation



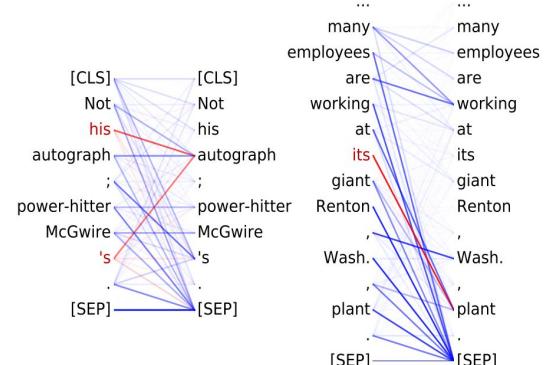
- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the **det** relation



- **Prepositions** attend to their objects
- 76.3% accuracy at the **pobj** relation



- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the **poss** relation



- **Coreferent mentions** attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

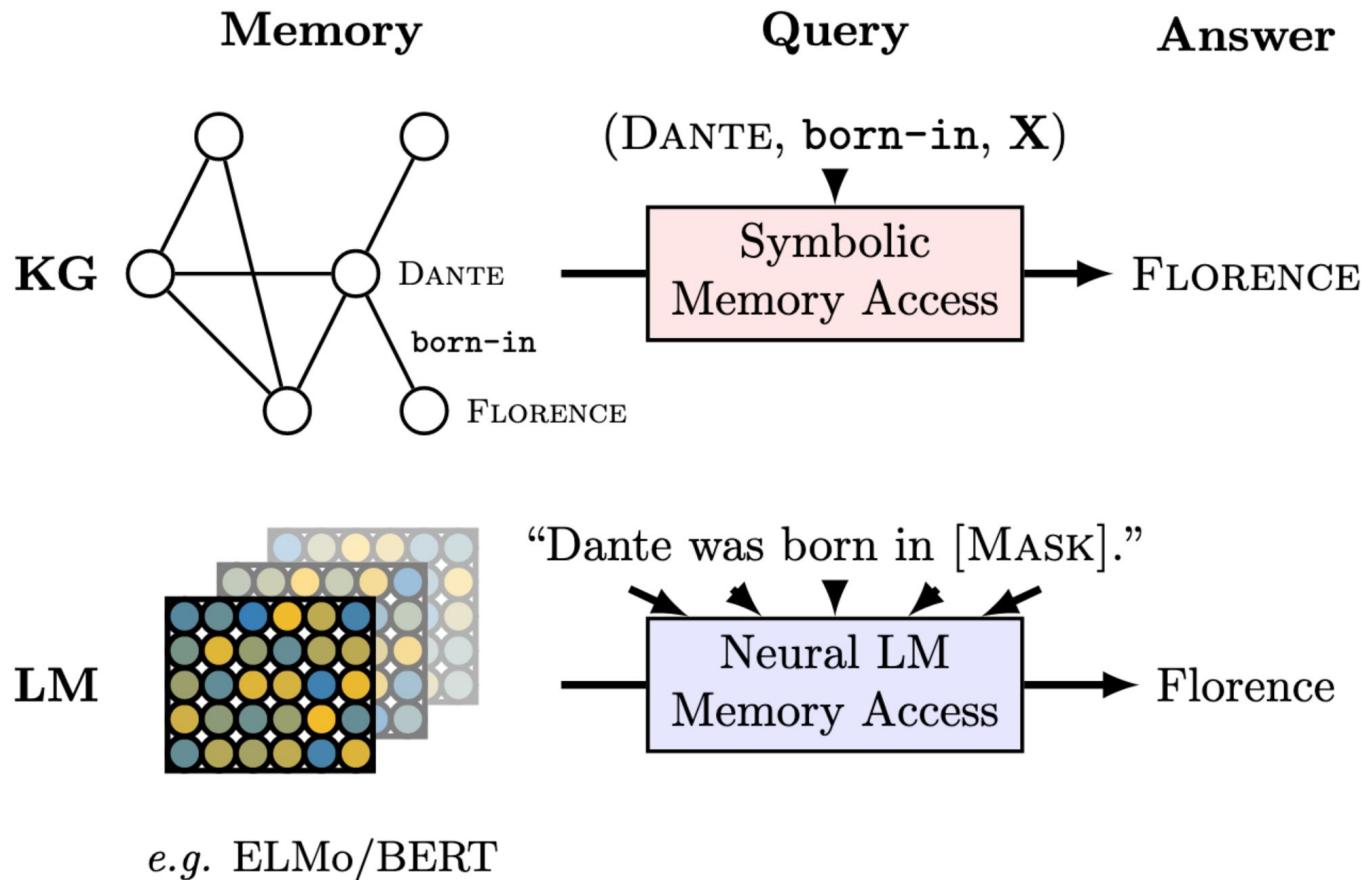


from "What does BERT look at? An Analysis of BERT's Attention" by Clark et al.

# *What do language models learn from language modeling?*

Language modeling seems to teach models

- Word meanings, syntax, and grammar
- **World knowledge**
- How to perform tasks



from “Language Models as Knowledge Bases” by Petroni et al.

President Franklin <M> born <M> January 1882.

Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.

Our <M> hand-picked and sun-dried <M> orchard in Georgia.

T5

D. Roosevelt was <M> in

believe her eyes <M> piece <M> she had ever

peaches are <M> at our

Pre-training

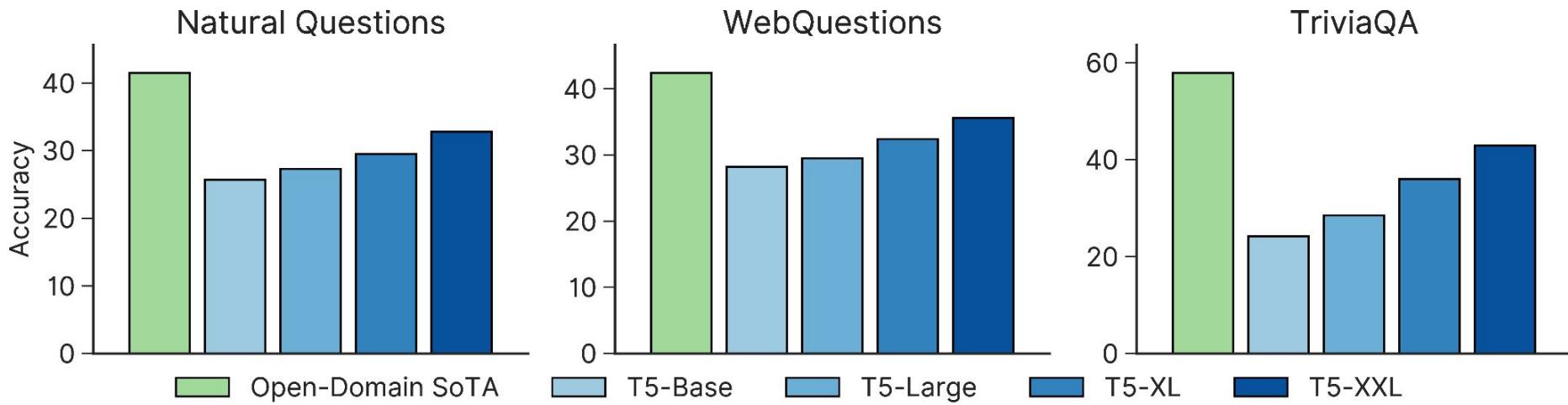
Evaluation

President Franklin D.  
Roosevelt was born  
in January 1882.

When was Franklin D.  
Roosevelt born?

T5

1882

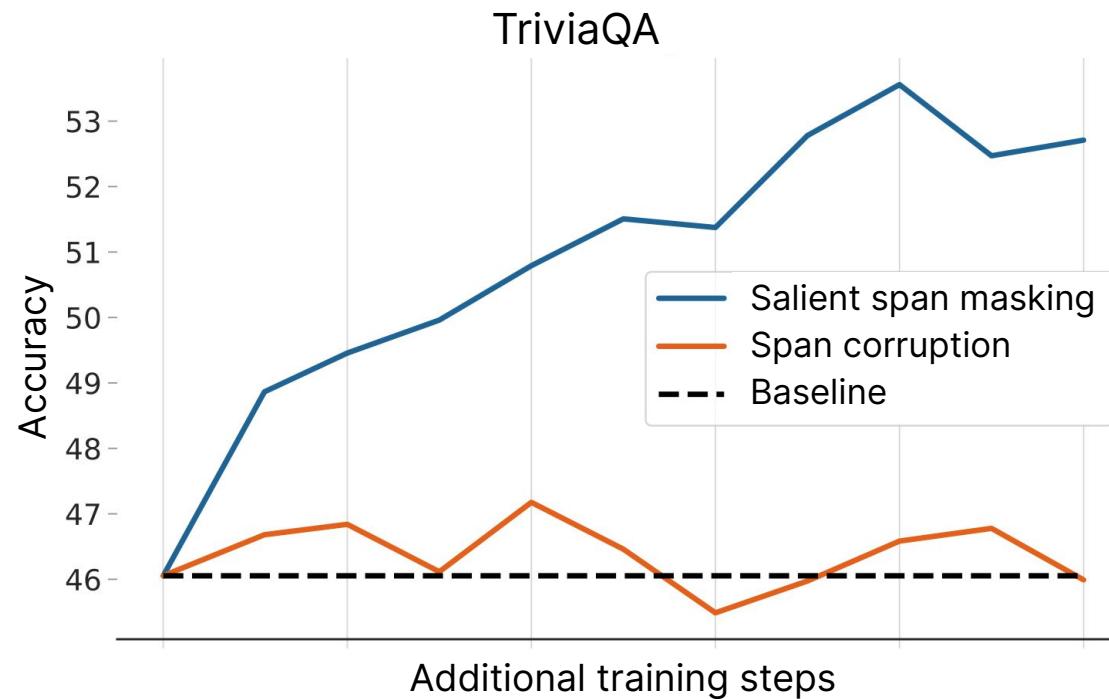


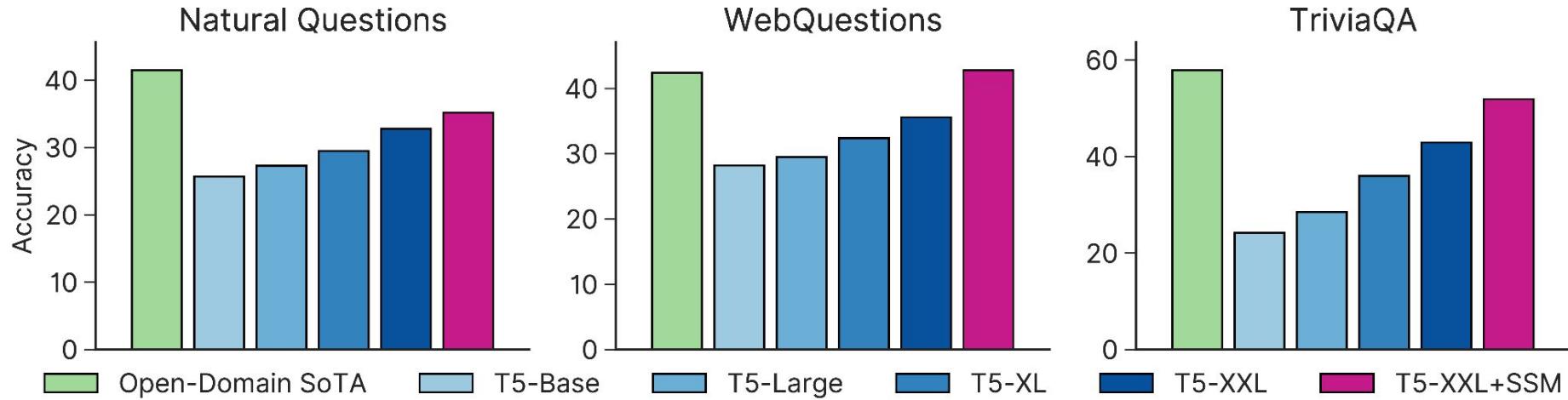
*from "How Much Knowledge Can You Pack Into the Parameters of a Language Model" by Roberts et al.*

<M> (born 1957) is a Spanish librarian who has been the director of the National Library of Spain since February 2013.

T5

Ana Santos Aramburo



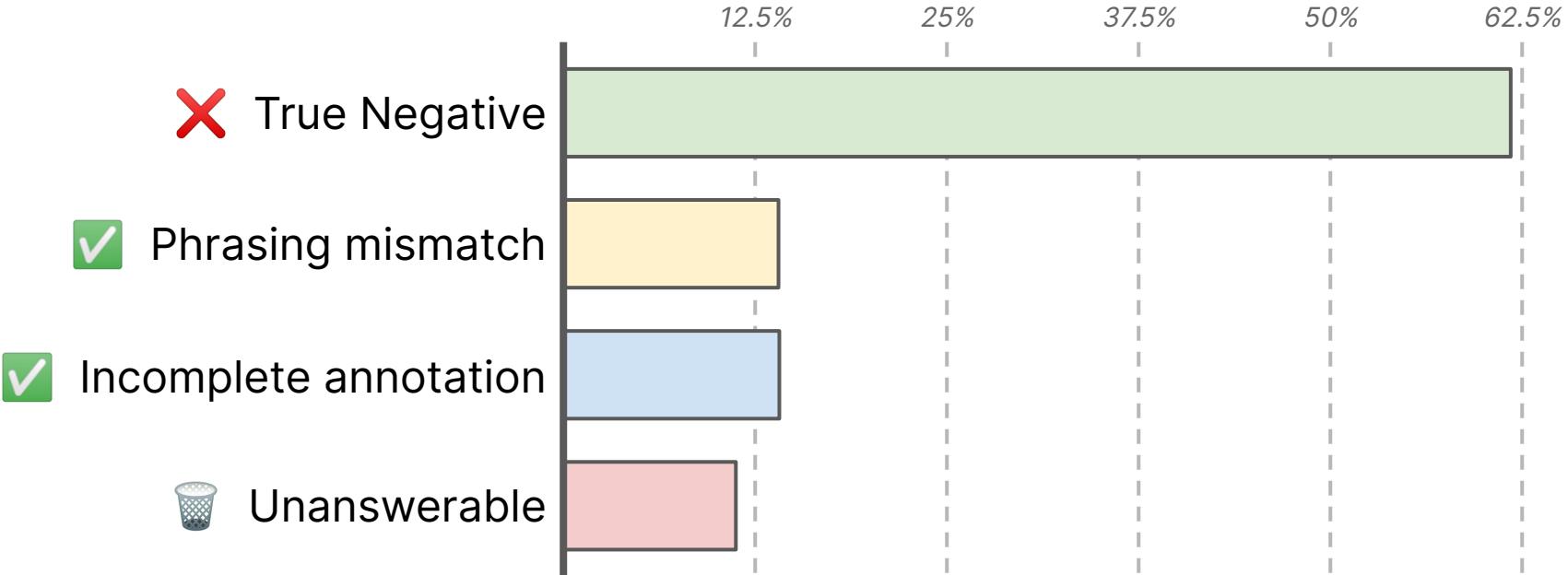


from "How Much Knowledge Can You Pack Into the Parameters of a Language Model" by Roberts et al.

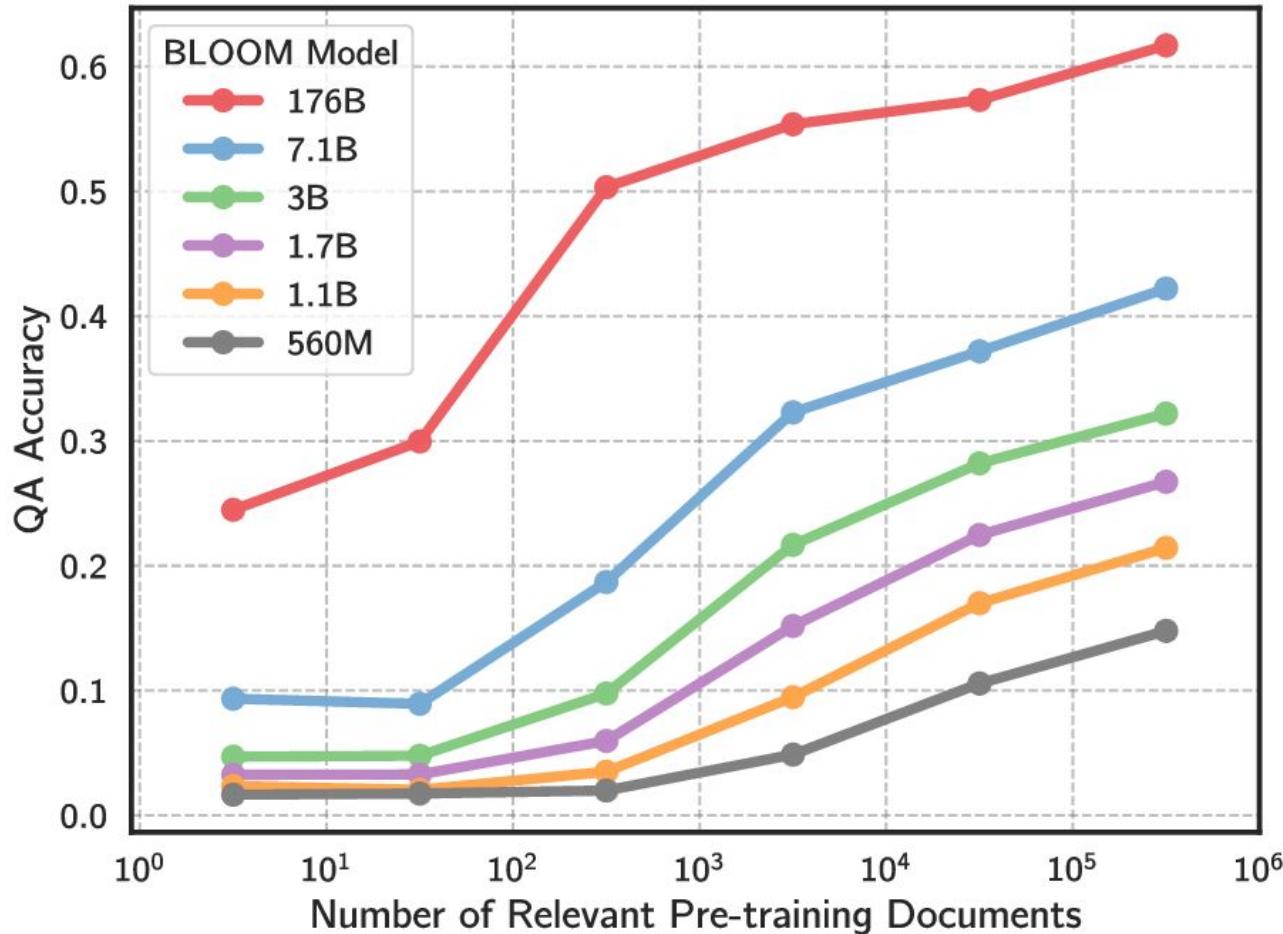
---

Category	Question	Target(s)	T5 Prediction
 True Negative	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
 Phrasing Mismatch	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
 Incomplete Annotation	where does the us launch space shuttles from	florida	kennedy lc39b
Unanswerable	who is the secretary of state for northern ireland	karen bradley	james brokenshire

---



Exact Match: 36.6 → 57.8%!



from "Large Language Models Struggle to Learn Long-Tail Knowledge" by Kandpal et al.

## 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\_Alichieri

Florence

Italy

### Document Indices

3 910 2472 ...

3 348 1032 ...

3 348 810 ...

Count Docs  
w/ Entities

## Question Answering Examples

In what city was the poet Dante born?

Florence

City of Florence

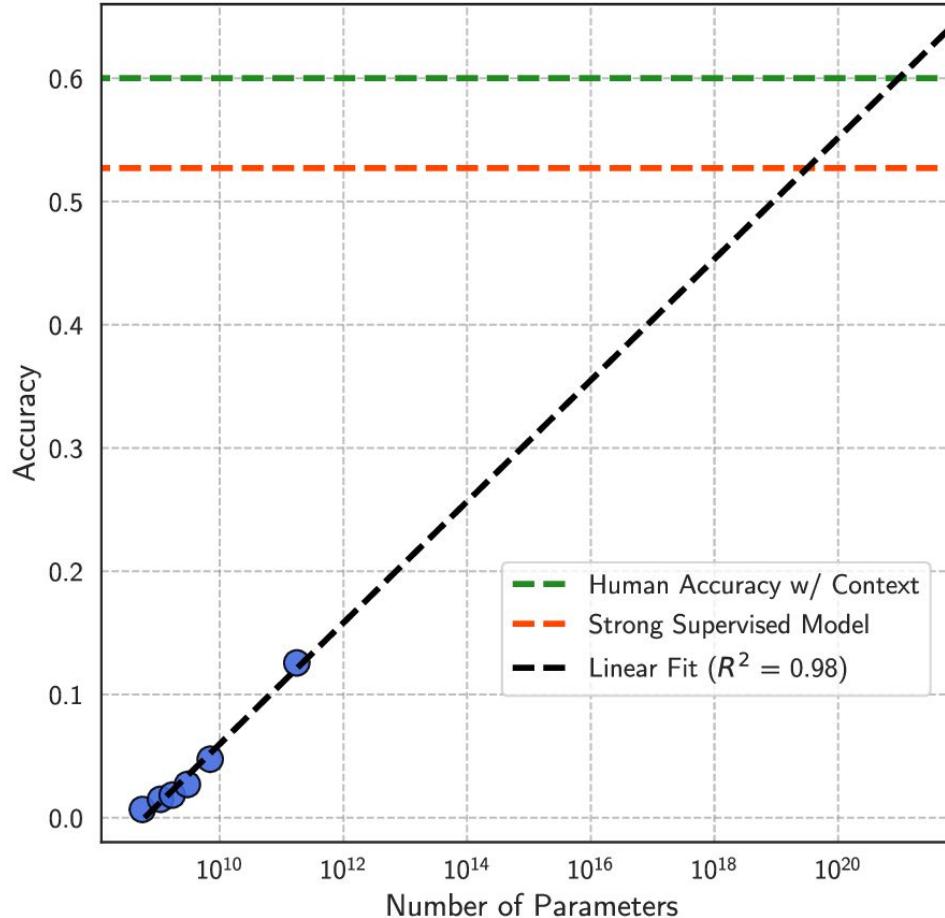
Italy

### Salient Question Entity

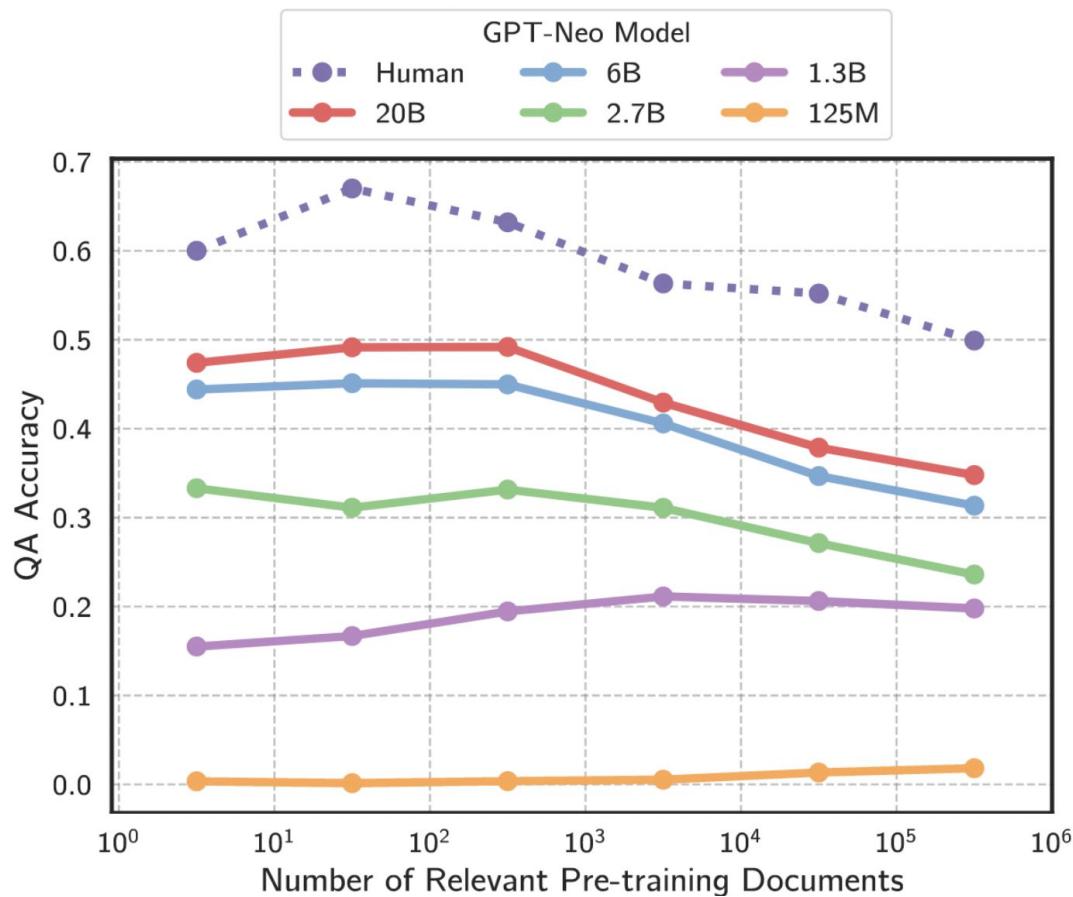
Dante\_Alichieri

### Salient Answer Entity

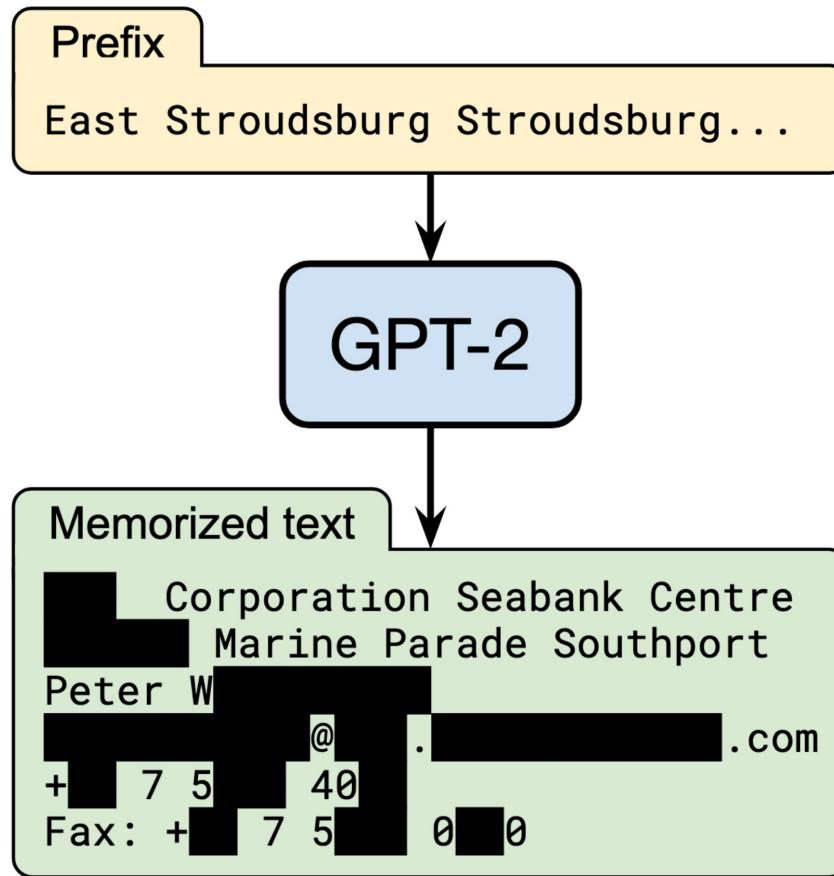
Florence



from “Large Language Models Struggle to Learn Long-Tail Knowledge” by Kandpal et al.



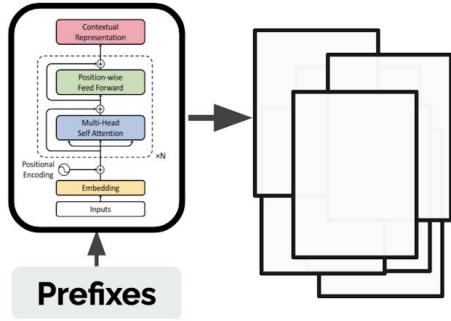
from "Large Language Models Struggle to Learn Long-Tail Knowledge" by Kandpal et al.



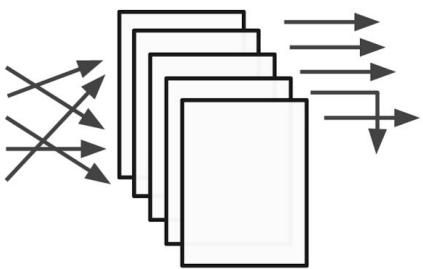
*from “Extracting Training Data from Large Language Models” by Carlini et al.*

## Training Data Extraction Attack

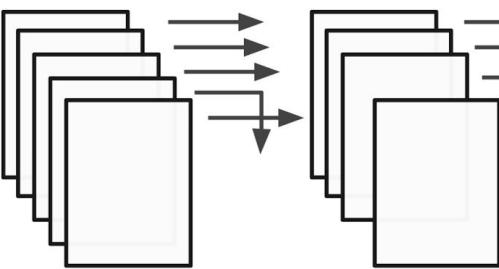
LM (GPT-2)  
200,000 LM Generations



Sorted Generations  
(using one of 6 metrics)



Deduplicate

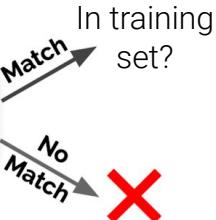


## Evaluation

Choose Top-100

Check Memorization

Internet Search

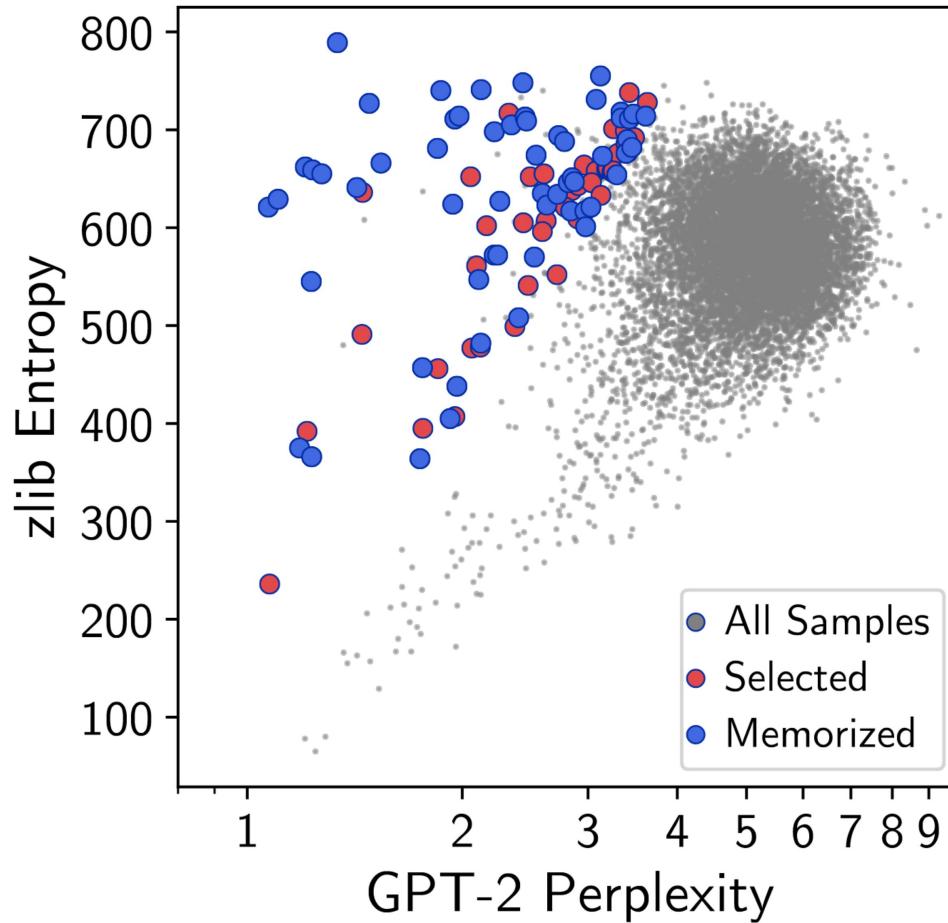


Prefixes

Top- $n$  sampling  
Decaying-temperature sampling  
Conditioning on Internet text

Perplexity  
... vs. different GPT  
... vs. zlib  
... vs. lowercased  
Windowed perplexity

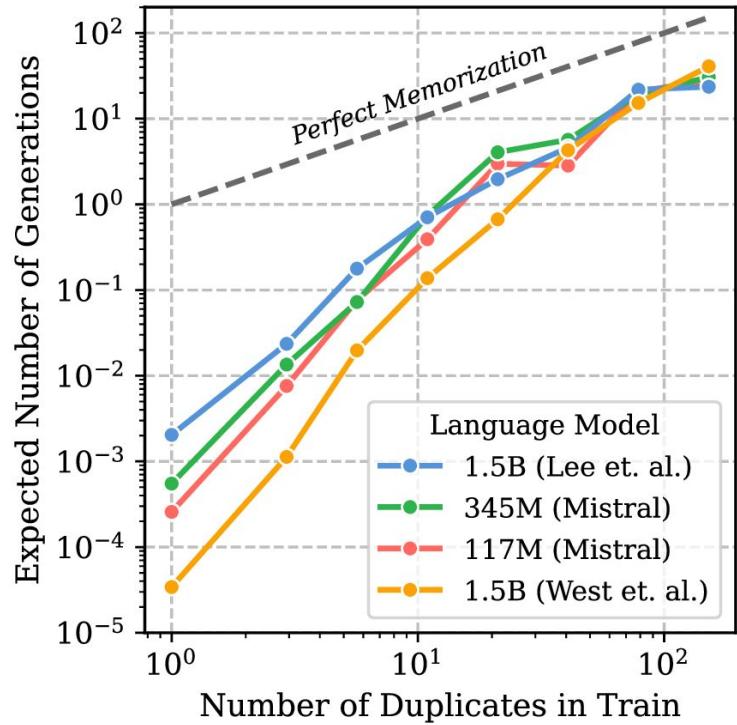
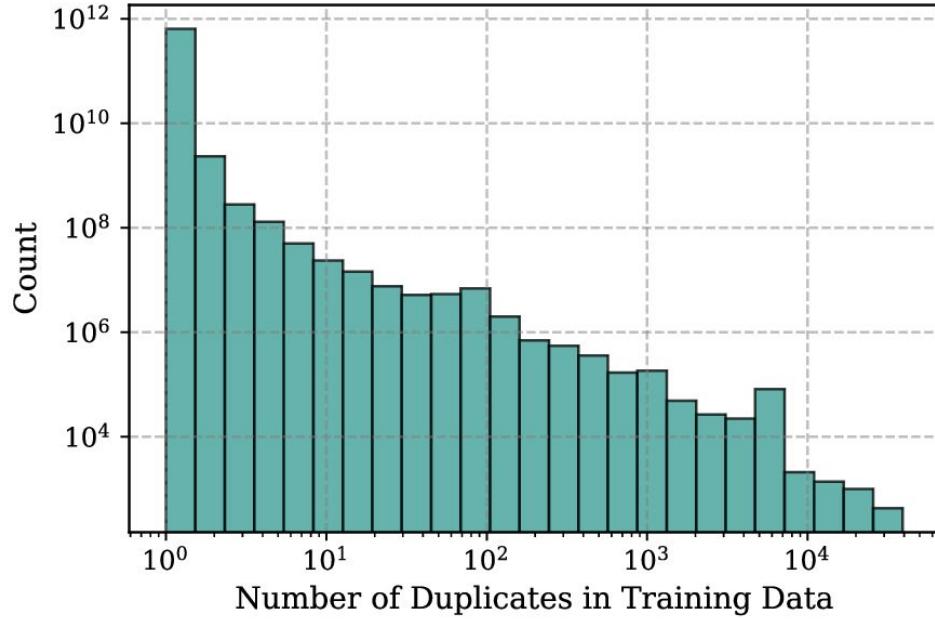
from "Extracting Training Data from Large Language Models" by Carlini et al.



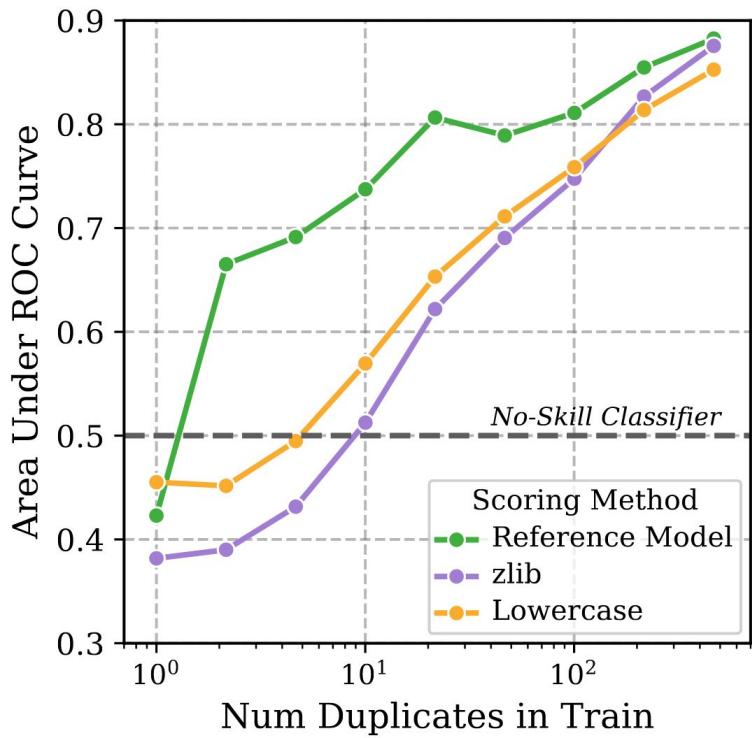
*from "Extracting Training Data from Large Language Models" by Carlini et al.*

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/[REDACTED]51y/milo_evacua...	1	359	✓	✓	1/2
/r/[REDACTED]zin/hi_my_name...	1	113	✓	✓	
/r/[REDACTED]7ne/for_all_yo...	1	76	✓	1/2	
/r/[REDACTED]5mj/fake_news_...	1	72	✓		
/r/[REDACTED]5wn/reddit_admi...	1	64	✓	✓	
/r/[REDACTED]lp8/26_evening...	1	56	✓	✓	
/r/[REDACTED]jla/so_pizzagat...	1	51	✓	1/2	
/r/[REDACTED]ubf/late_night...	1	51	✓	1/2	
/r/[REDACTED]eta/make_christ...	1	35	✓	1/2	
/r/[REDACTED]6ev/its_officia...	1	33	✓		
/r/[REDACTED]3c7/scott_adams...	1	17			
/r/[REDACTED]k2o/because_his...	1	17			
/r/[REDACTED]tu3/armynavy_ga...	1	8			

from “Extracting Training Data from Large Language Models” by Carlini et al.



from “Deduplicating Training Data Mitigates Privacy Risks in Language Models” by Kandpal et al.



	Normal Model	Deduped Model
Training Data Generated	Count 1,427,212 Percent 0.14	68,090 0.007
Mem. Inference AUROC	zlib Ref Model Lowercase	0.76 0.88 0.86
		0.67 0.87 0.68

# *What do language models learn from language modeling?*

Language modeling seems to teach models

- Word meanings, syntax, and grammar
- World knowledge
- **How to perform tasks**

## *Unsupervised pre-training*

The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular; even the Prince of Wales (the future King Edward VII) travelled in one. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. Passengers reported that the interior fittings were luxurious when compared to horse-drawn cabs but there were some complaints that the internal ...

lighting made them too conspicuous to those outside the cab. The fleet peaked at around 75 cabs, all of which needed to return to the single depot at Lambeth to switch batteries.

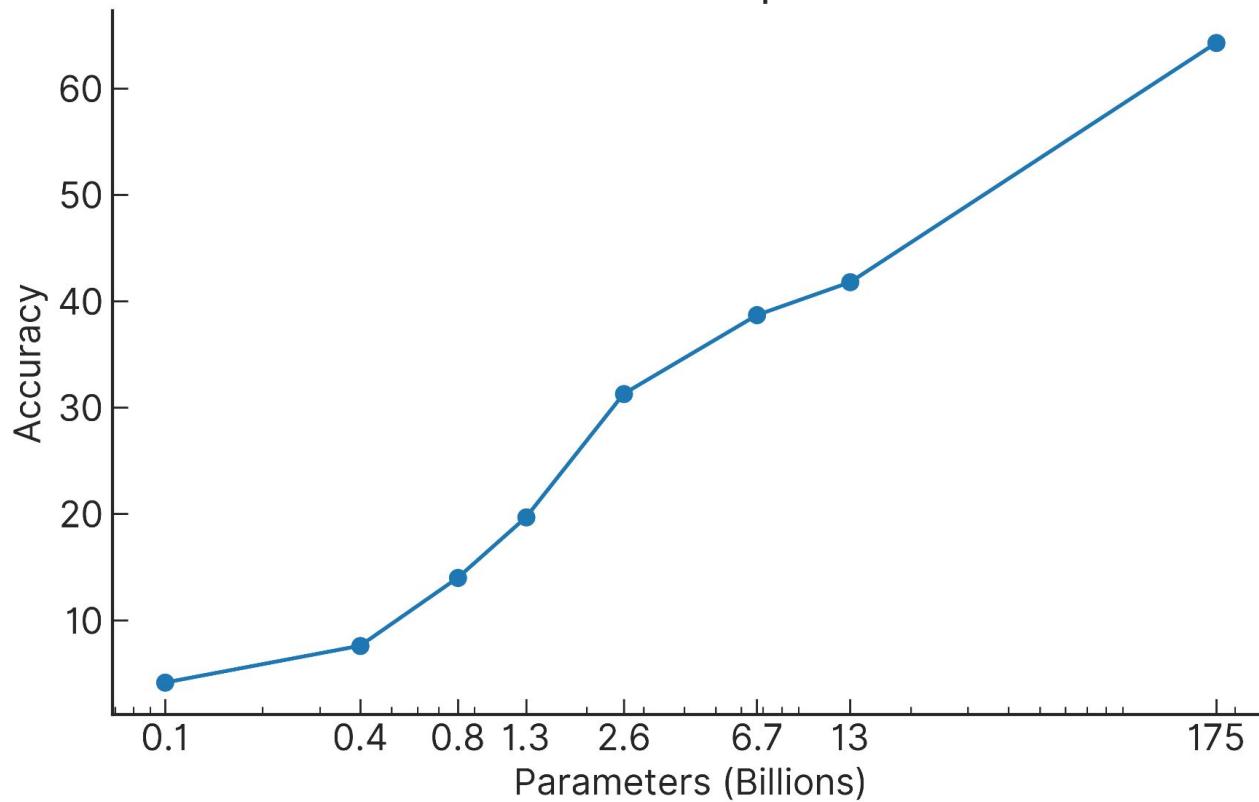


## *"Zero-shot" prompting*

Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

yes

## TriviaQA zero-shot performance



*from "Language Models are Few-Shot Learners" by Brown et al.*

## Closed-book question answering

<http://www.autos weblog.com/cat/trivia-questions-from-the-50s>

who was frank sinatra? a: an american singer, actor, and producer.

## Paraphrase identification

<https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same>

Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

## Natural Language Inference

<https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply>

If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

## Summarization

<https://blog.nytsoi.net/tag/reddit>

... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.

## Pronoun resolution

<https://nursecheung.com/ati-teas-guide-to-english-language-usage-understanding-pronouns/>

Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.

## Summarization

*The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?*

## Paraphrase identification

"How is air traffic controlled?" "How do you become an air traffic controller?"  
Pick one: these questions are duplicates or not duplicates.

## Question answering

I know that the answer to "What team did the Panthers defeat?" is in "The Panthers finished the regular season [...]" . Can you tell me what it is?

## Multi-task training

## Zero-shot generalization

## Natural language inference

Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

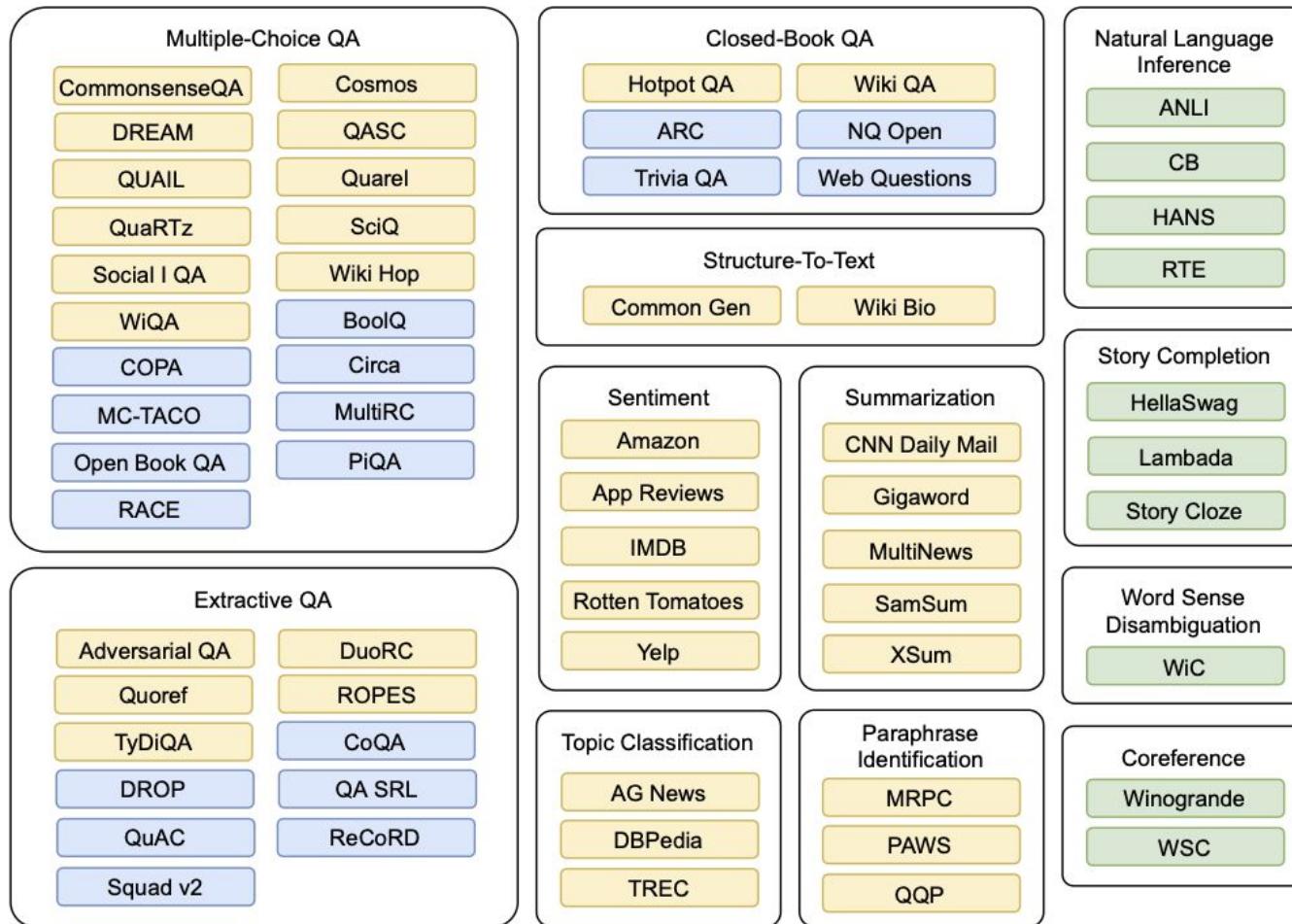
T0

Graffiti artist Banksy is believed to be behind [...]

Not duplicates

Arizona Cardinals

Yes



from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.

### QQP (Paraphrase)

Question1	How is air traffic controlled?
Question2	How do you become an air traffic controller?
Label	0

{Question1} {Question2}  
Pick one: These questions  
are duplicates or not  
duplicates.

{Choices[label]}

I received the questions  
"{Question1}" and  
"{Question2}". Are they  
duplicates?

{Choices[label]}

### XSum (Summary)

Document	The picture appeared on the wall of a Poundland store on Whymark Avenue...
Summary	Graffiti artist Banksy is believed to be behind...

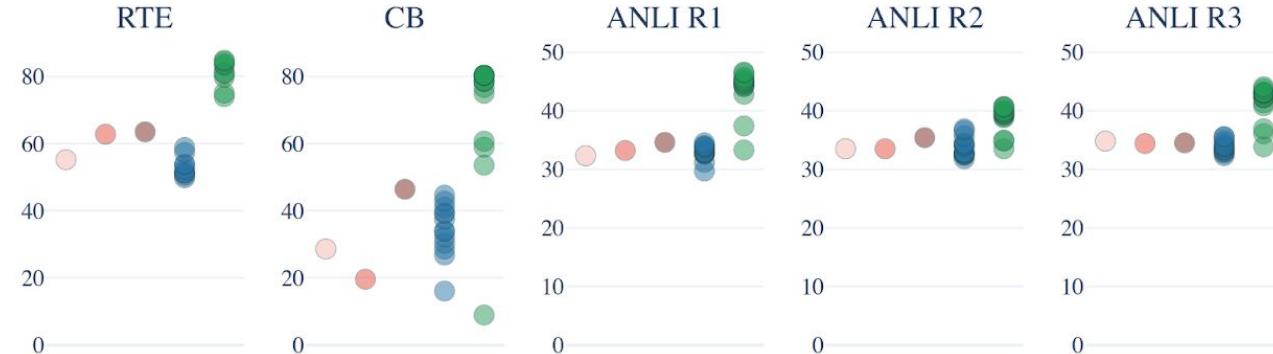
{Document}  
How would you  
rephrase that in  
a few words?

{Summary}

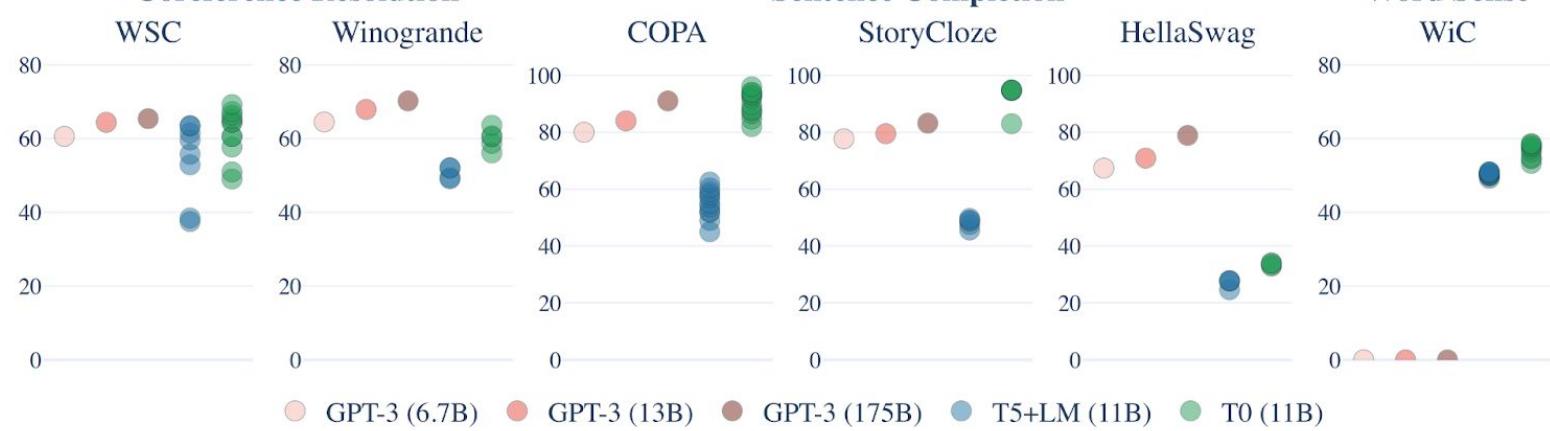
First, please read the article:  
{Document}  
Now, can you write me an  
extremely short abstract for it?

{Summary}

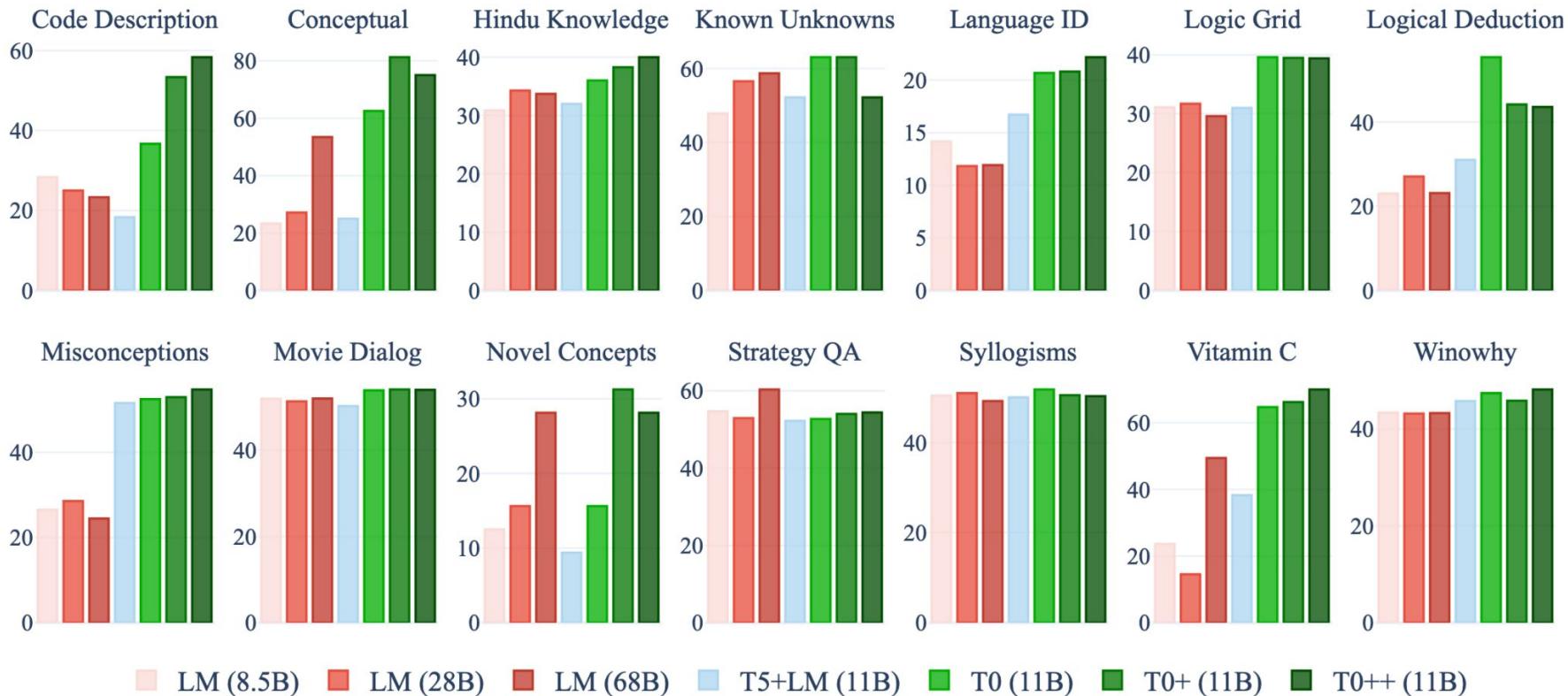
## Natural Language Inference



## Coreference Resolution



from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.



from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.

# *What do language models learn from language modeling?*

Language modeling seems to teach models

- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks

*What happens when we make language models larger?*

Larger language models learn more

- World knowledge
- Esoteric facts
- Tasks (maybe?)

*What happens when we make language models larger?*

Larger language models learn more

- World knowledge
- Esoteric facts
- Tasks (maybe?)

*Mostly a function  
of the data?*

# Index of /wikidatawiki/entities/

---

[latest-all.json.bz2](#)  
[latest-all.json.gz](#)

27-Jul-2021 11:32      68418560489  
27-Jul-2021 04:58      102963487951

*~68GB compressed = 17B float32 parameters*

*The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both application areas, the training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status.*

# co:here

# API

OpenAI API Beta

ABOUT

EXAMPLES

DOCS

PRICING

LOG IN

JOIN >

OpenAI technology, just an HTTPS call away

Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.

JOIN THE WAITLIST >

</> EXPLORE THE DOCS

AI21 studio

# Custom language models built for scale

Build sophisticated language applications on top of AI21's language models

## Microsoft Megatron-Turing NLG 530B

The World's Largest and Most Powerful Generative Language Model

SambaNova<sup>®</sup>  
SYSTEMS

PRODUCTS ▾ SOLUTIONS ▾ RESOURC

Enterprise-Grade  
Large Language Models  
Made Simple & Accessible

Introducing Dataflow-as-a-Service™ GPT

Introducing the LightOn Muse API



# Create. Process. Understand. Learn.



Production-ready intelligence primitives powered by state-of-the-art language models.

For the first time natively in French, Spanish, Italian, and more. Now in private beta!



# AI, 모두의 능력이 되다. HyperCLOVA

AI가 모두의 능력이 되는 새로운 시대.

그 시작이 될 HyperCLOVA를 소개합니다.

네이버 클로바와 함께 새로운 시대를 시작하세요.

Thanks.

Please give me feedback:

<http://bit.ly/colin-talk-feedback>