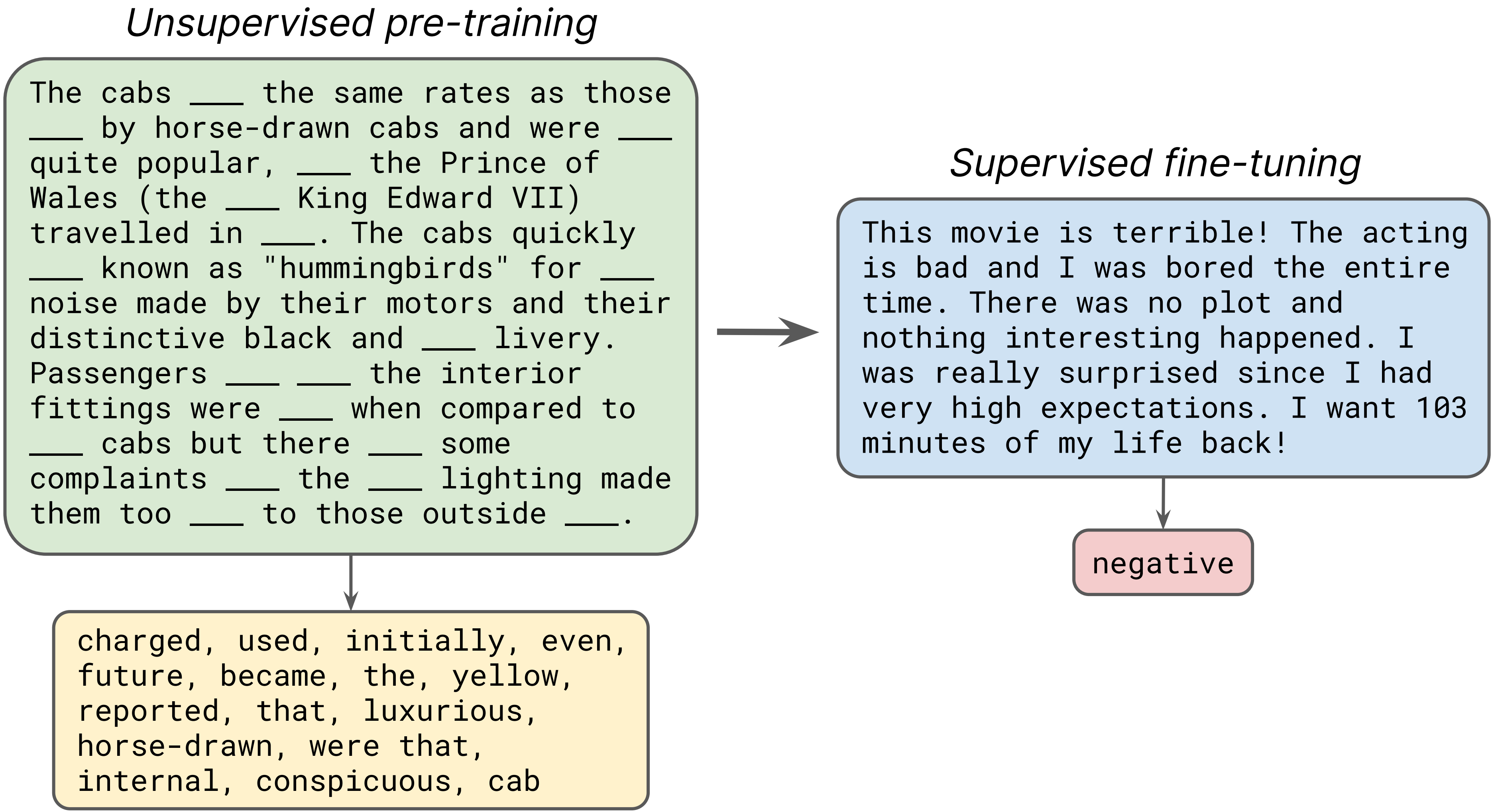
What do language models learn from language modeling?

Colin Raffel



Unsupervised 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 ___.

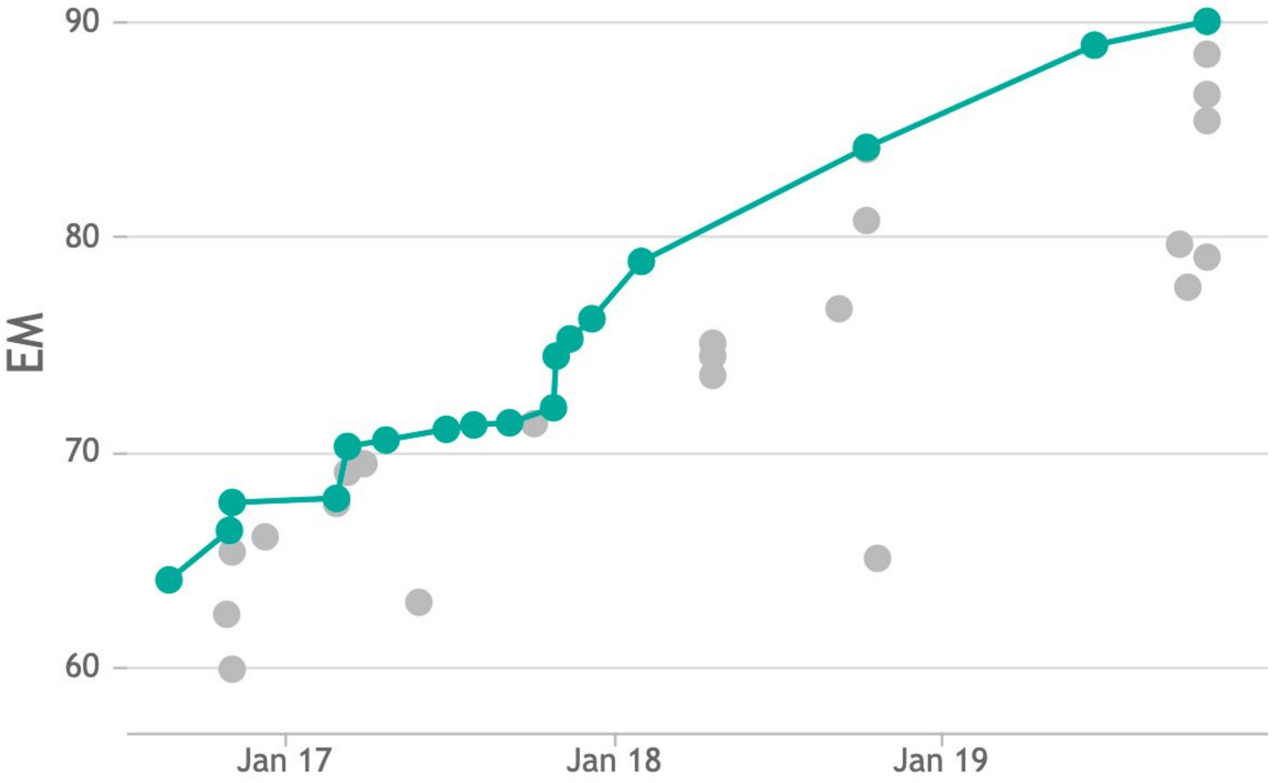
charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab

Supervised fine-tuning

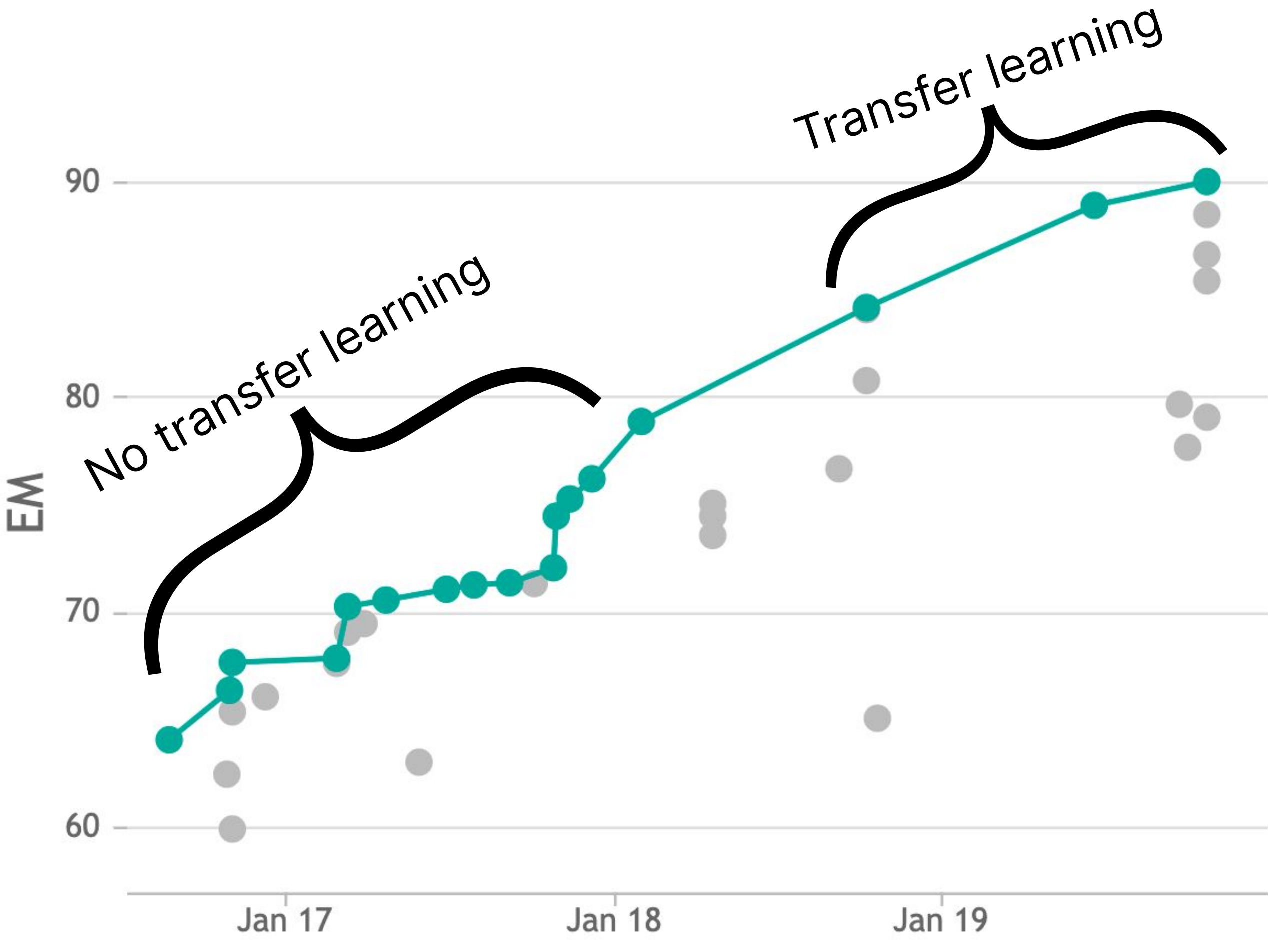
This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!

negative

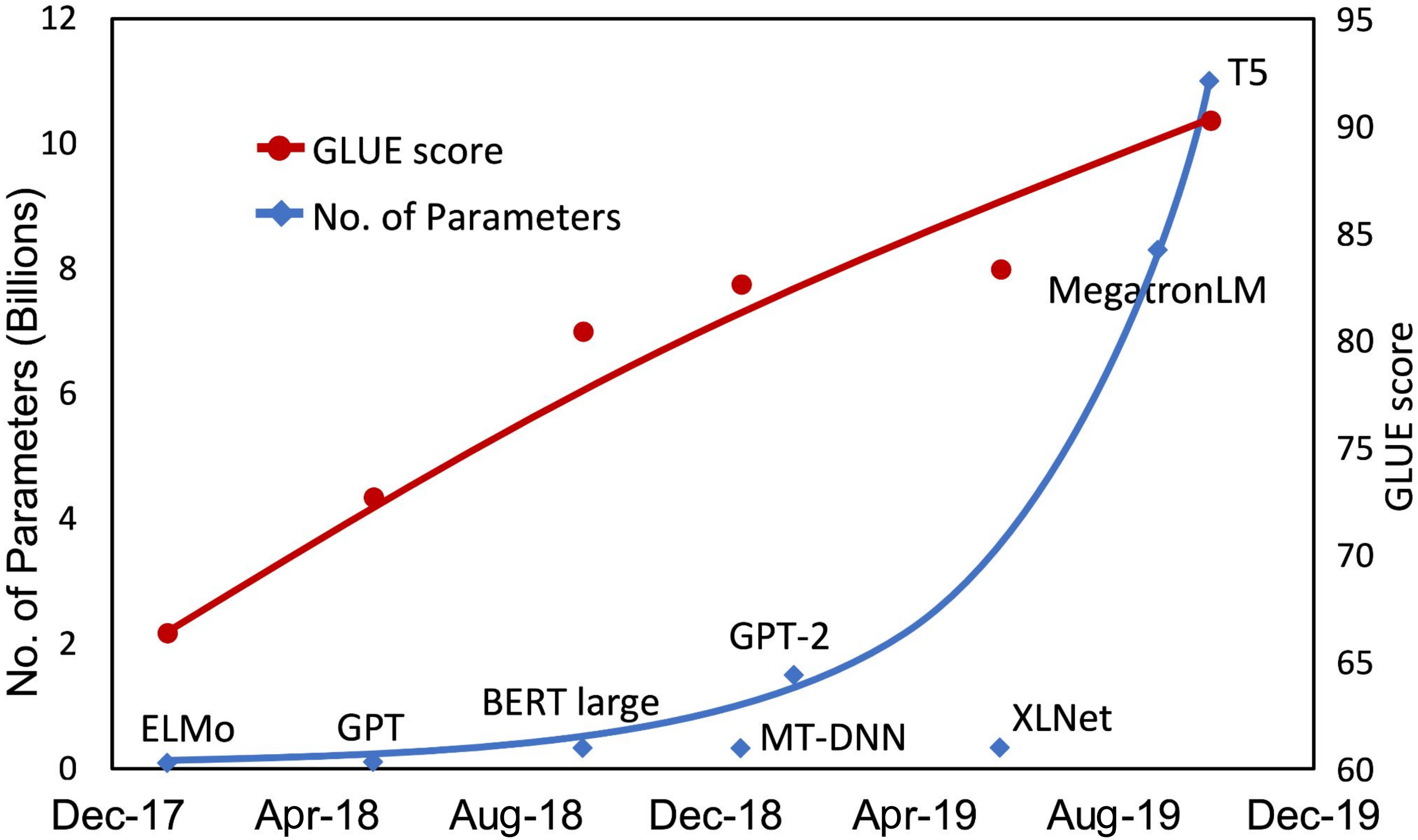
SQuAD Exact Match score (validation set)



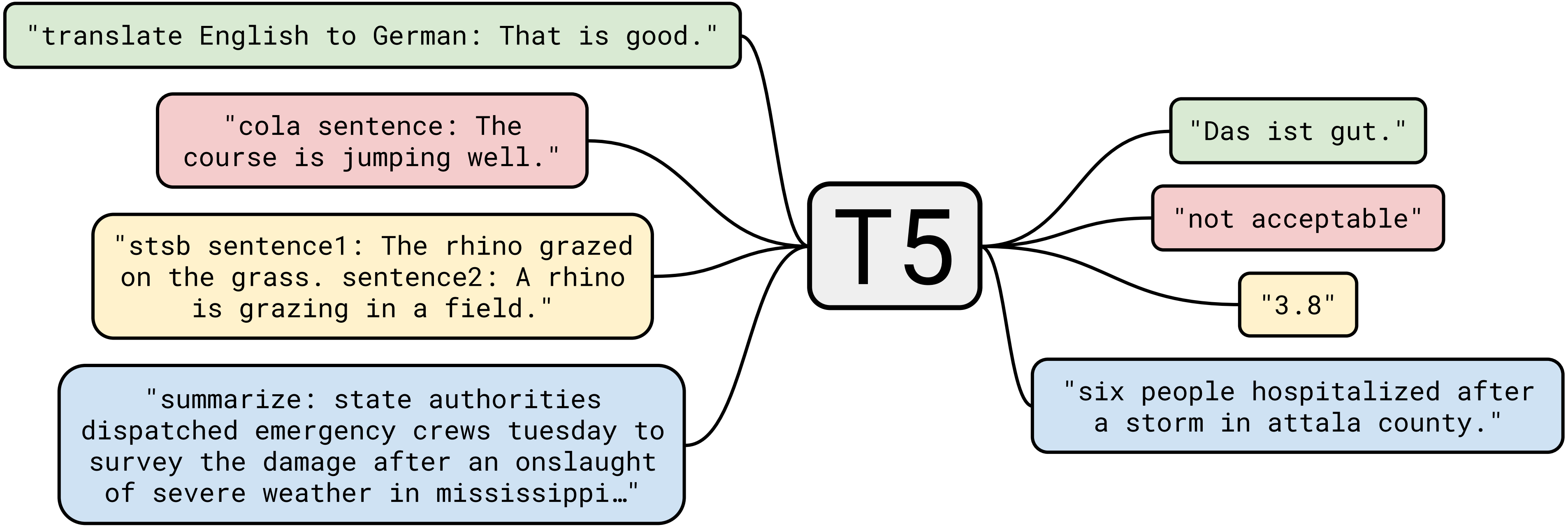
from https://paperswithcode.com/sota/question-answering-on-squad11-dev



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



from "Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach" by Ahmet and Abdullah



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

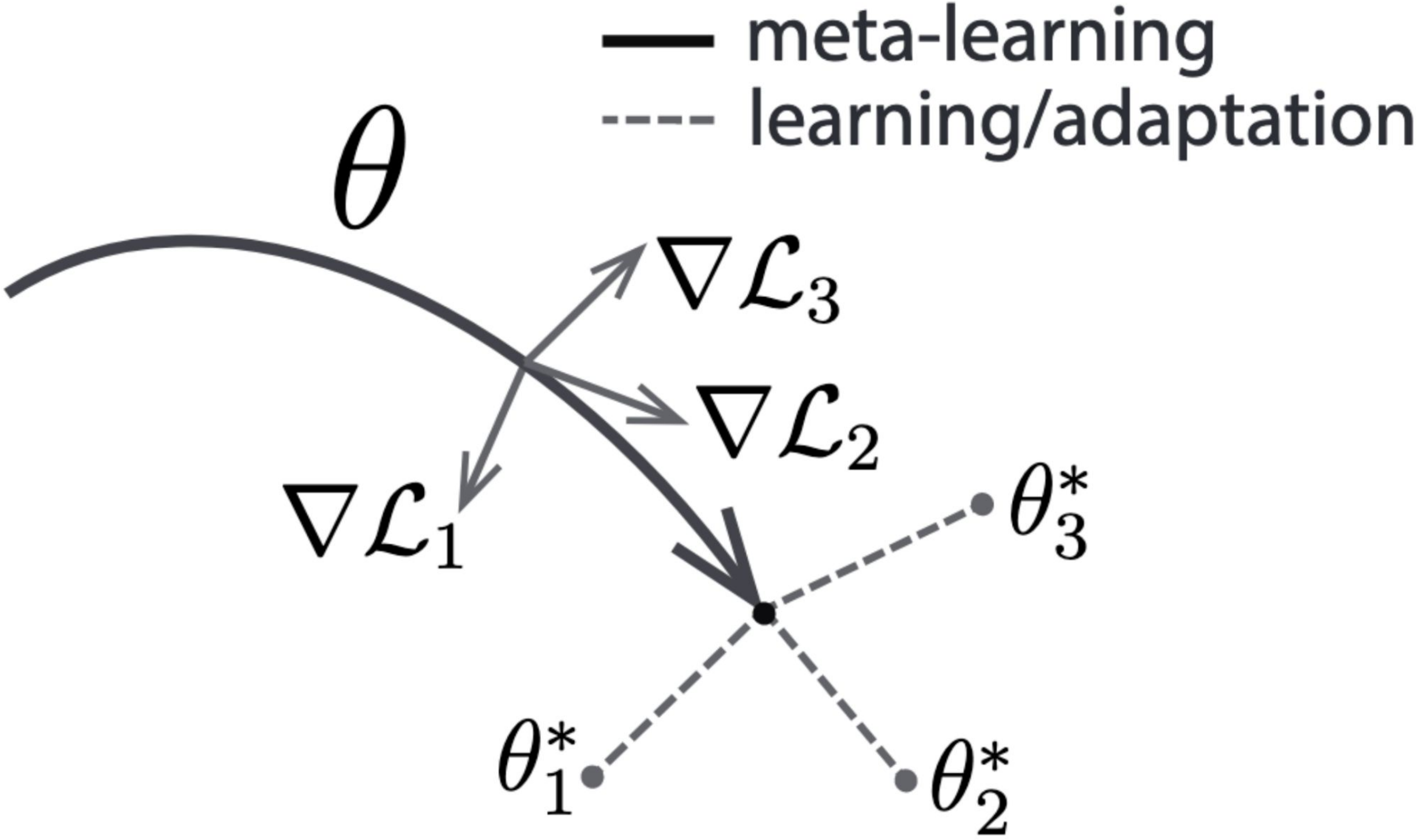
"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

from “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” by Raffel et al.



from “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks” by Finn et al.

Why does language modeling effectively result in meta-learning?

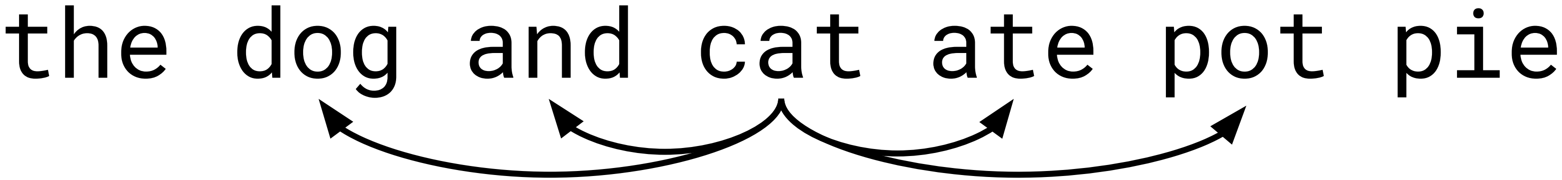
Language modeling seems to teach models

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

Why does language modeling effectively result in meta-learning?

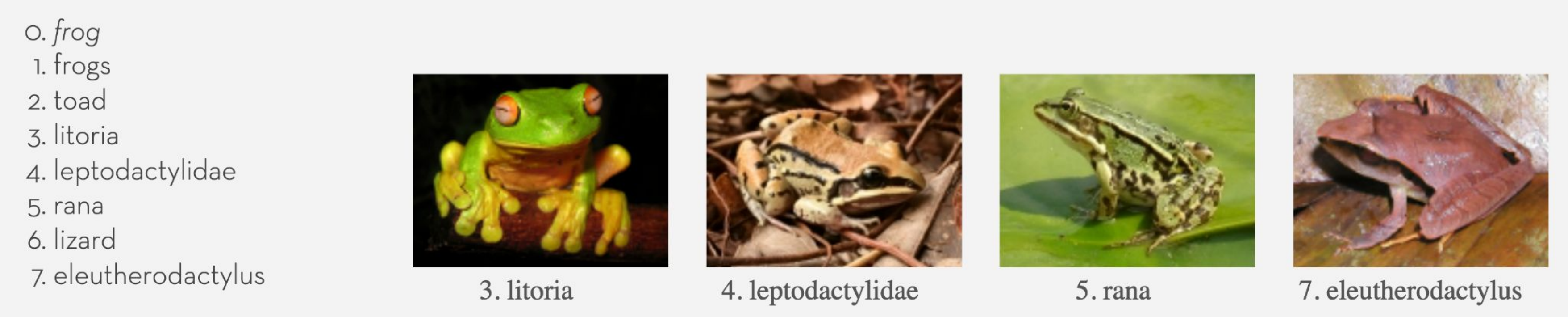
Language modeling seems to teach models

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



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

Skip-gram word vector model



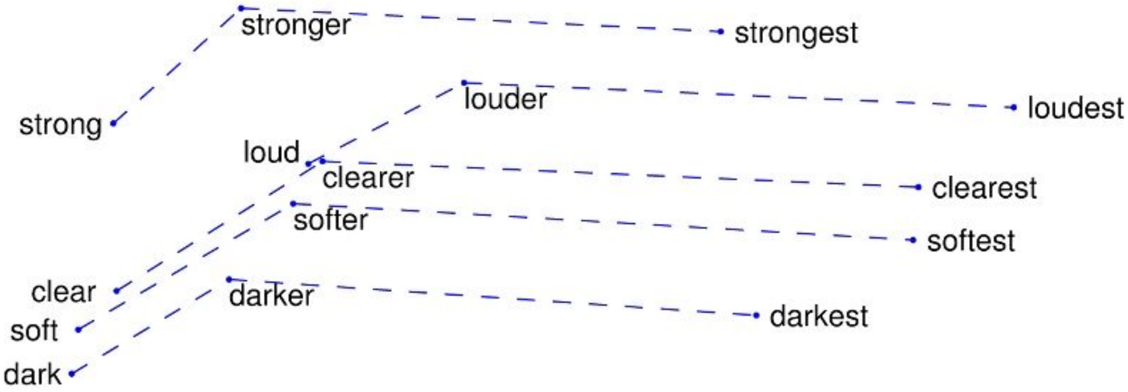
0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

3. litoria

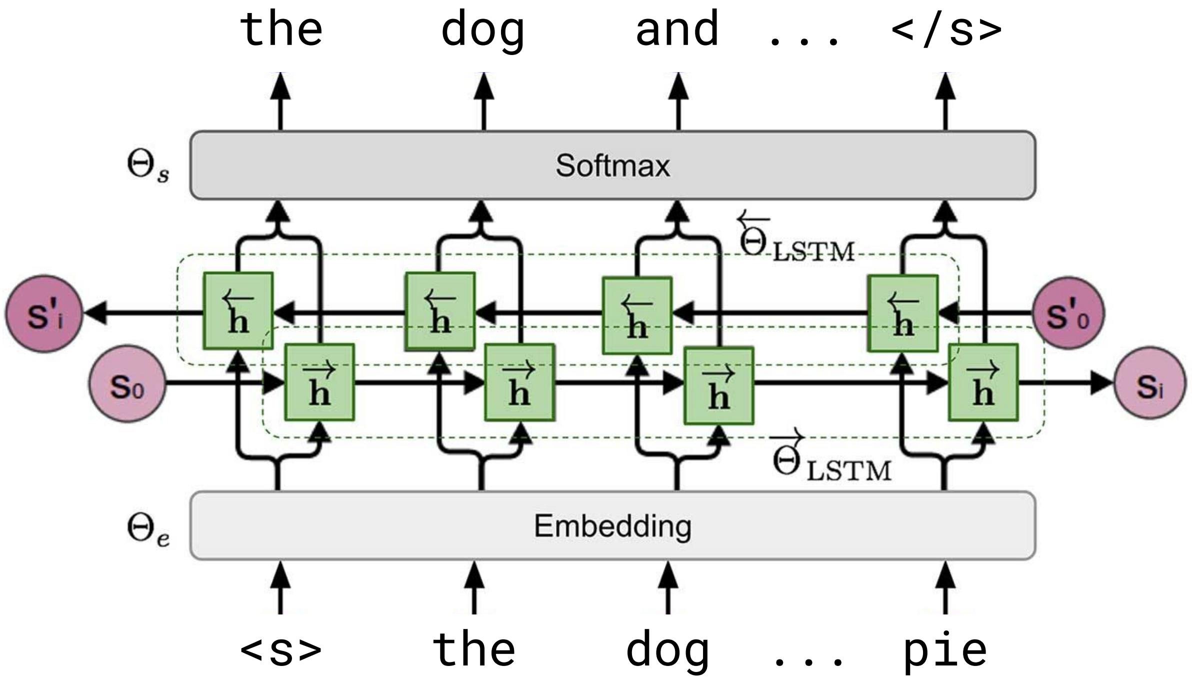
4. leptodactylidae

5. rana

7. eleutherodactylus

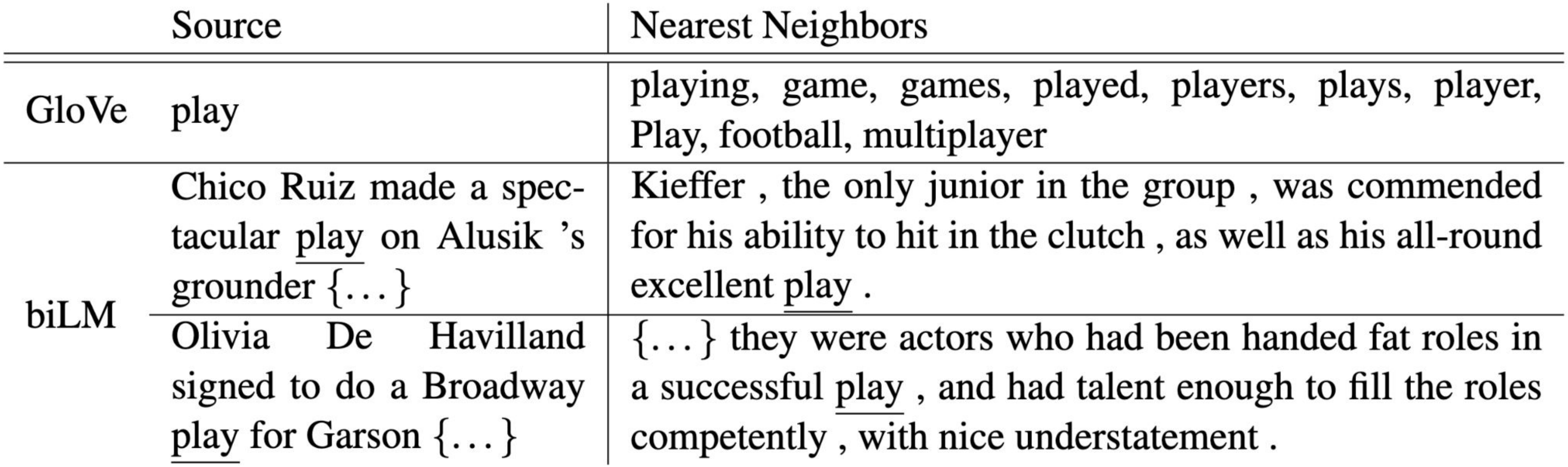


from https://nlp.stanford.edu/projects/glove/



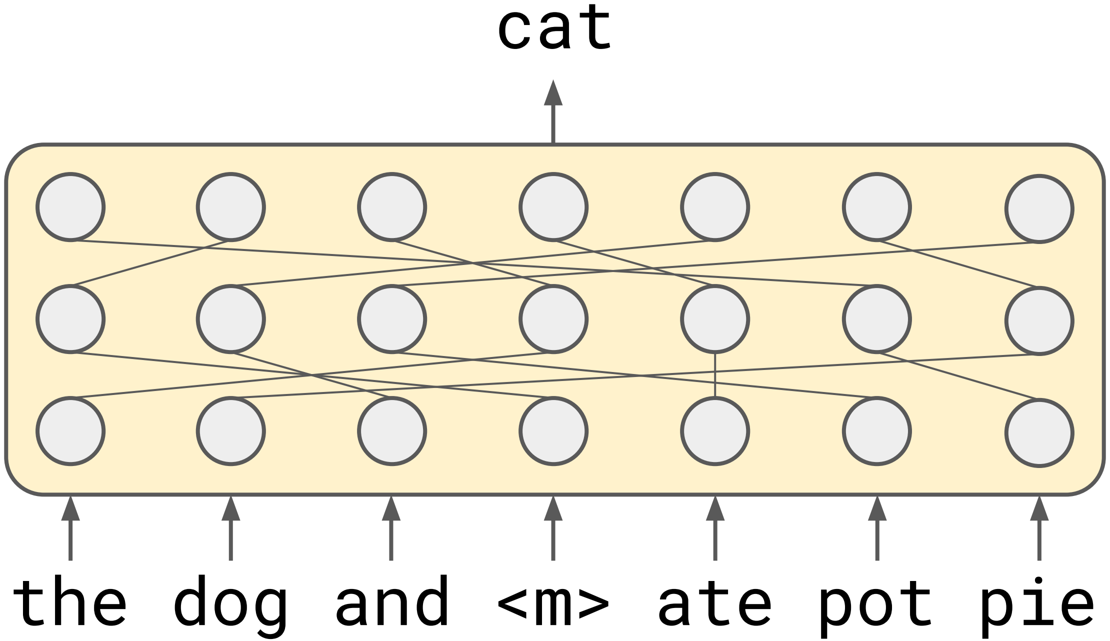
$$p\left(\text{cat}\,|\,\text{the dog and, } \overrightarrow{\Theta}_{\text{LSTM}}\right) + p\left(\text{cat}\,|\,\text{ate pot pie, } \overleftarrow{\Theta}_{\text{LSTM}}\right)$$

from https://www.topbots.com/generalized-language-models-cove-elmo/



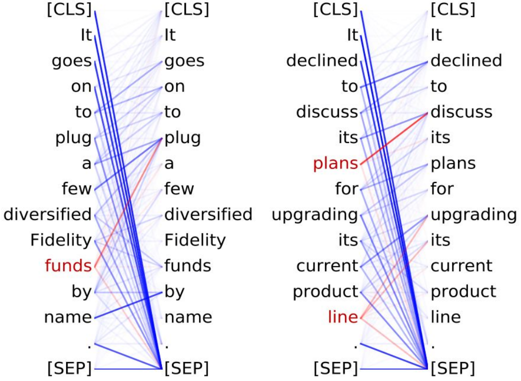
	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {. . . }	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round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

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

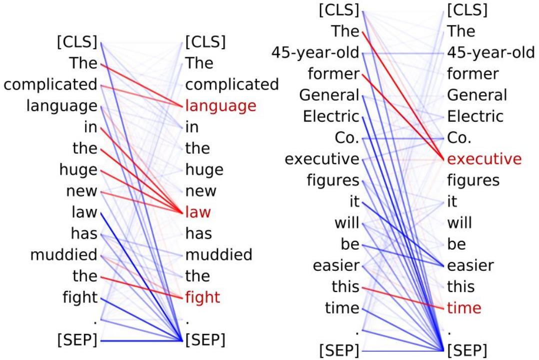


$$p(\text{cat}|\text{the dog and} < \text{m} > \text{ate pot pie})$$

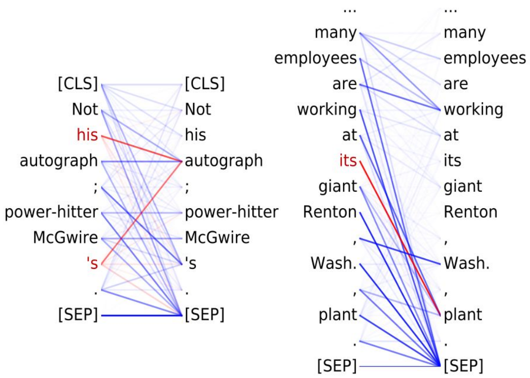
- **Direct objects** attend to their verbs
- 86.8% accuracy at the `dobj` relation


[CLS] It goes on to plug a few diversified Fidelity funds by name . [SEP]
[CLS] It goes on to plug a few diversified Fidelity funds by name . [SEP]
[CLS] It declined to discuss its plans for upgrading its current product line . [SEP]
[CLS] It declined to discuss its plans for upgrading its current product line . [SEP]

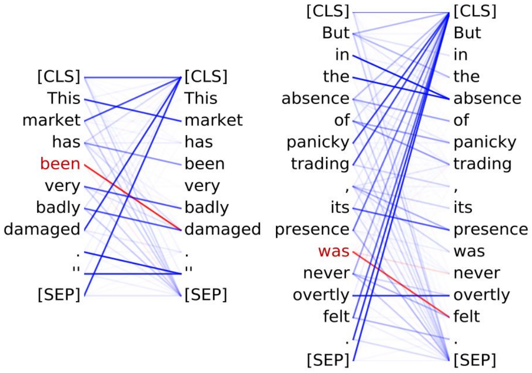
- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation


[CLS] The complicated language in the huge new law has muddied the fight . [SEP]
[CLS] The complicated language in the huge new law has muddied the fight . [SEP]
[CLS] The 45-year-old former General Electric Co. executive figures it will be easier this time . [SEP]
[CLS] The 45-year-old former General Electric Co. executive figures it will be easier this time . [SEP]

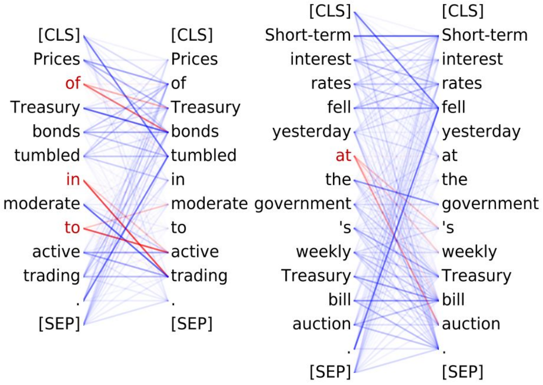
- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the `poss` relation


[CLS] Not his autograph ; power-hitter McGwire 's . [SEP]
[CLS] Not his autograph ; power-hitter McGwire 's . [SEP]
... many employees are working at its giant Renton , Wash. , plant . [SEP]
... many employees are working at its giant Renton , Wash. , plant . [SEP]

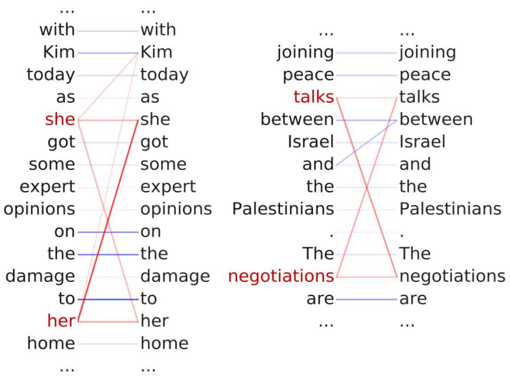
- **Passive auxiliary verbs** attend to the verb they modify
- 82.5% accuracy at the `auxpass` relation


[CLS] This market has been very badly damaged . '' [SEP]
[CLS] This market has been very badly damaged . '' [SEP]
[CLS] But in the absence of panicky trading , its presence was never overtly felt . [SEP]
[CLS] But in the absence of panicky trading , its presence was never overtly felt . [SEP]

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


[CLS] Prices of Treasury bonds tumbled in moderate to active trading . [SEP]
[CLS] Prices of Treasury bonds tumbled in moderate to active trading . [SEP]
[CLS] Short-term interest rates fell yesterday at the government 's weekly Treasury bill auction . [SEP]
[CLS] Short-term interest rates fell yesterday at the government 's weekly Treasury bill auction . [SEP]

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

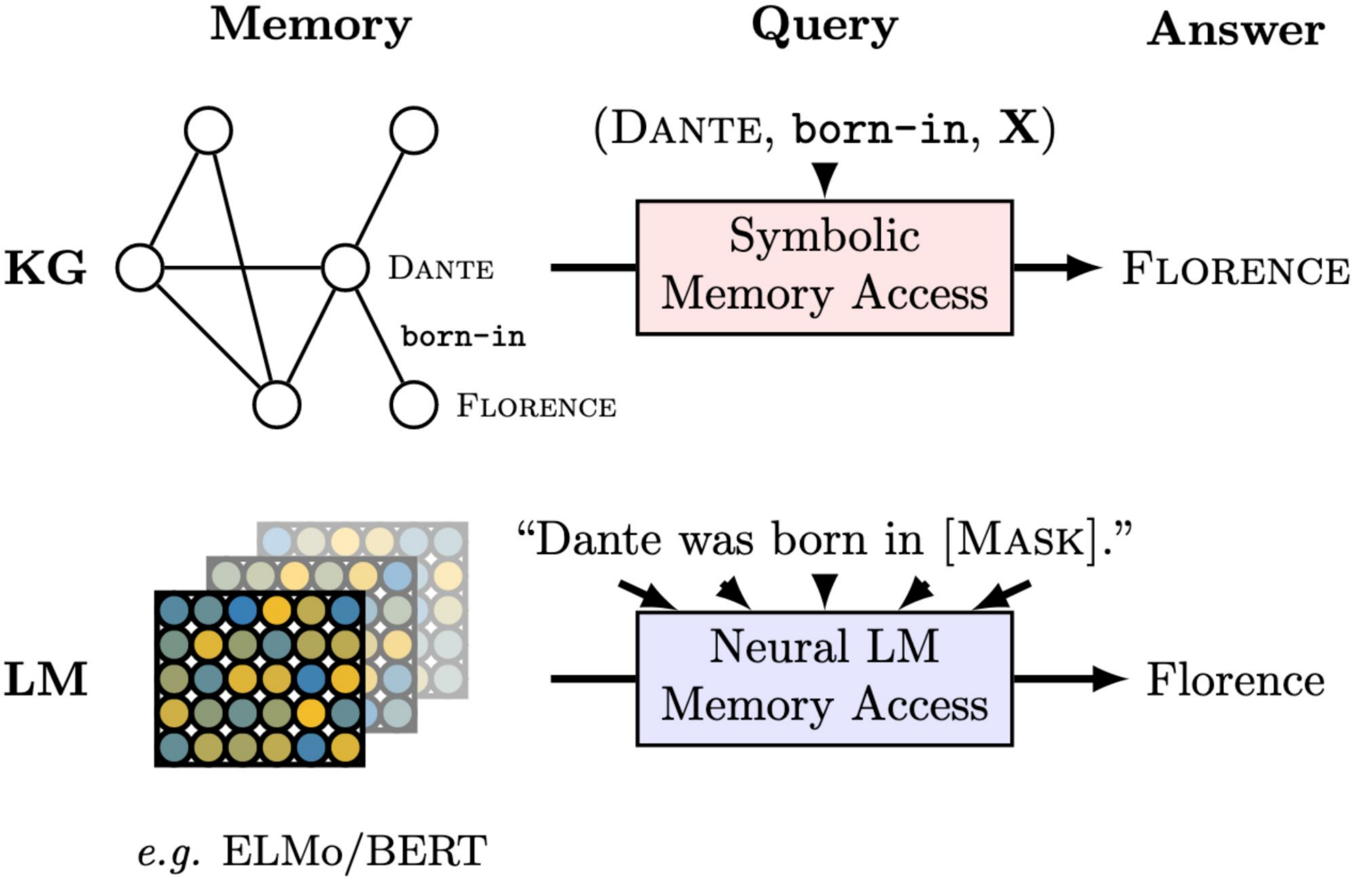

... with Kim today as she got some expert opinions on the damage to her home ...
... with Kim today as she got some expert opinions on the damage to her home ...
... joining peace talks between Israel and the Palestinians . The negotiations are ...
... joining peace talks between Israel and the Palestinians . The negotiations are ...

from “What does BERT look at? An Analysis of BERT’s Attention” by Clark et al.

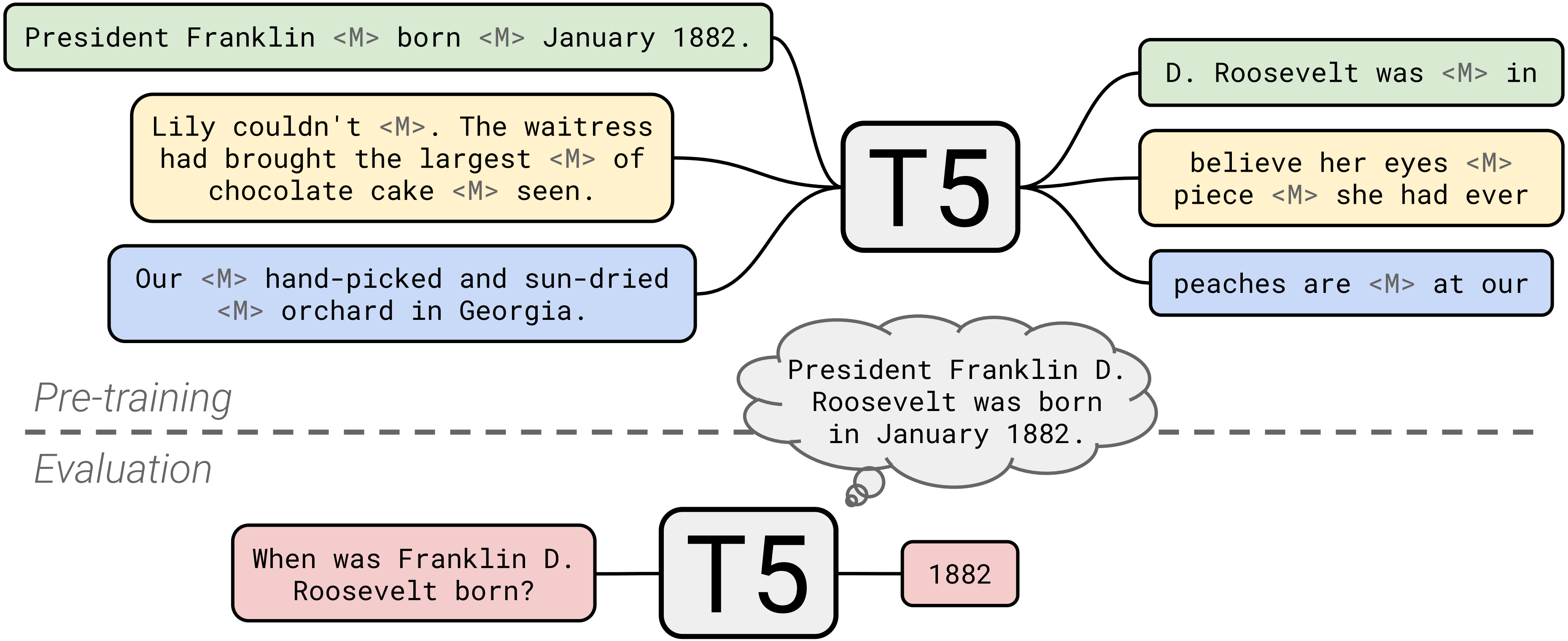
Why does language modeling effectively result in meta-learning?

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

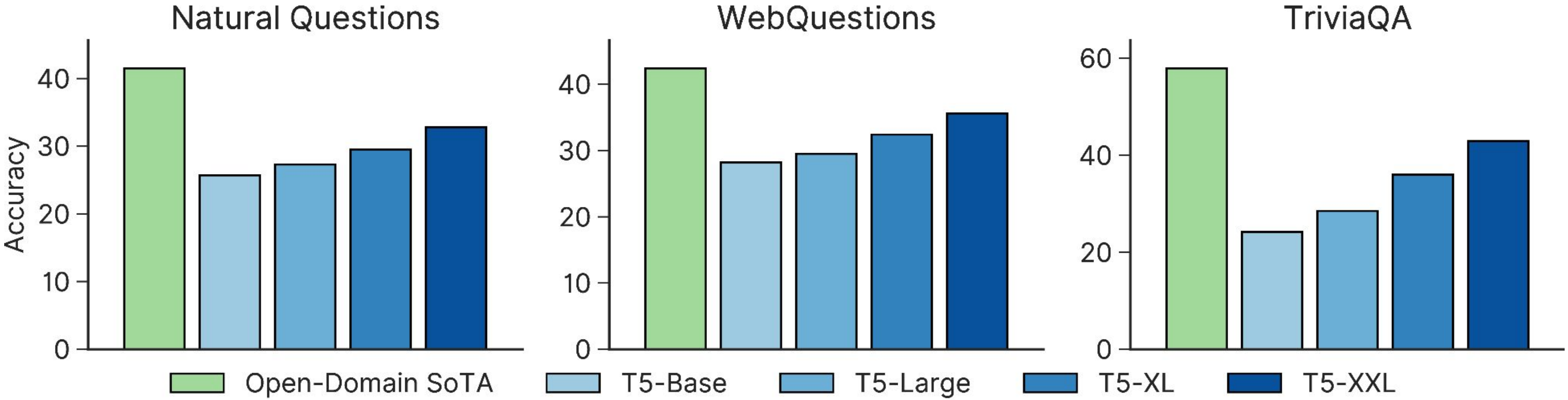
President Franklin D. Roosevelt was born in January 1882.

When was Franklin D. Roosevelt born?

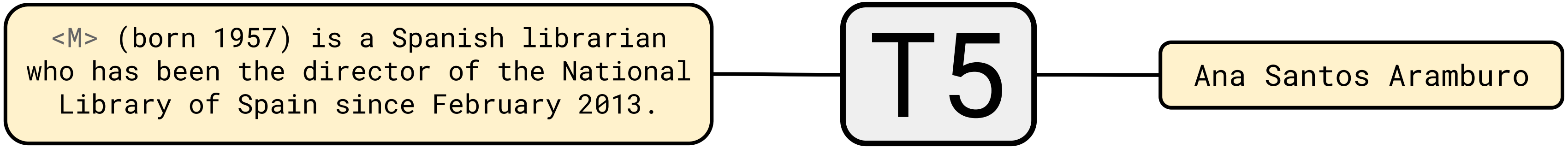
T5

1882

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



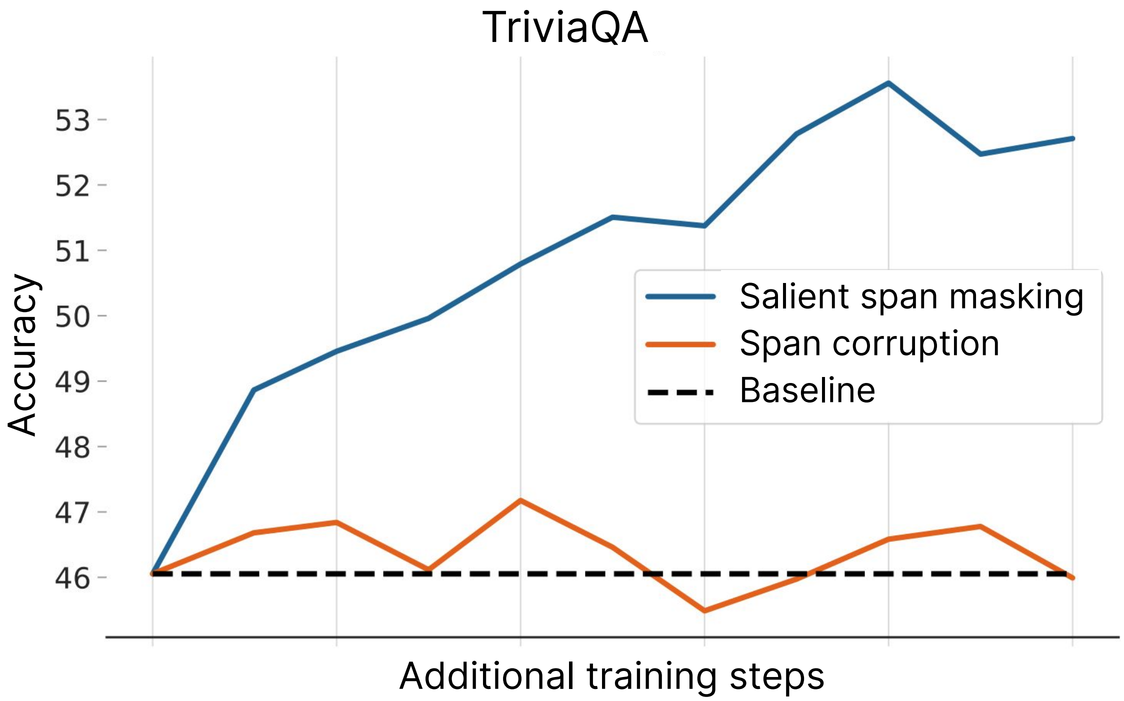
from “How Much Knowledge Can You Pack Into the Parameters of a Language Model” by Roberts et al.



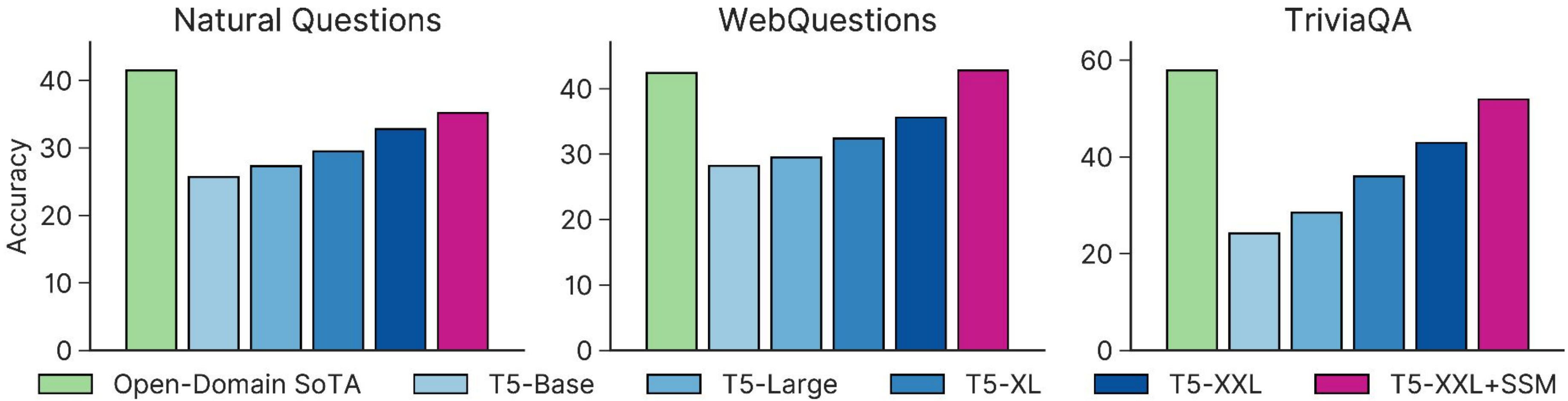
<M> (born 1957) is a Spanish librarian who has been the director of the National Library of Spain since February 2013.

T5

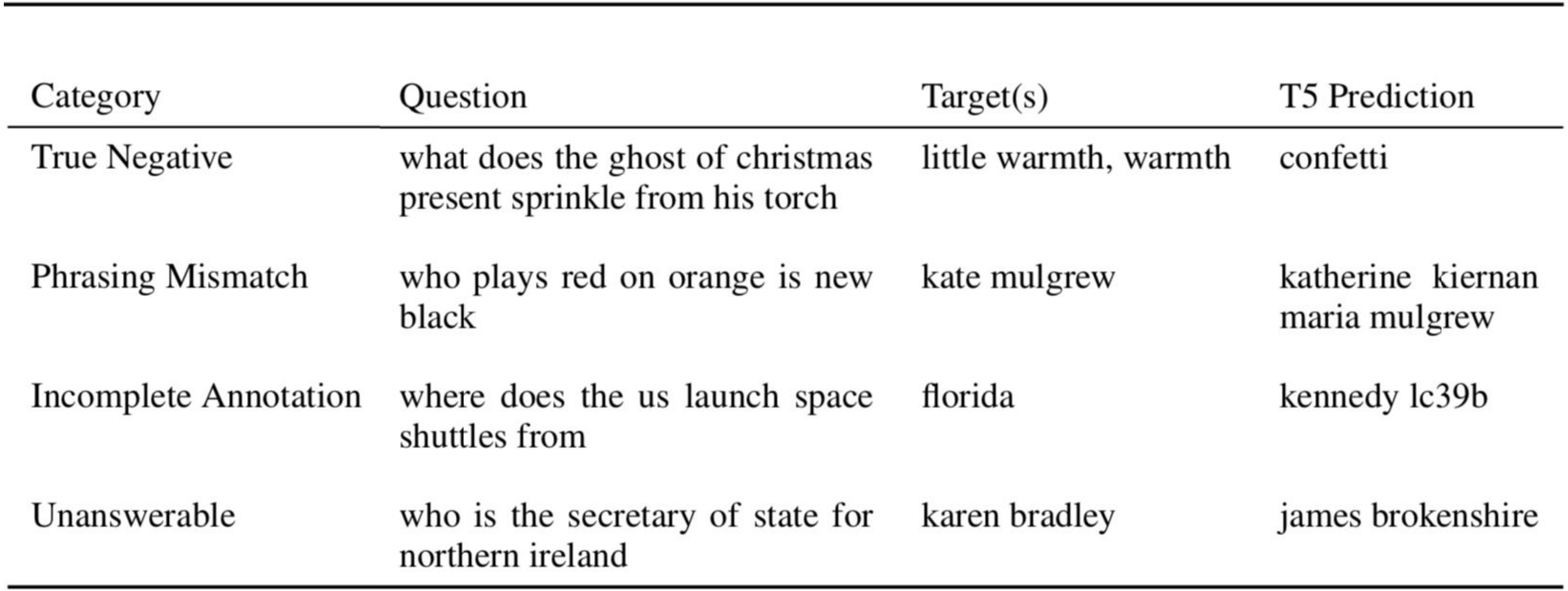
Ana Santos Aramburo



SSM data from "REALM: Retrieval-Augmented Language Model Pre-Training" by Guu et al.

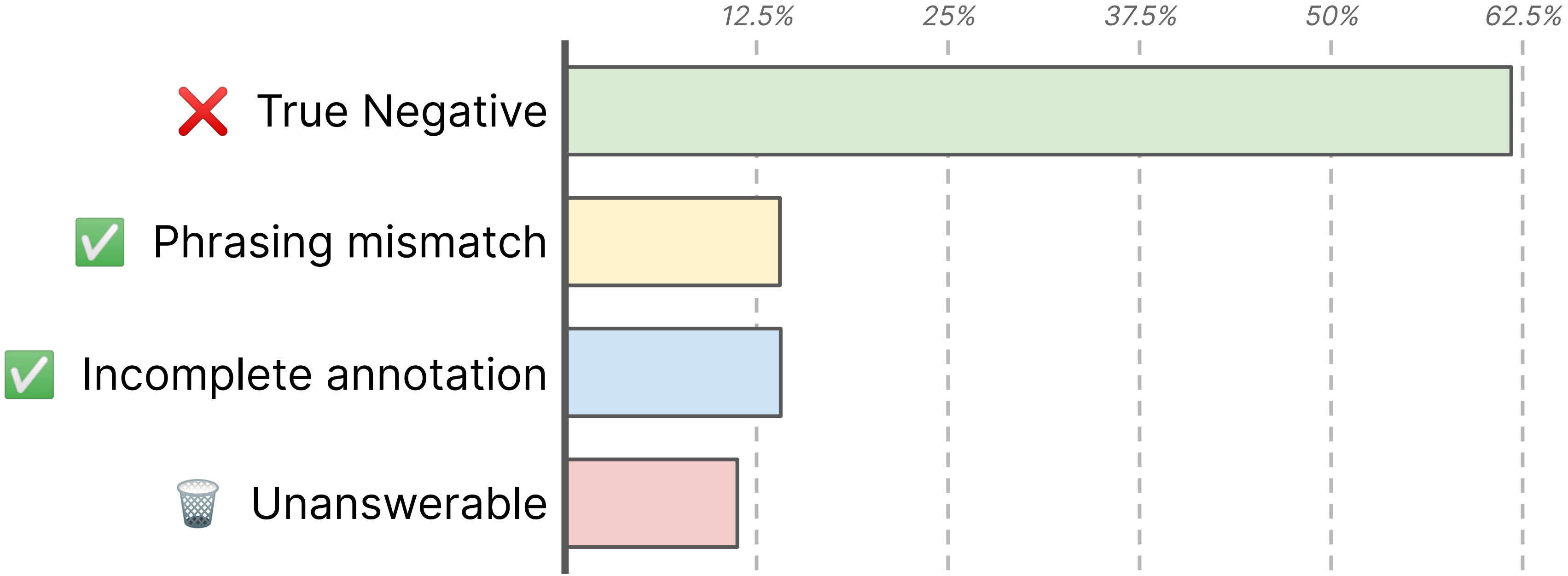


from "How Much Knowledge Can You Pack Into the Parameters of a Language Model" by Roberts et al.



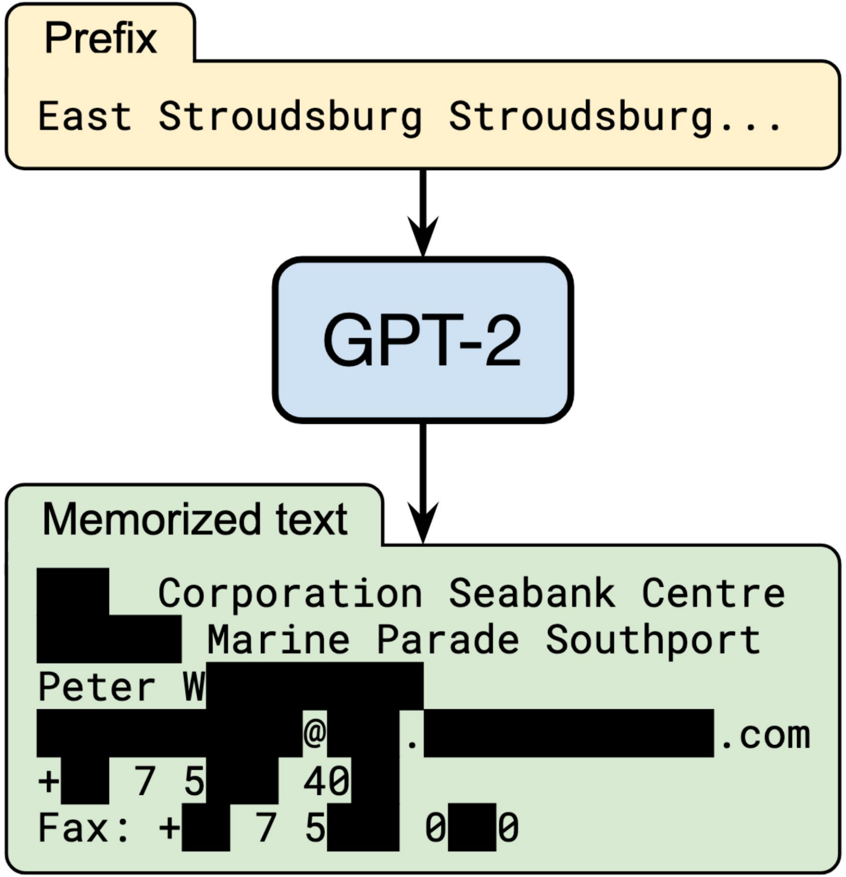
	Category	Question	Target(s)	T5 Prediction
❌	True Negative	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
✅	Phrasing Mismatch	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
✅	Incomplete Annotation	where does the us launch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

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



Exact Match: 36.6 → 57.8%!

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



Prefix

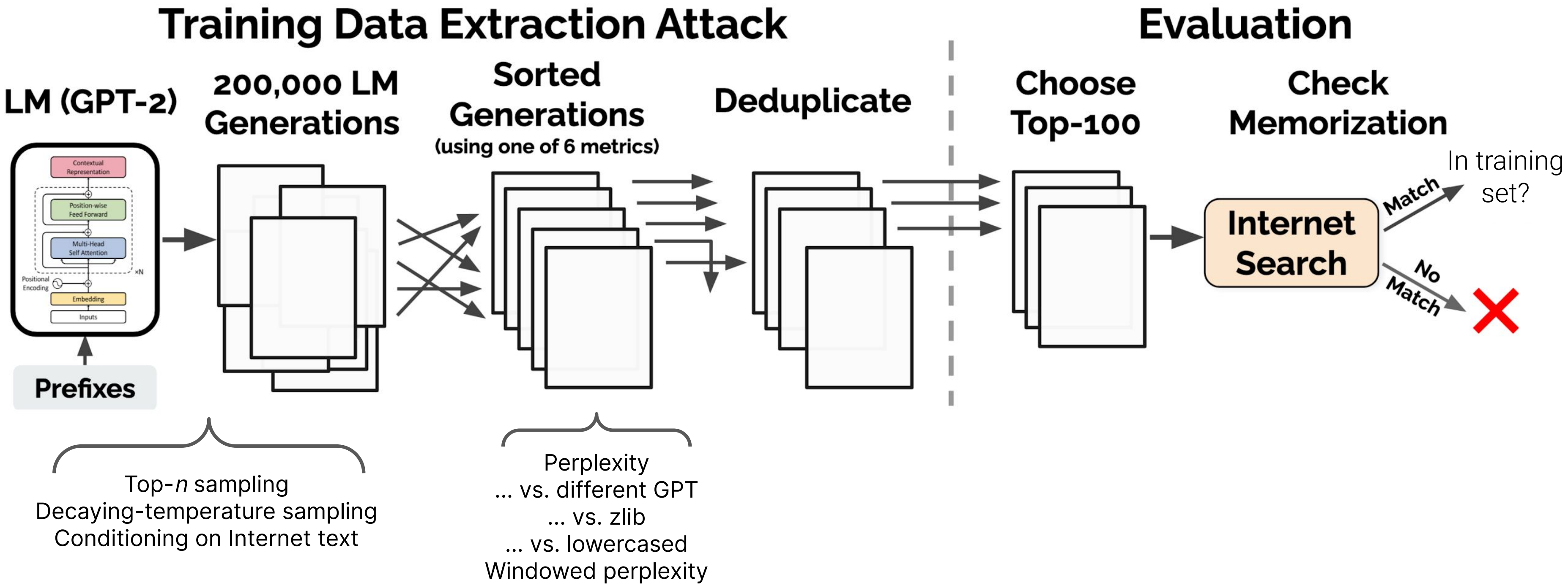
```
East Stroudsburg Stroudsburg...
```

GPT-2

Memorized text

```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

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


Contextual Representation
Position-wise Feed Forward
Multi-Head Self Attention
×N
Positional Encoding
Embedding
Inputs
Match
No Match

Training Data Extraction Attack

Evaluation

LM (GPT-2)

Prefixes

200,000 LM Generations

Sorted Generations (using one of 6 metrics)

Deduplicate

Choose Top-100

Check Memorization

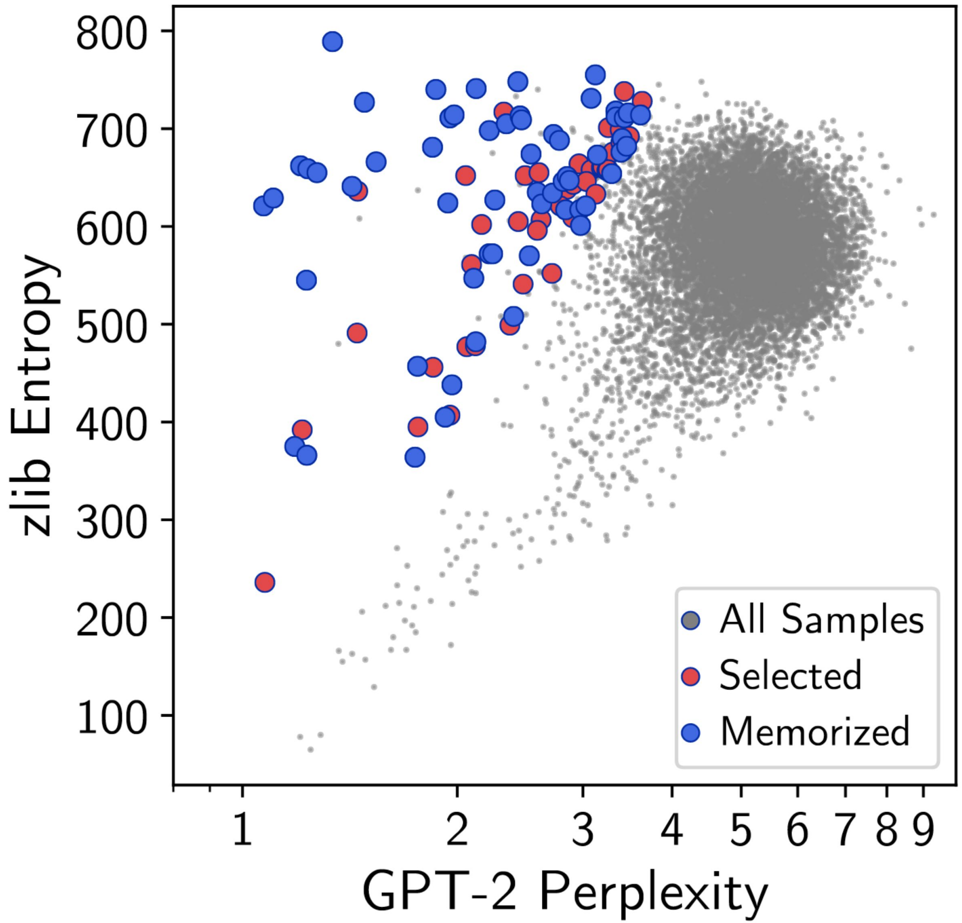
Internet Search

In training set?

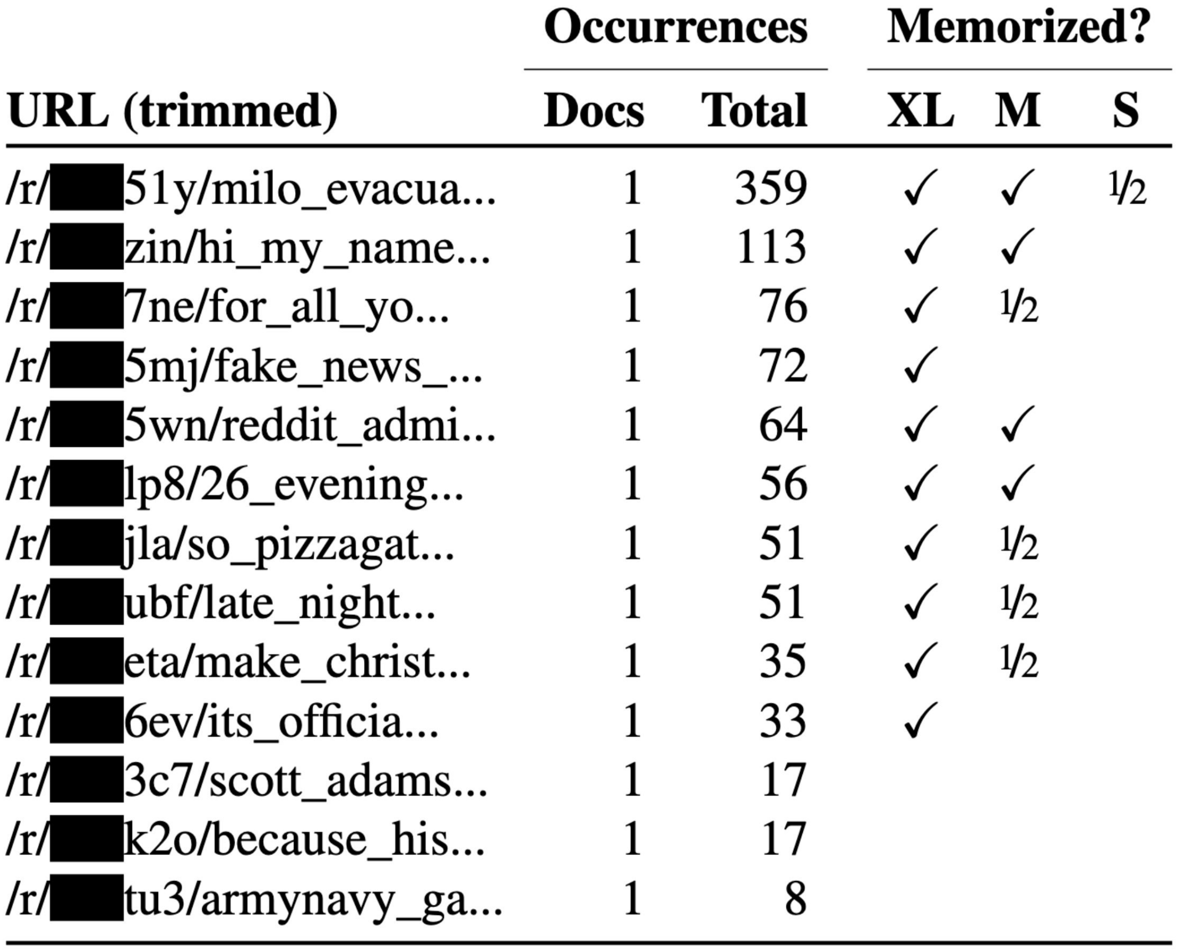
Top-*n* sampling
Decaying-temperature sampling
Conditioning on Internet text

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.



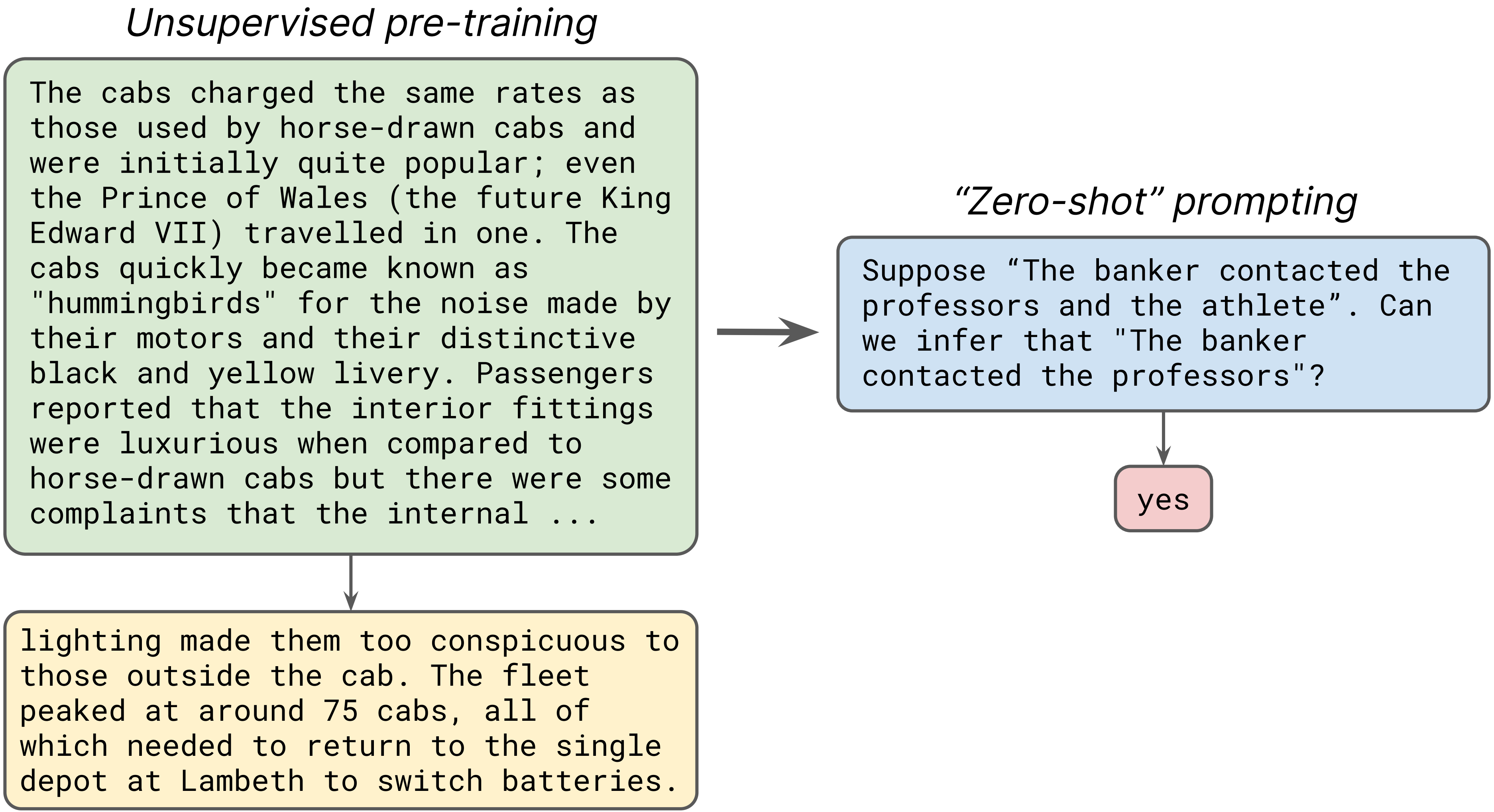
	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/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.

Why does language modeling effectively result in meta-learning?

Language modeling seems to teach models

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



Unsupervised pre-training

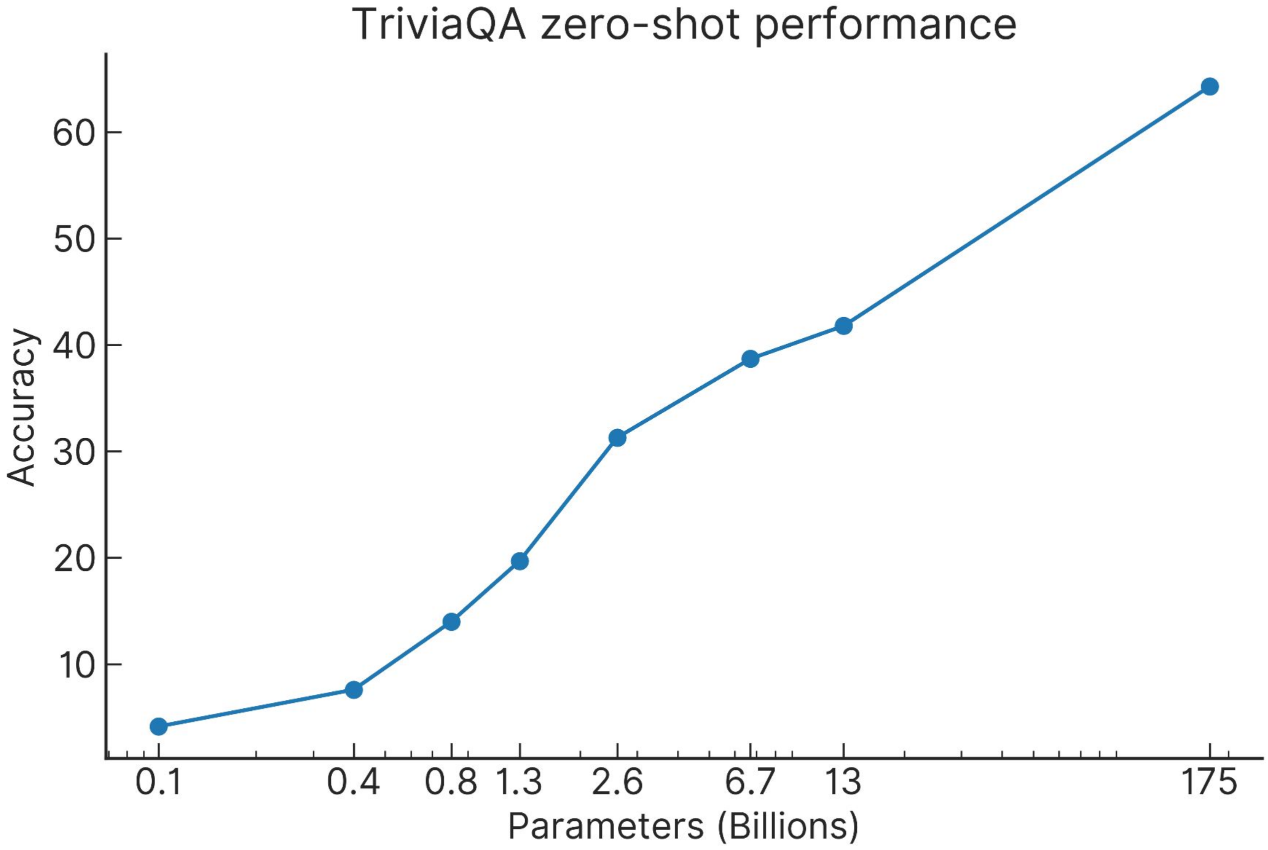
The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular; even the Prince of Wales (the future King Edward VII) travelled in one. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. 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



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

Closed-book question answering

http://www.autosweblog.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

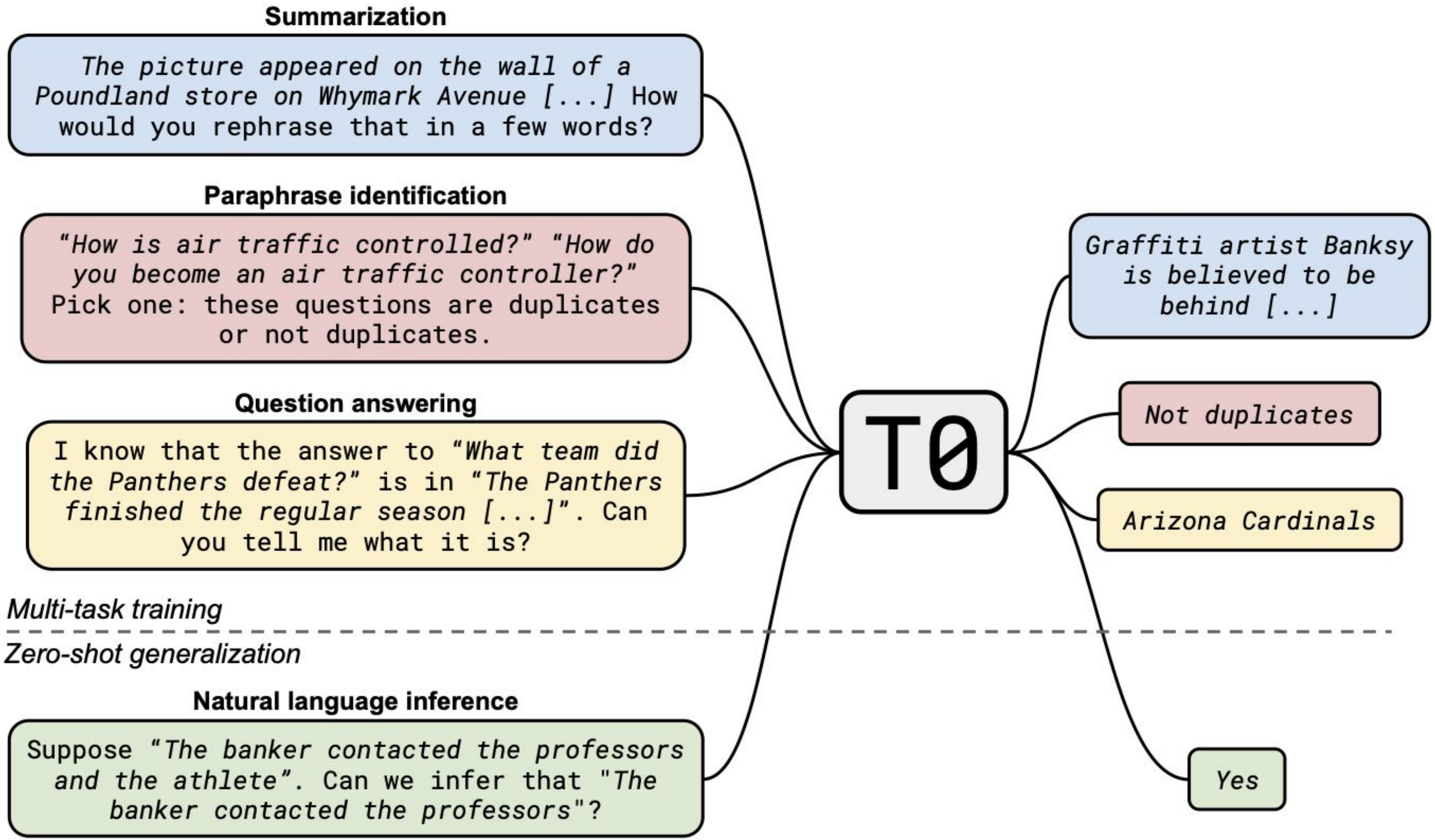
https://blog.nytsoi.net/tag/reddit

... 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

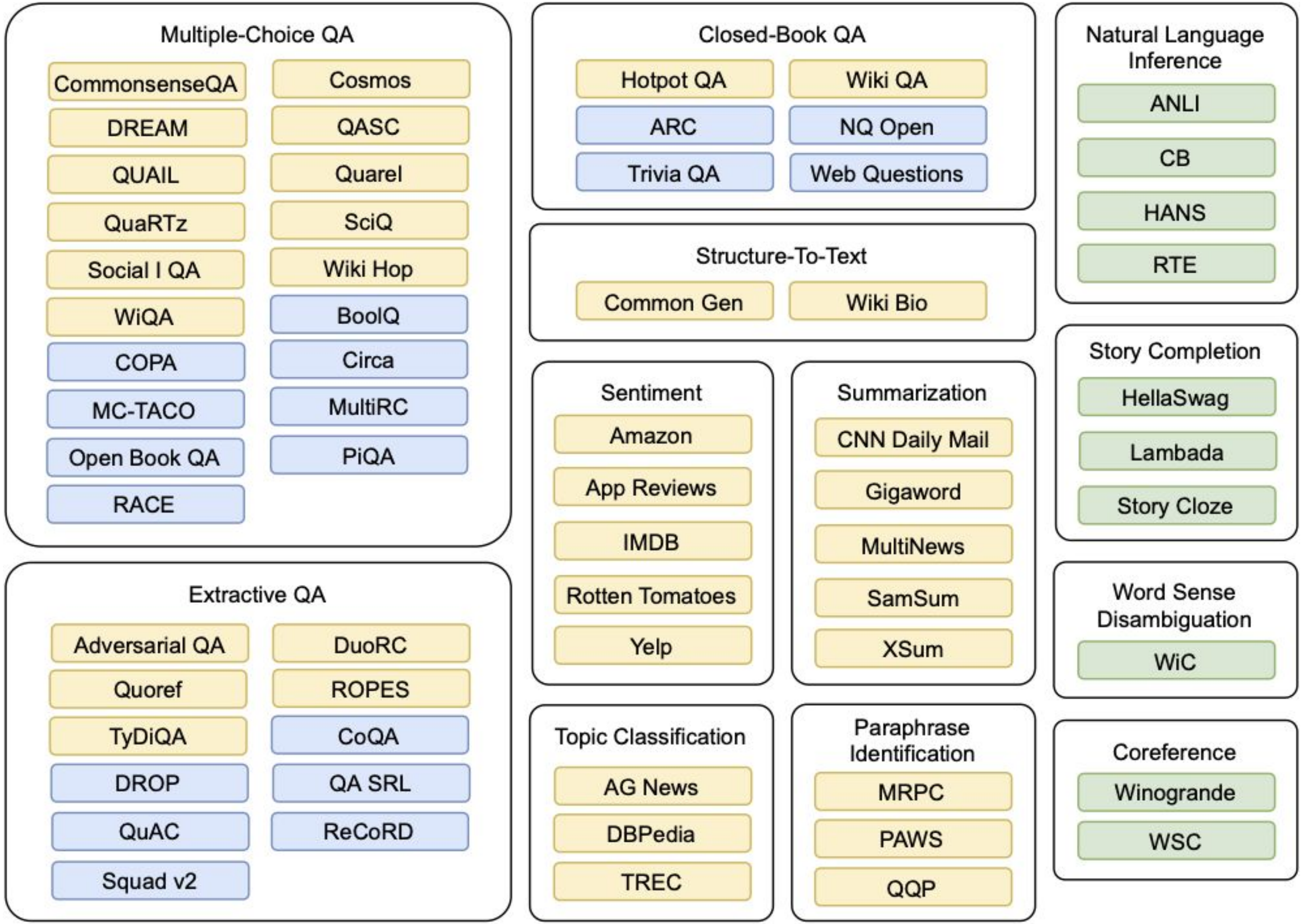
Graffiti artist Banksy is believed to be behind [...]

Not duplicates

Arizona Cardinals

Yes

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



Multiple-Choice QA

CommonsenseQA, DREAM, QUAIL, QuaRTz, Social I QA, WiQA, COPA, MC-TACO, Open Book QA, RACE

Cosmos, QASC, Quarel, SciQ, Wiki Hop, BoolQ, Circa, MultiRC, PiQA

Extractive QA

Adversarial QA, Quoref, TyDiQA, DROP, QuAC, Squad v2

DuoRC, ROPES, CoQA, QA SRL, ReCoRD

Closed-Book QA

Hotpot QA, ARC, Trivia QA

Wiki QA, NQ Open, Web Questions

Structure-To-Text

Common Gen, Wiki Bio

Sentiment

Amazon, App Reviews, IMDB, Rotten Tomatoes, Yelp

Summarization

CNN Daily Mail, Gigaword, MultiNews, SamSum, XSum

Topic Classification

AG News, DBPedia, TREC

Paraphrase Identification

MRPC, PAWS, QQP

Natural Language Inference

ANLI, CB, HANS, RTE

Story Completion

HellaSwag, Lambada, Story Cloze

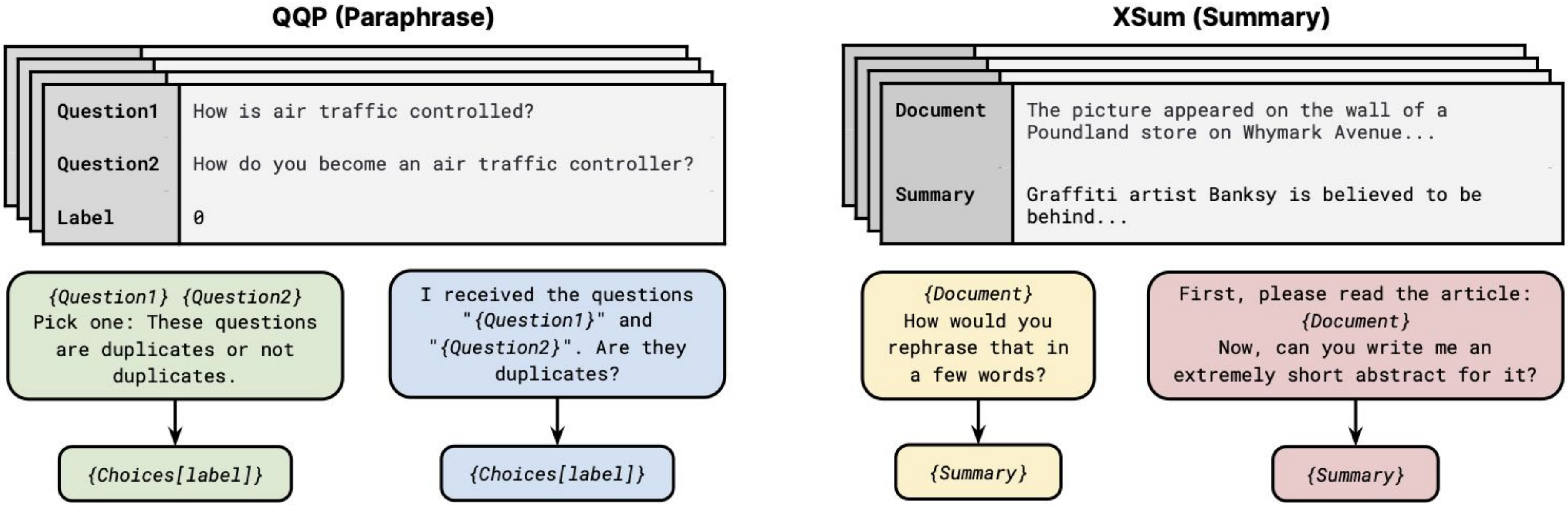
Word Sense Disambiguation

WiC

Coreference

Winogrande, WSC

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...

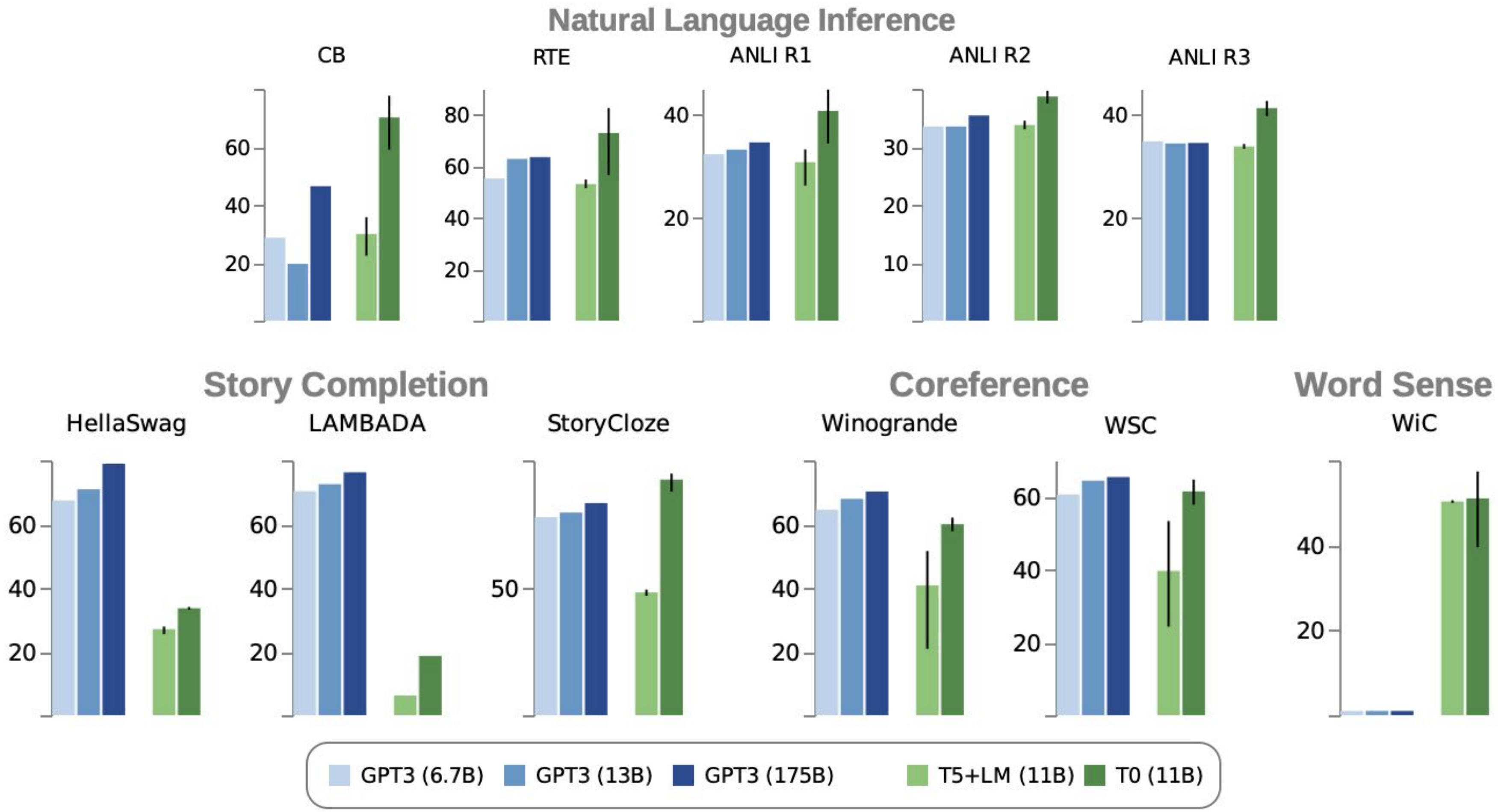
```
{Document}
How would you
rephrase that in
a few words?
```

```
{Summary}
```

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

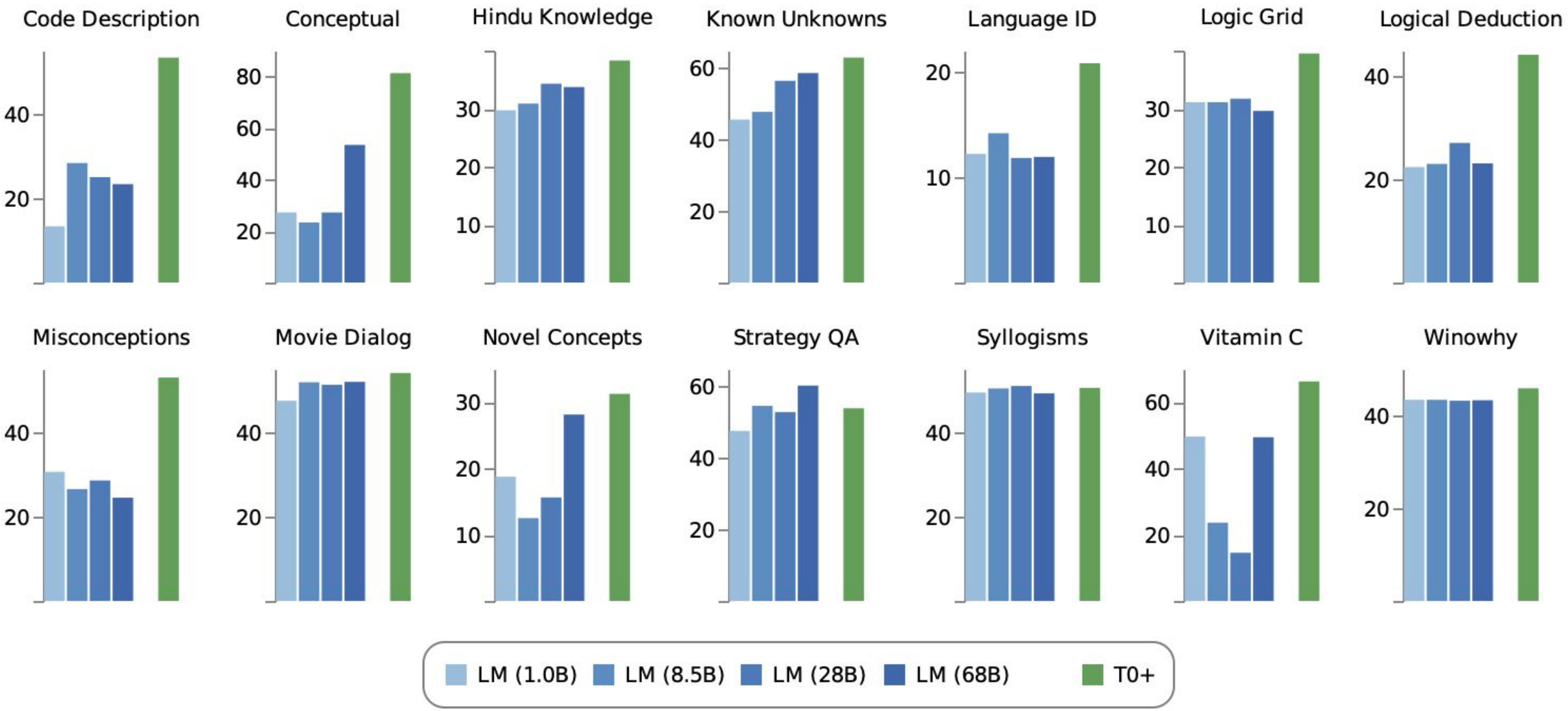
```
{Summary}
```

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.

Big Bench



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

Why does language modeling effectively result in meta-learning?

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

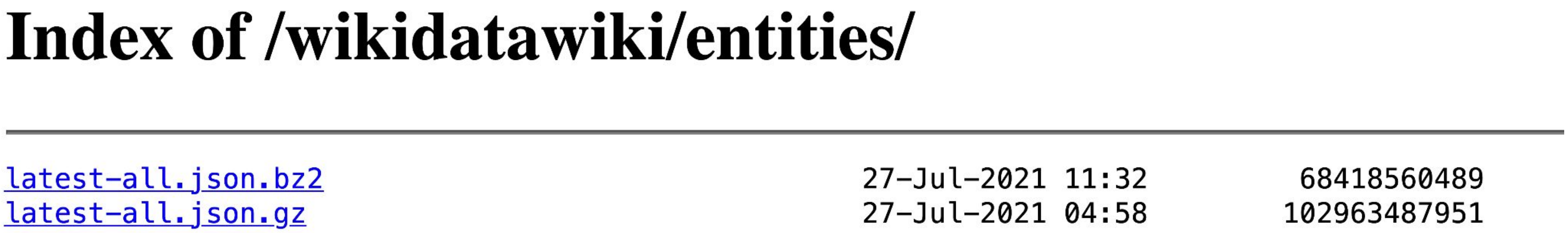
- World knowledge
- Esoteric facts
- Tasks (maybe?)

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?



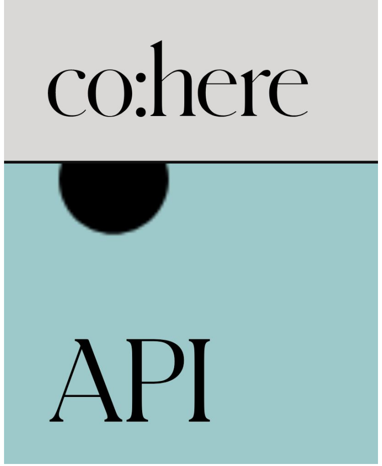
Index of /wikidatawiki/entities/

```
latest-all.json.bz2                                  27-Jul-2021 11:32         68418560489
latest-all.json.gz                                   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.

From “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜” by Bender et al.



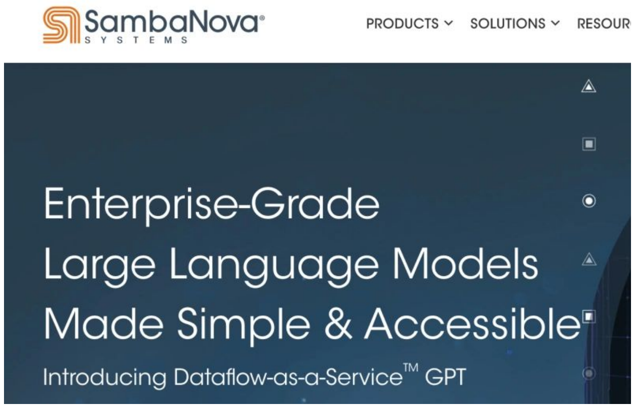
co:here

API

AI21 studio

Custom language models built for scale

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

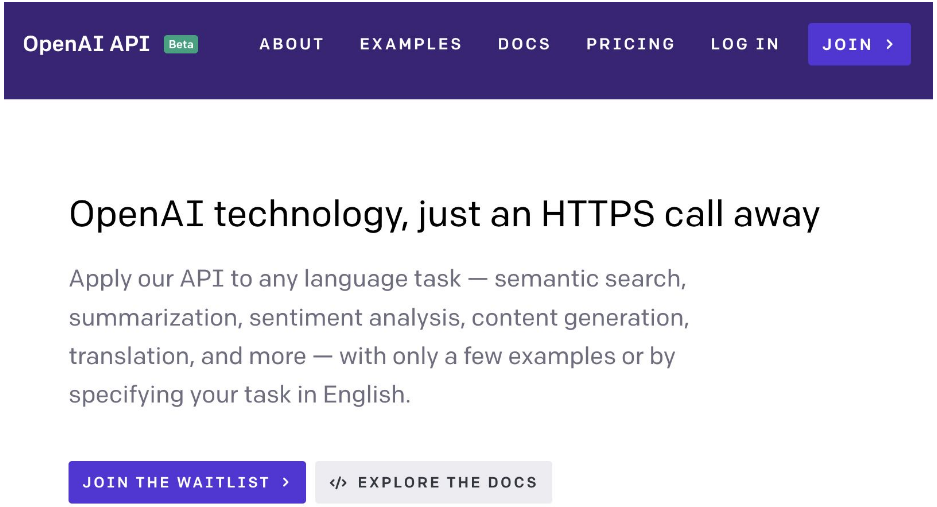


Enterprise-Grade
Large Language Models
Made Simple & Accessible

Introducing Dataflow-as-a-Service™ GPT

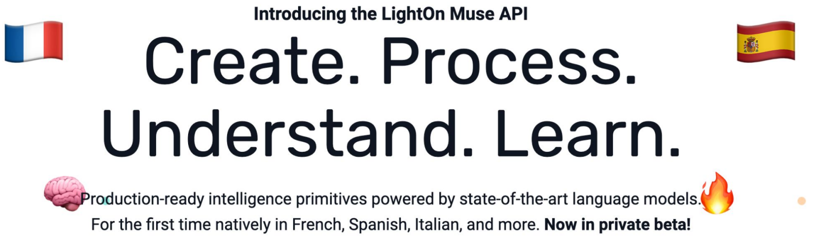
Microsoft Megatron-Turing NLG 530B

The World's Largest and Most Powerful Generative Language Model



OpenAI technology, just an HTTPS call away

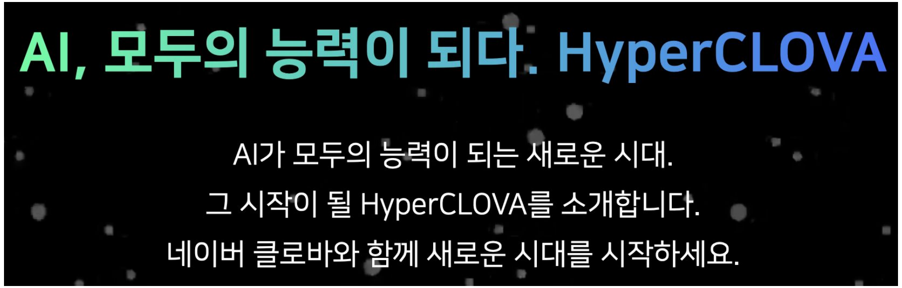
Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.



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