

Building Better Language Models

Colin Raffel

<word

form="appalling"

cornetto_synset_id="n_a-513370"

wordnet_id="a-00193367"

pos="JJ"

sense="causing consternation"

polarity="-0.6"

subjectivity="1.0"

intensity="1.0"

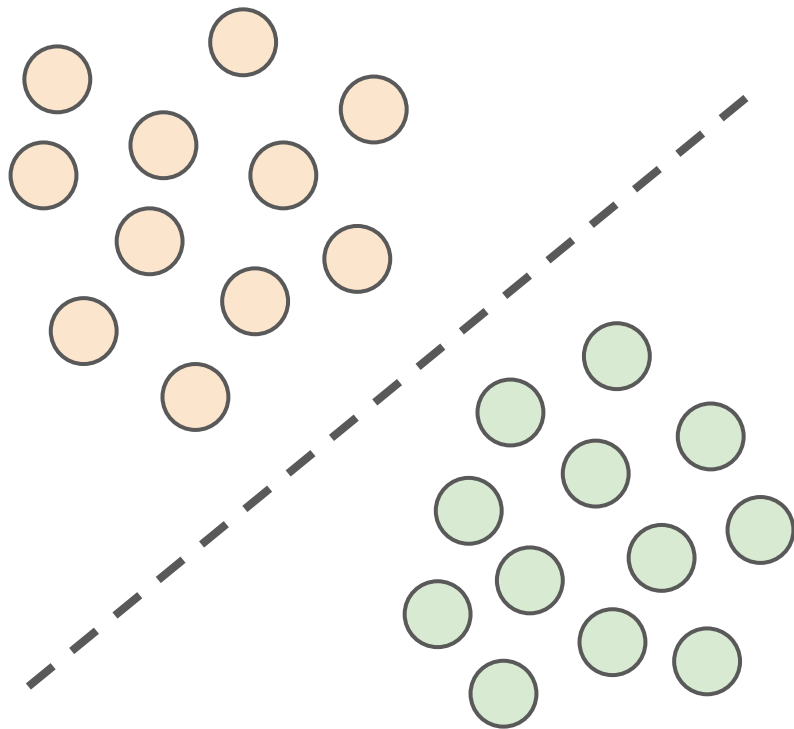
confidence="0.8"

/>

```

a = []
for w, pos in words:
    if w in self and pos in self[w]:
        # Known word not preceded by a modifier ("good").
        if m is None:
            a.append(dict(w=[w], p=p, s=s, i=i, n=1))
        # Known word preceded by a modifier ("really good").
        if m is not None:
            a[-1]["w"].append(w)
            a[-1]["p"] = max(-1.0, min(p * a[-1]["i"], +1.0))
            a[-1]["s"] = max(-1.0, min(s * a[-1]["i"], +1.0))
            a[-1]["i"] = i
        # Known word preceded by a negation ("not really good").
        if n is not None:
            a[-1]["w"].insert(0, n)
            a[-1]["i"] = 1.0 / a[-1]["i"]
            a[-1]["n"] = -1
polarity = avg([(w, p) for w, p, s, x in a])

```



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?

Natural language inference

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

LM

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

Not duplicates

Arizona Cardinals

Yes

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.

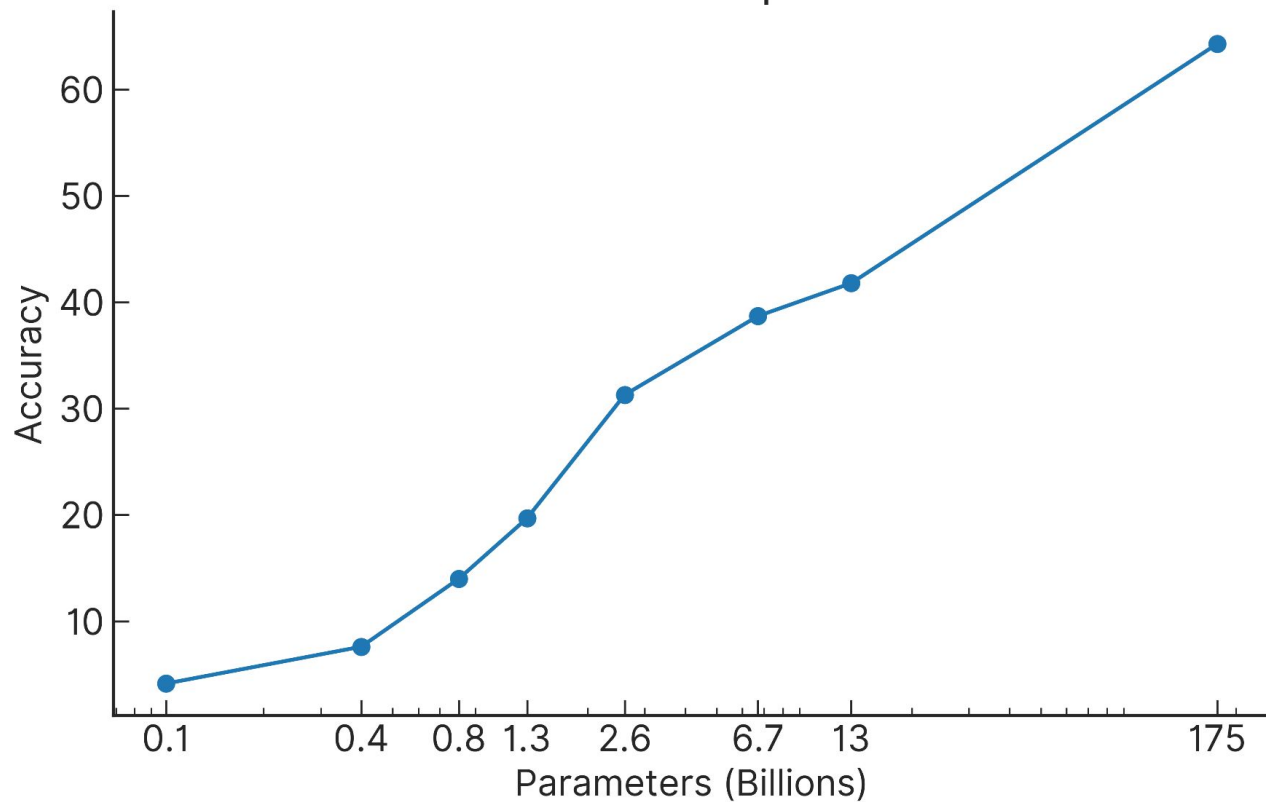


"Zero-shot" prompting

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.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

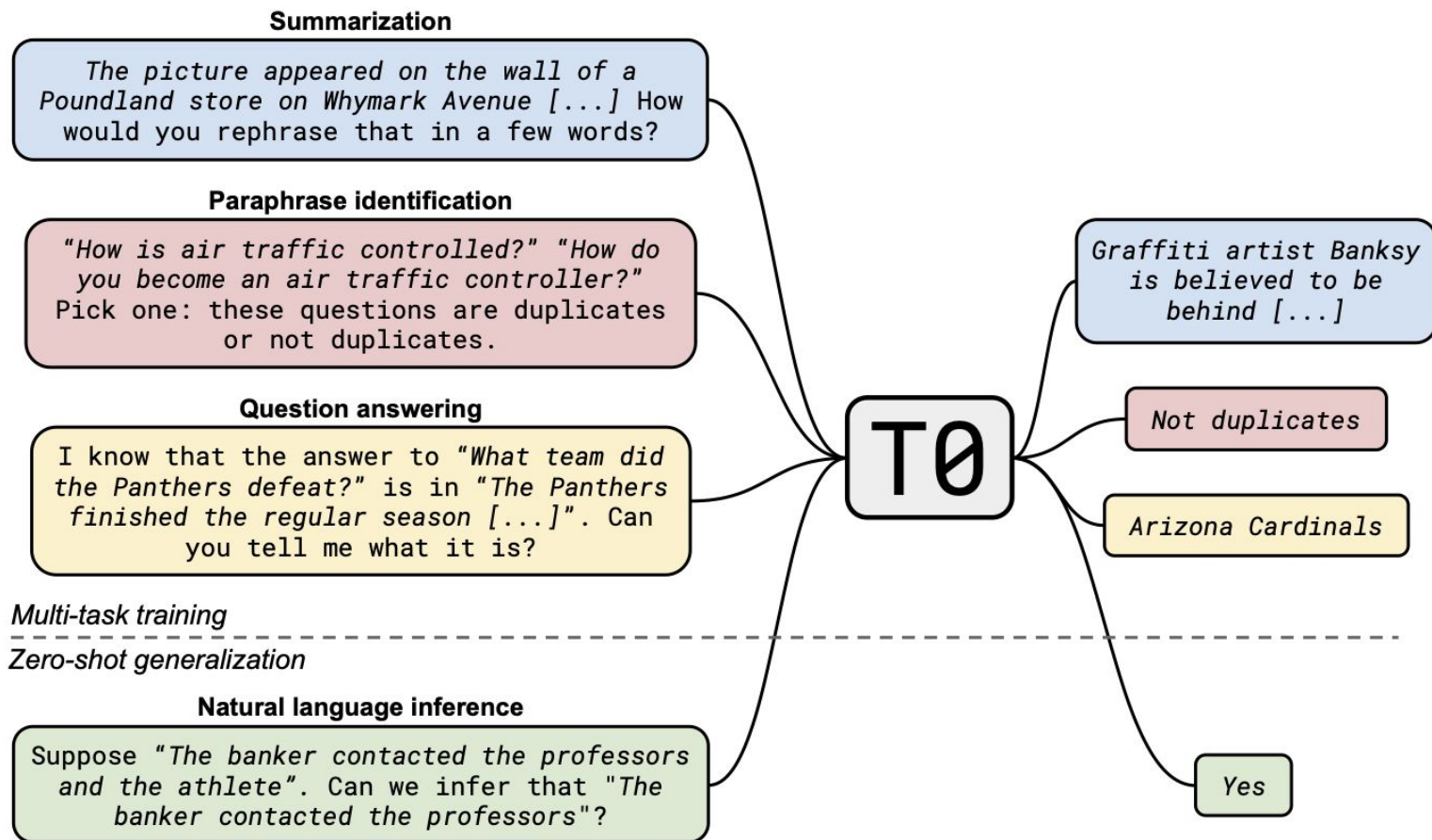
<https://blog.nytsai.net/tag/reddit>

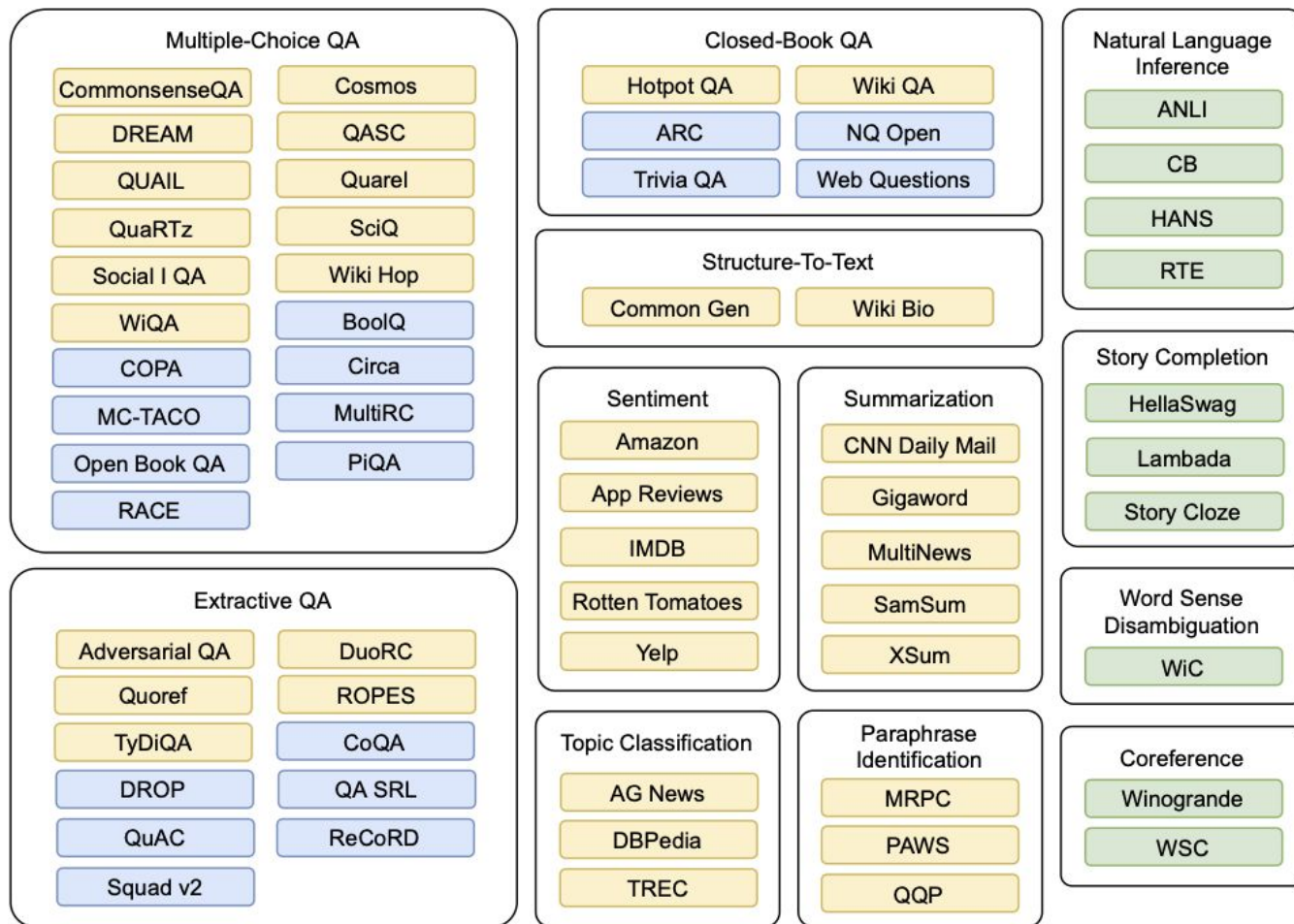
... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade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





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...

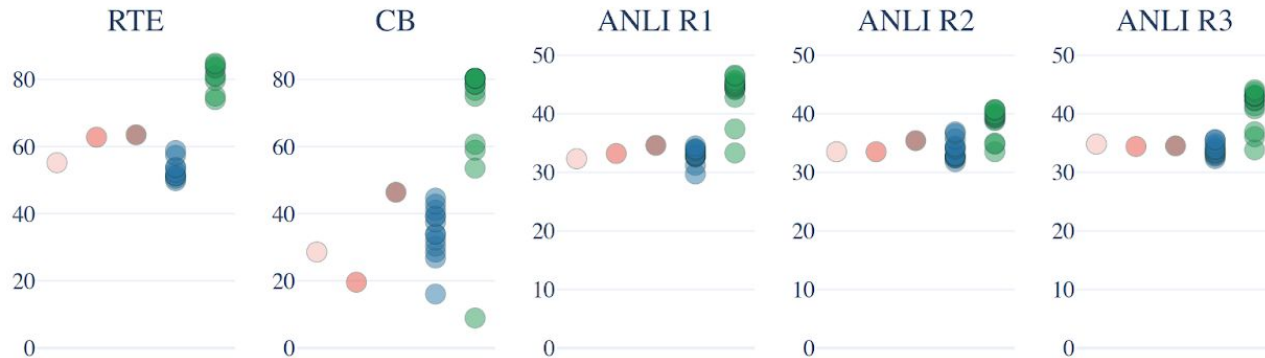
{Document}
How would you
rephrase that in
a few words?

{Summary}

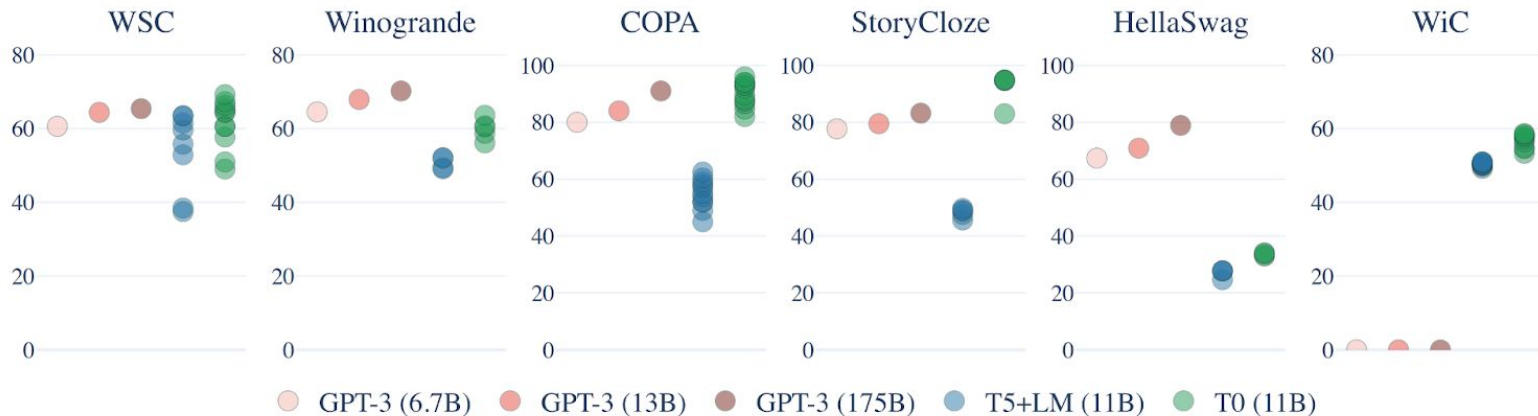
First, please read the article:
{Document}
Now, can you write me an
extremely short abstract for it?

{Summary}

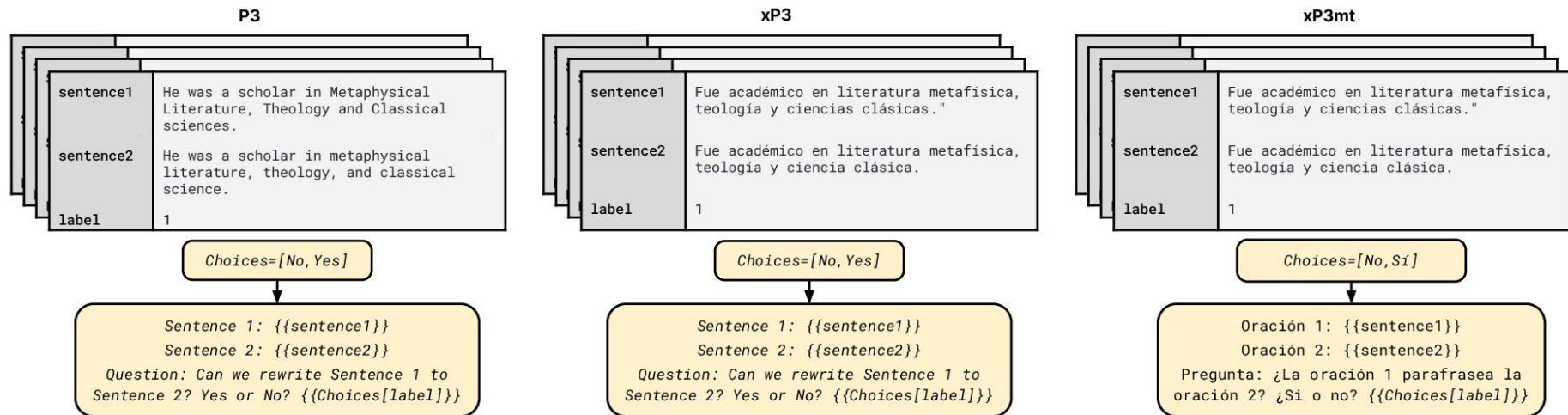
Natural Language Inference



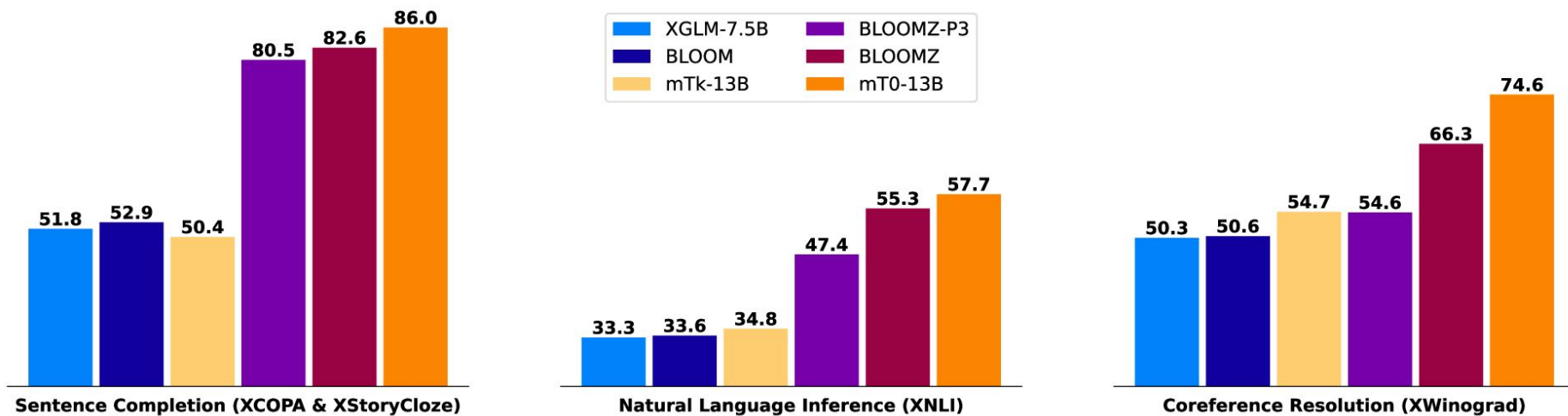
Coreference Resolution



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

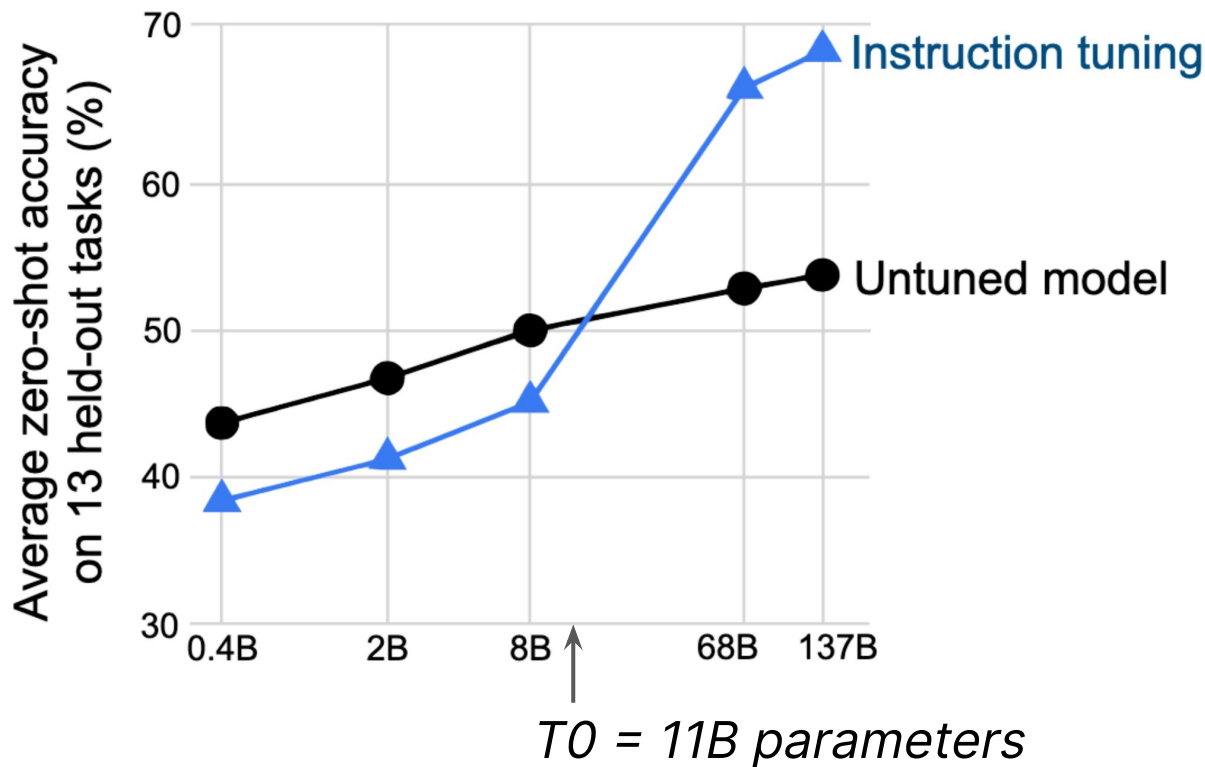


Multilingual Multitask Generalization

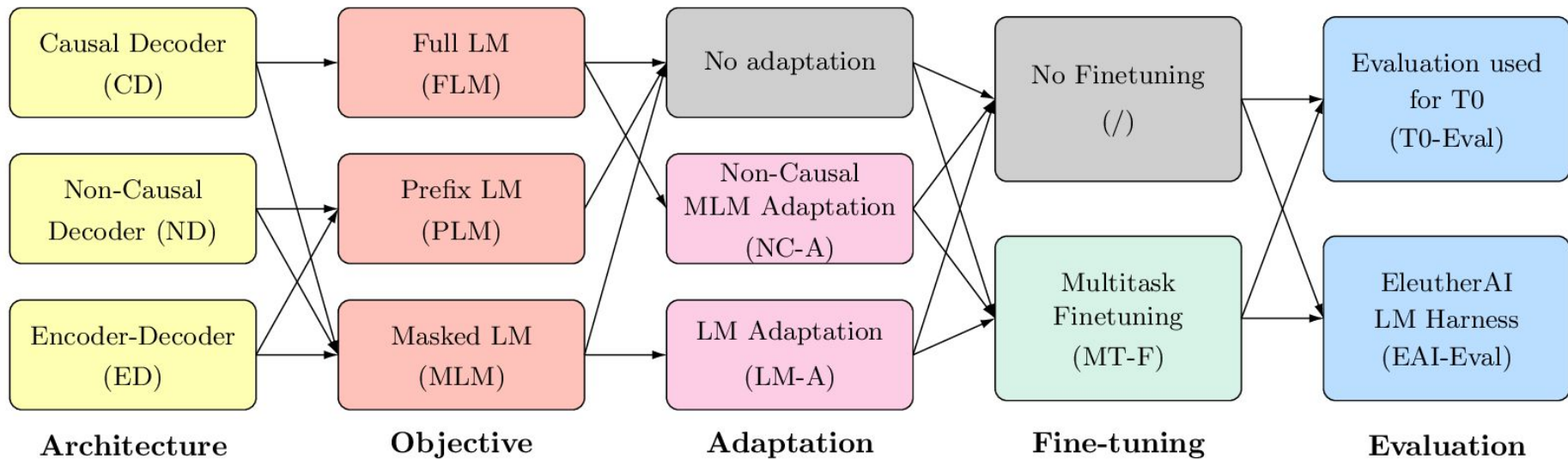


From "Crosslingual Generalization through Multitask Finetuning" by Muennighoff et al.

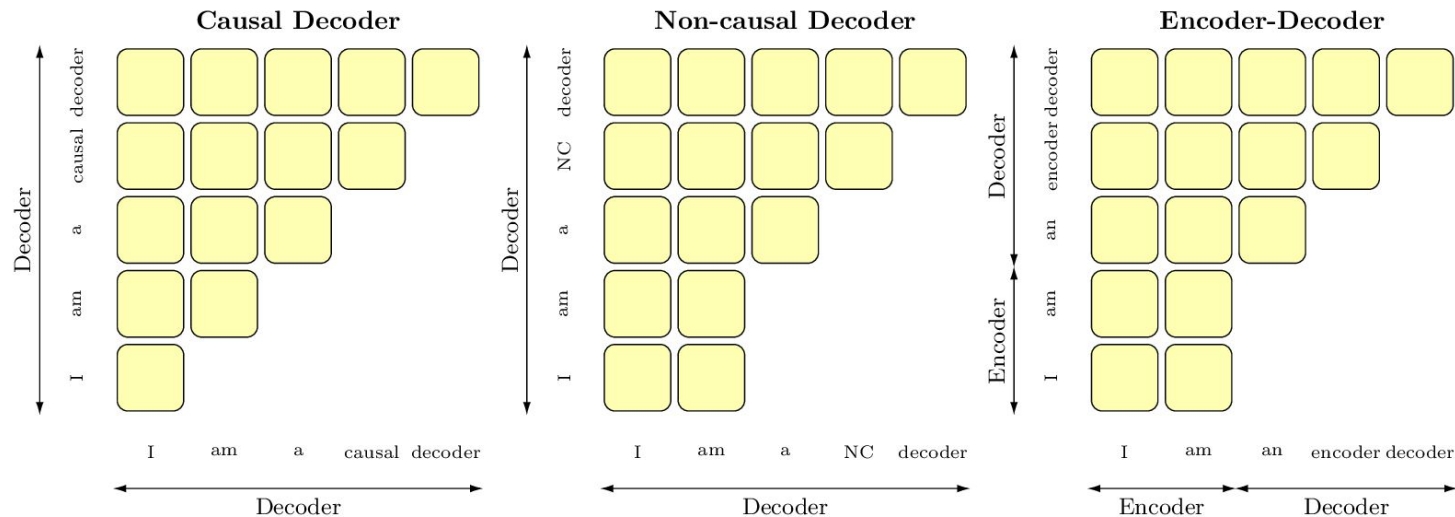
Performance on held-out tasks



From "Fine-Tuned Language Models are Zero-Shot Learners" by Wei et al.



From "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?" by Wang et al.



Full Language Modeling

May

the force be with you

Prefix Language Modeling

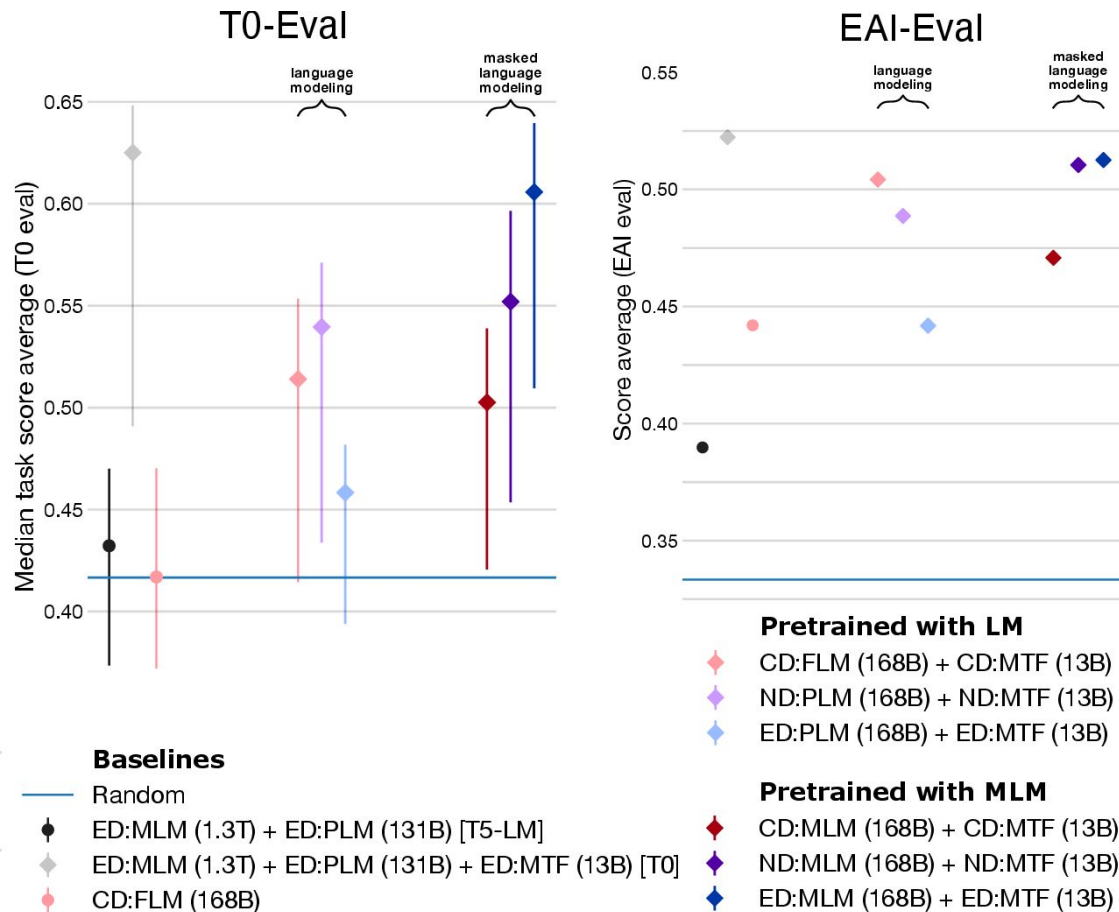
May the force

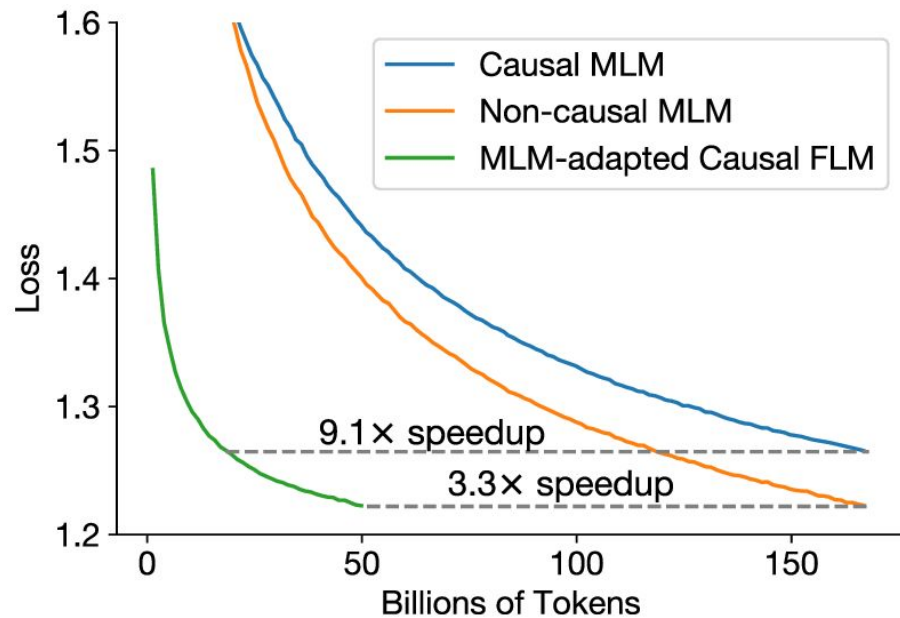
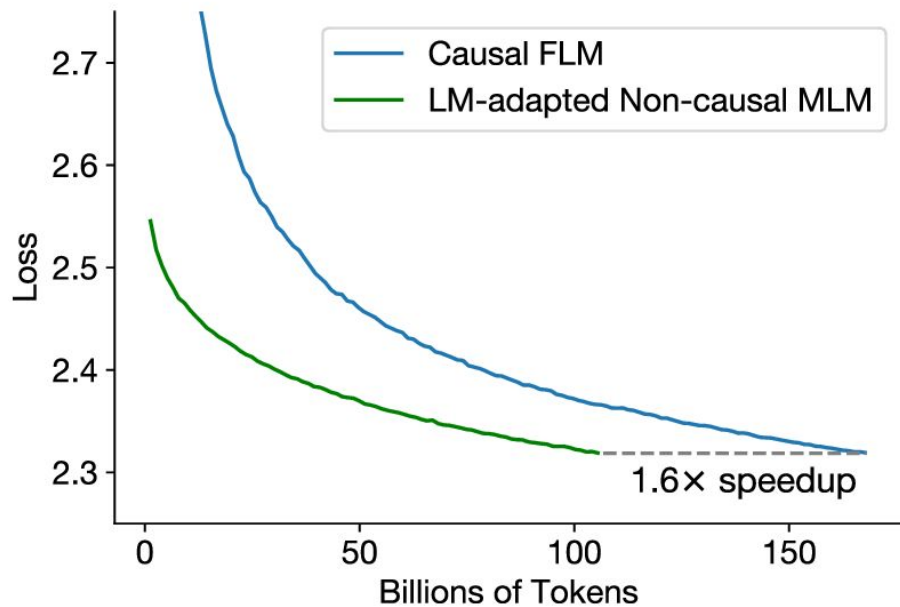
be with you

Masked Language Modeling

May

the force be with you





Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
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
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1 sea otter => loutre de mer ← example #1
```



gradient update



```
1 peppermint => menthe poivrée ← example #2
```



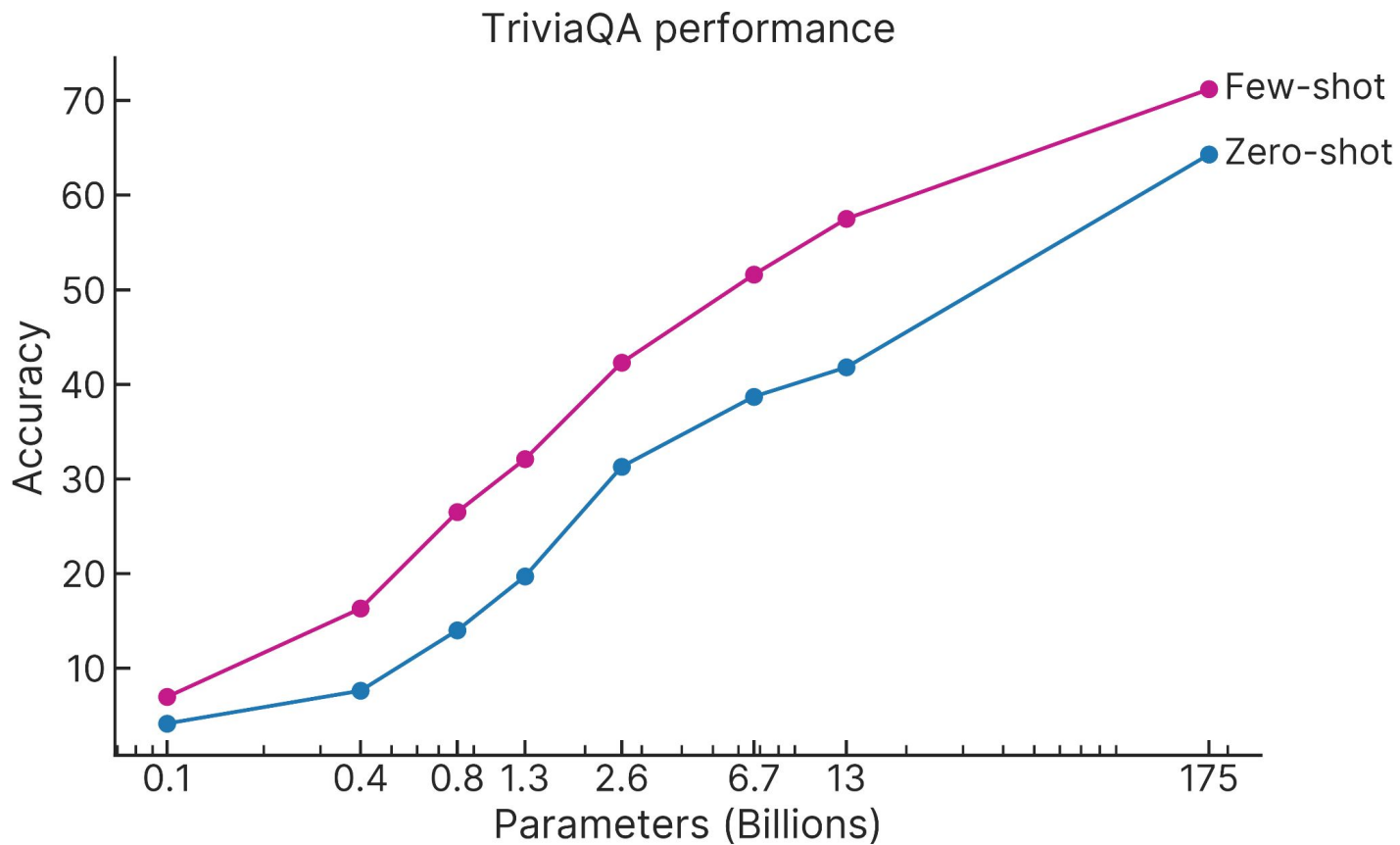
gradient update



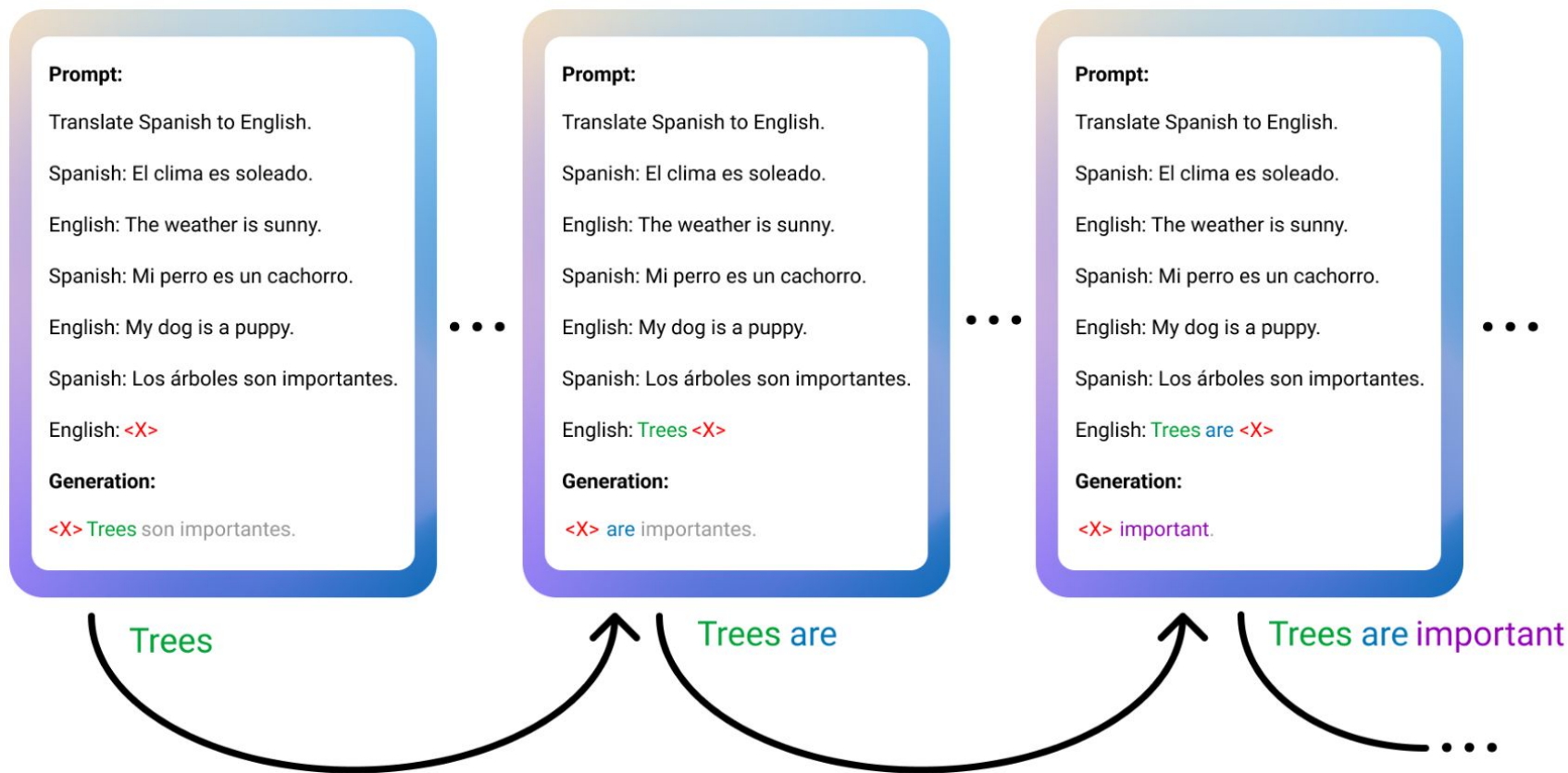
```
1 plush giraffe => girafe peluche ← example #N
```

gradient update

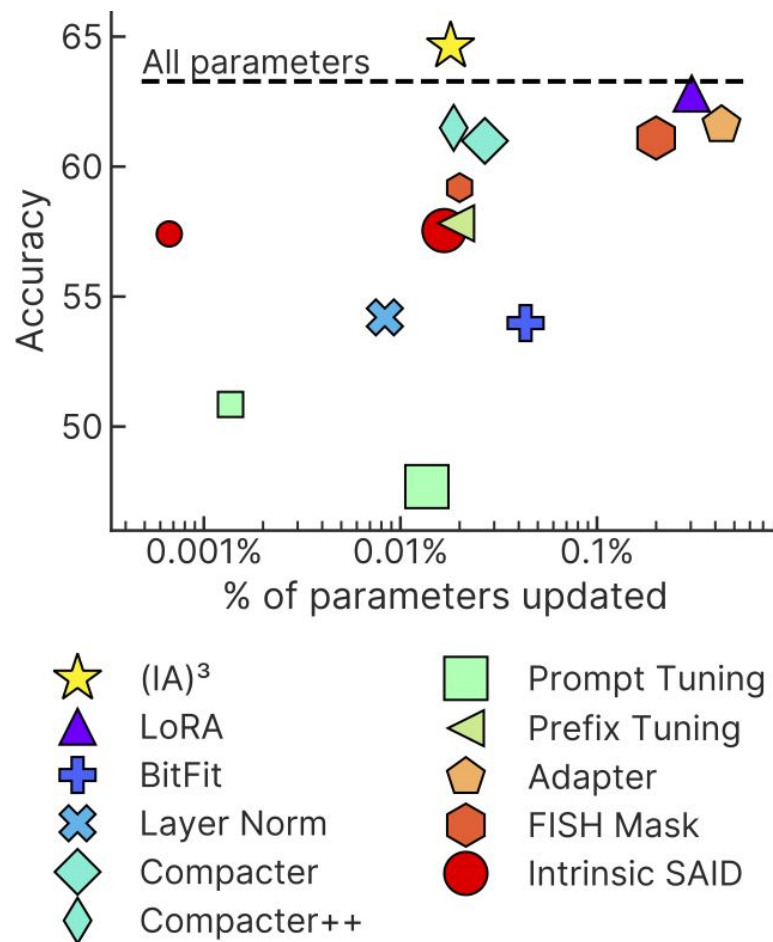
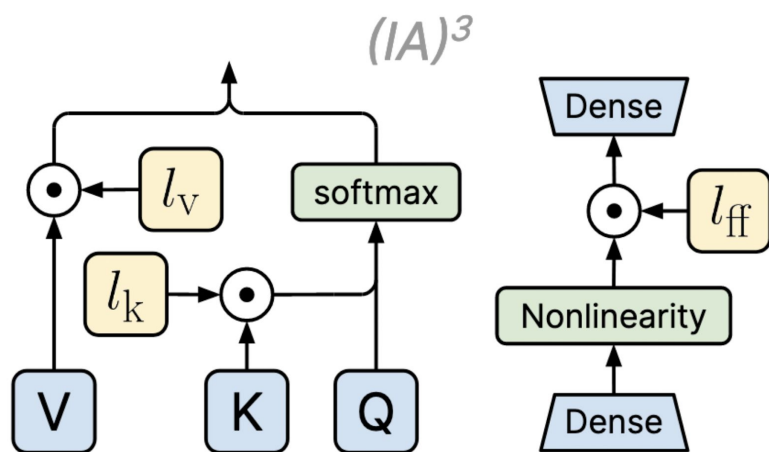
```
1 cheese => ..... ← prompt
```



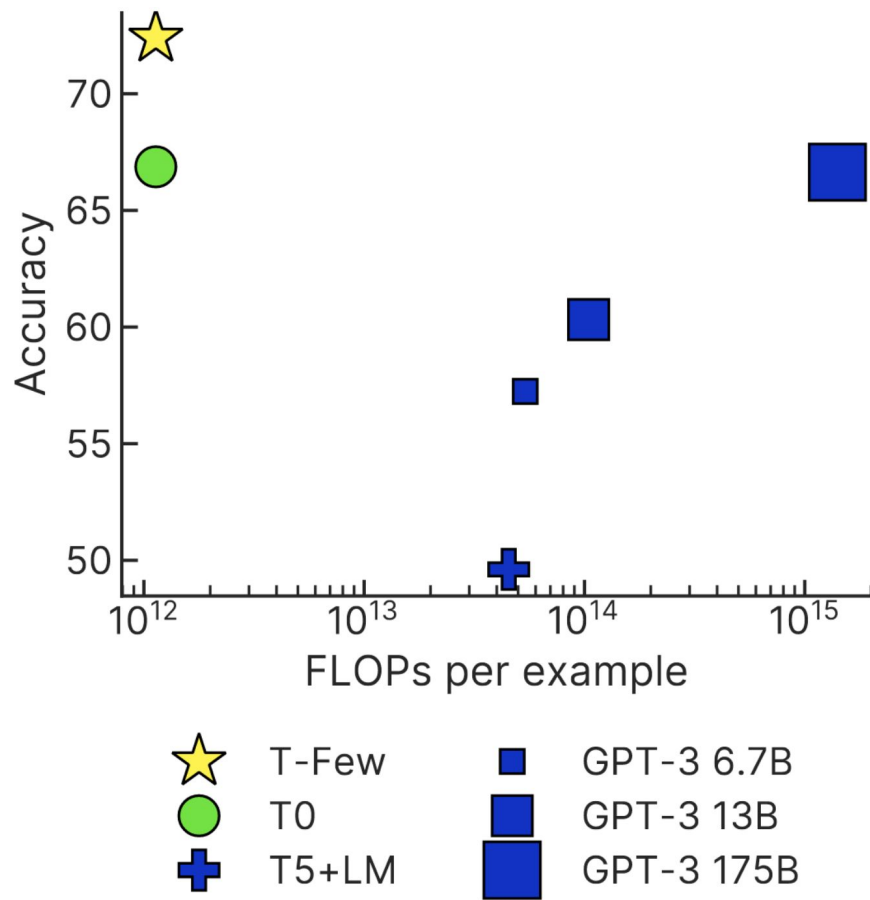
From "Language Models are Few-Shot Learners" by Brown et al.



From "Bidirectional Language Models Are Also Few-shot Learners" by Patel et al.



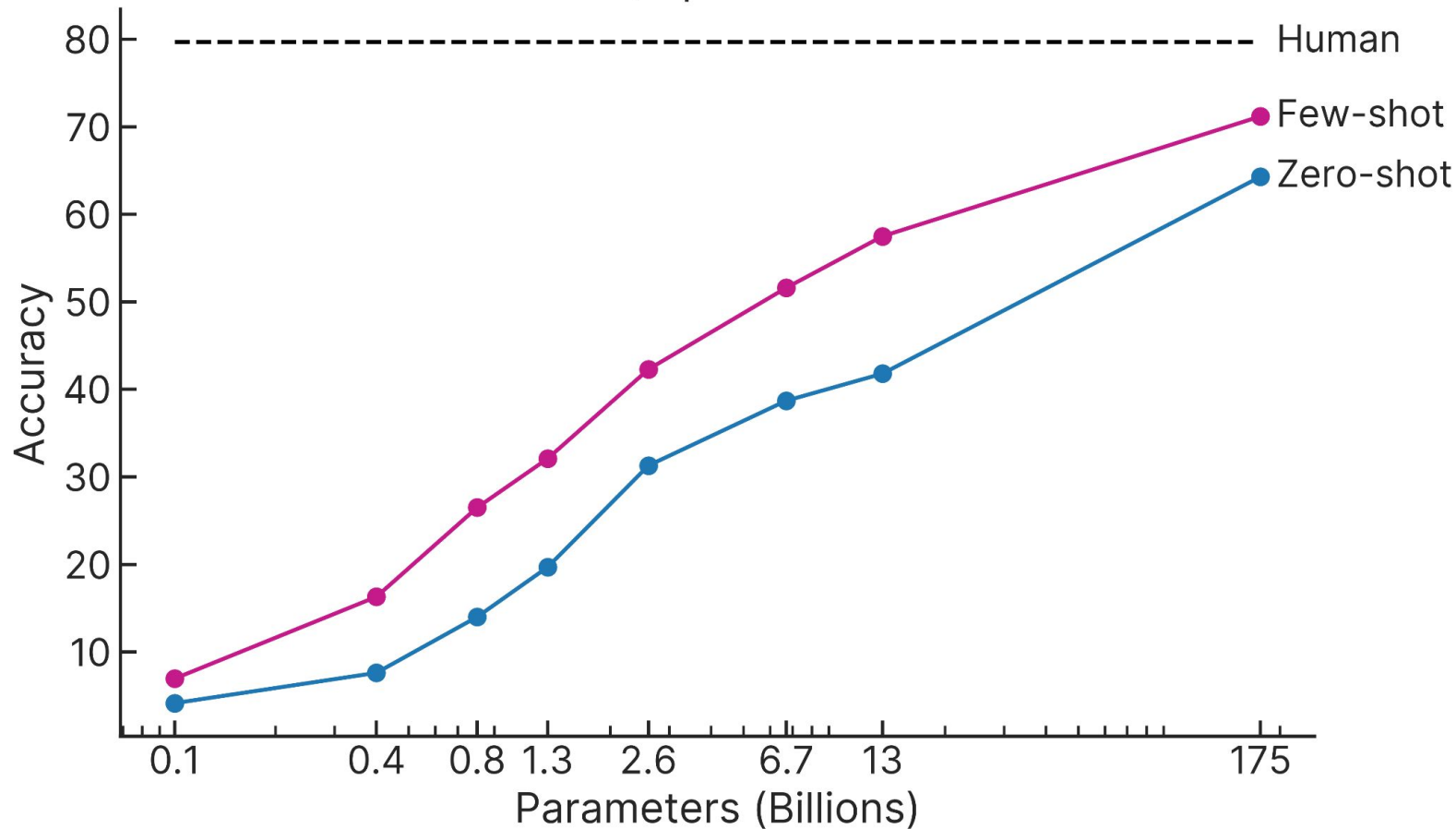
Method	Inference FLOPs	Training FLOPs	Disk space
T-Few	1.1e12	2.7e16	4.2 MB
T0 [1]	1.1e12	0	0 B
T5+LM [14]	4.5e13	0	16 kB
GPT-3 6.7B [4]	5.4e13	0	16 kB
GPT-3 13B [4]	1.0e14	0	16 kB
GPT-3 175B [4]	1.4e15	0	16 kB



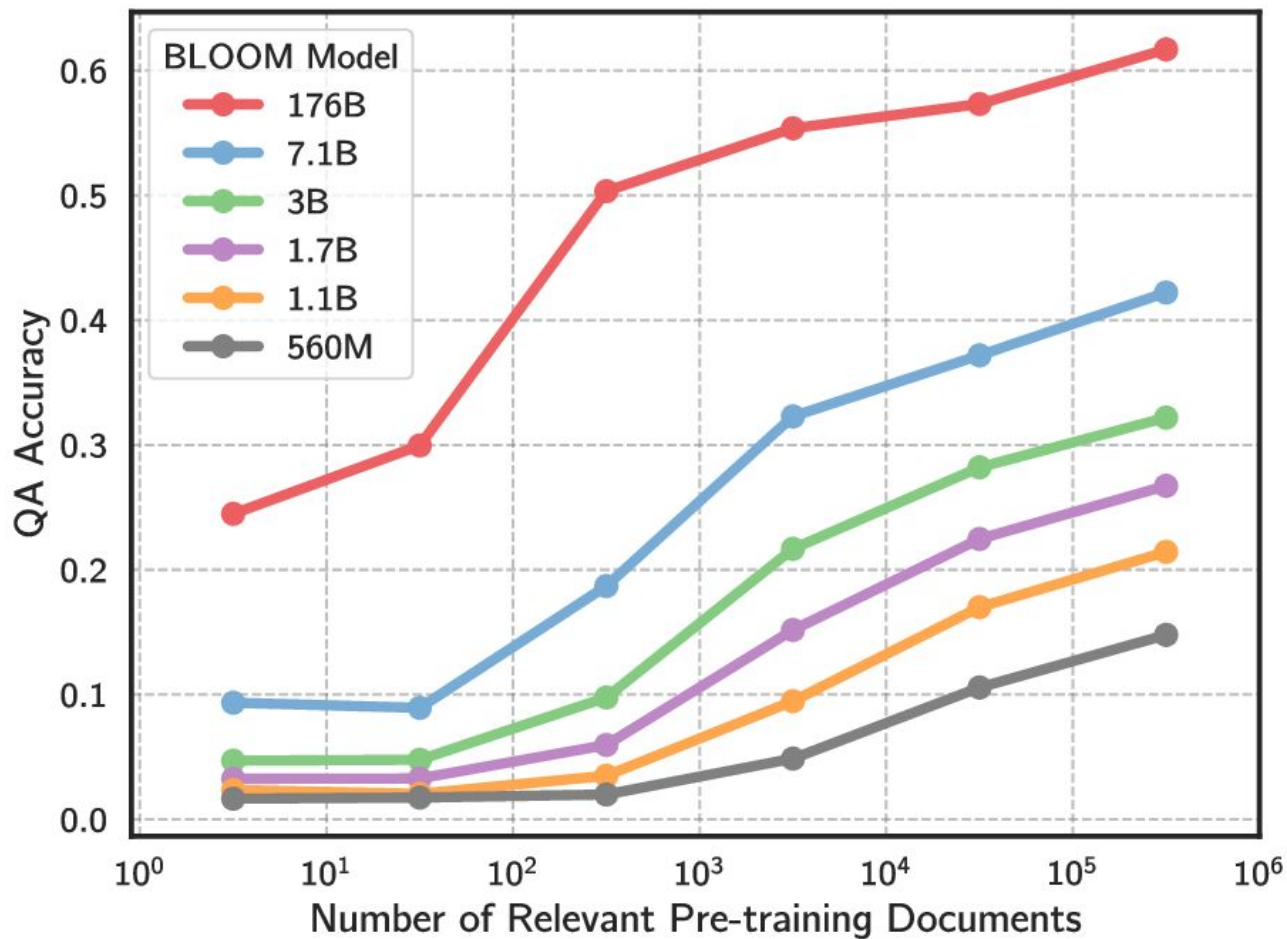
Method	Acc.
T-Few	75.8%
Human baseline [2]	73.5%
PET [50]	69.6%
SetFit [51]	66.9%
GPT-3 [4]	62.7%

Table 2: Top-5 best methods on RAFT as of writing. T-Few is the first method to outperform the human baseline and achieves over 6% higher accuracy than the next-best method.

TriviaQA performance



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



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_Alighieri

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

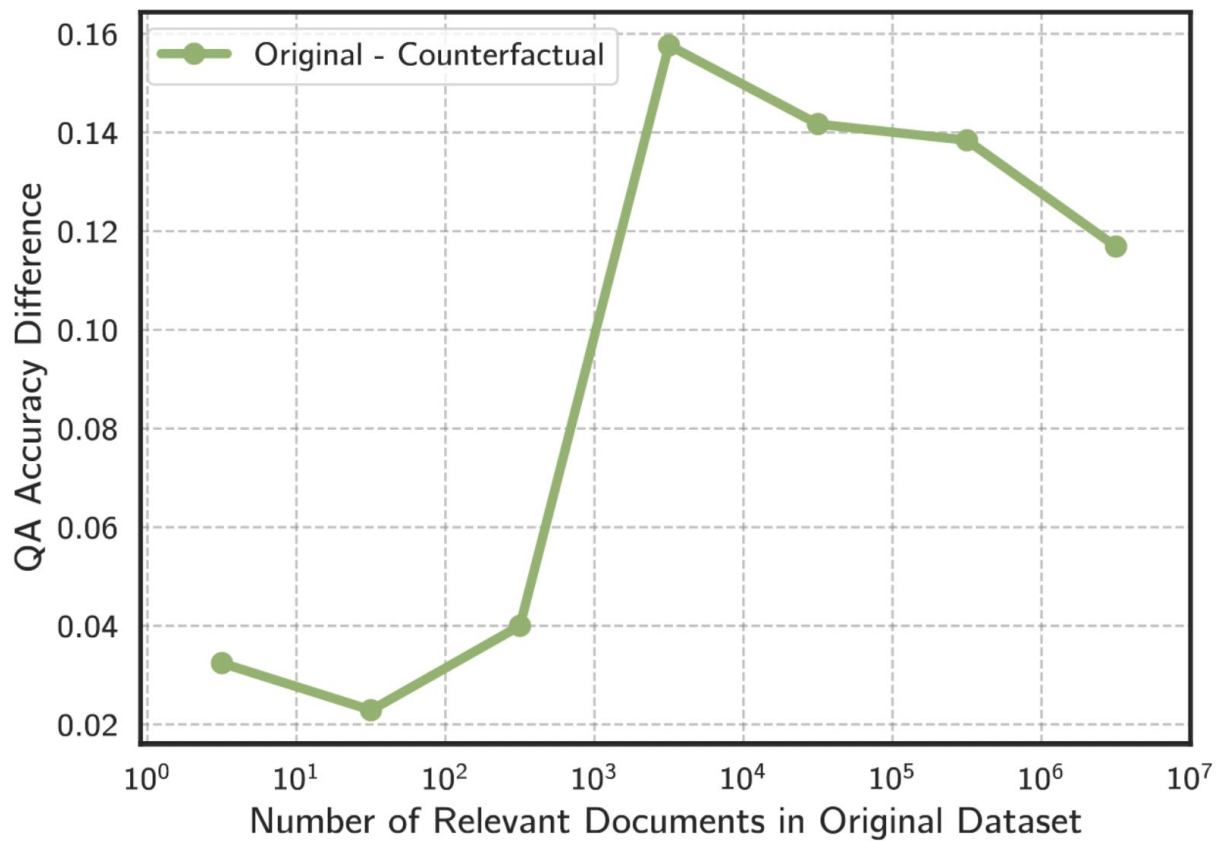
Italy

Salient Question Entity

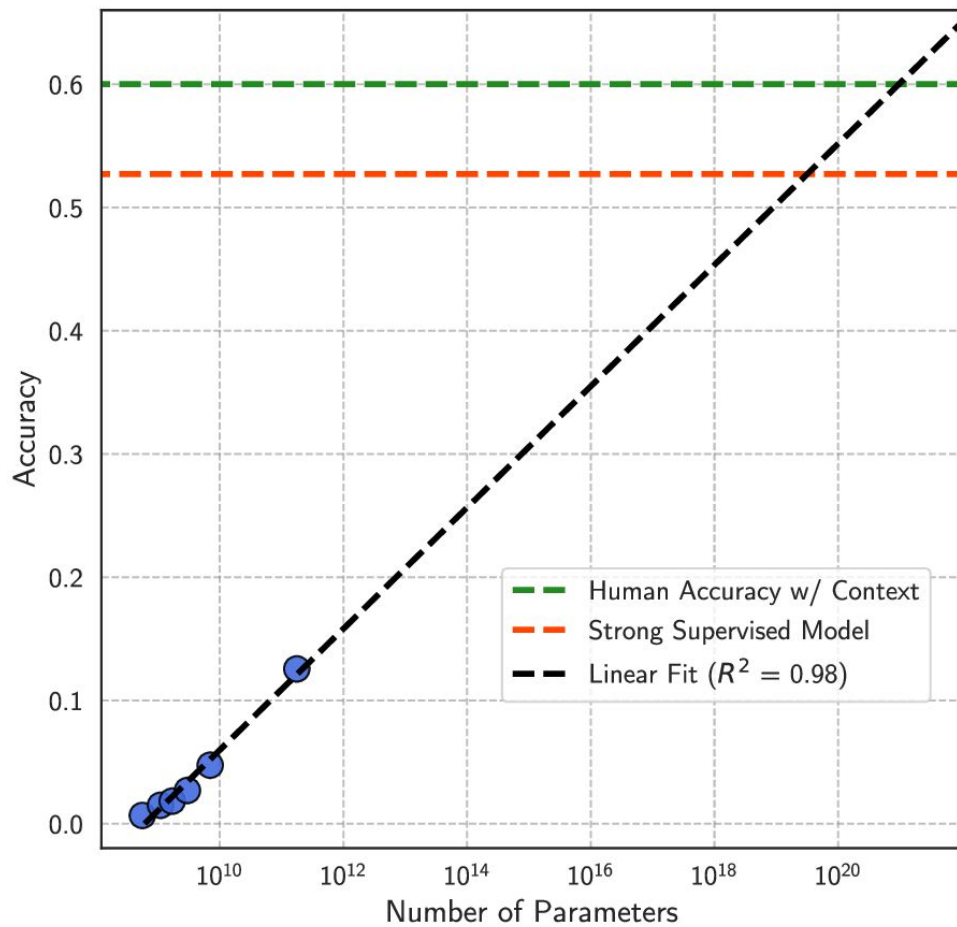
Dante_Alighieri

Salient Answer Entity

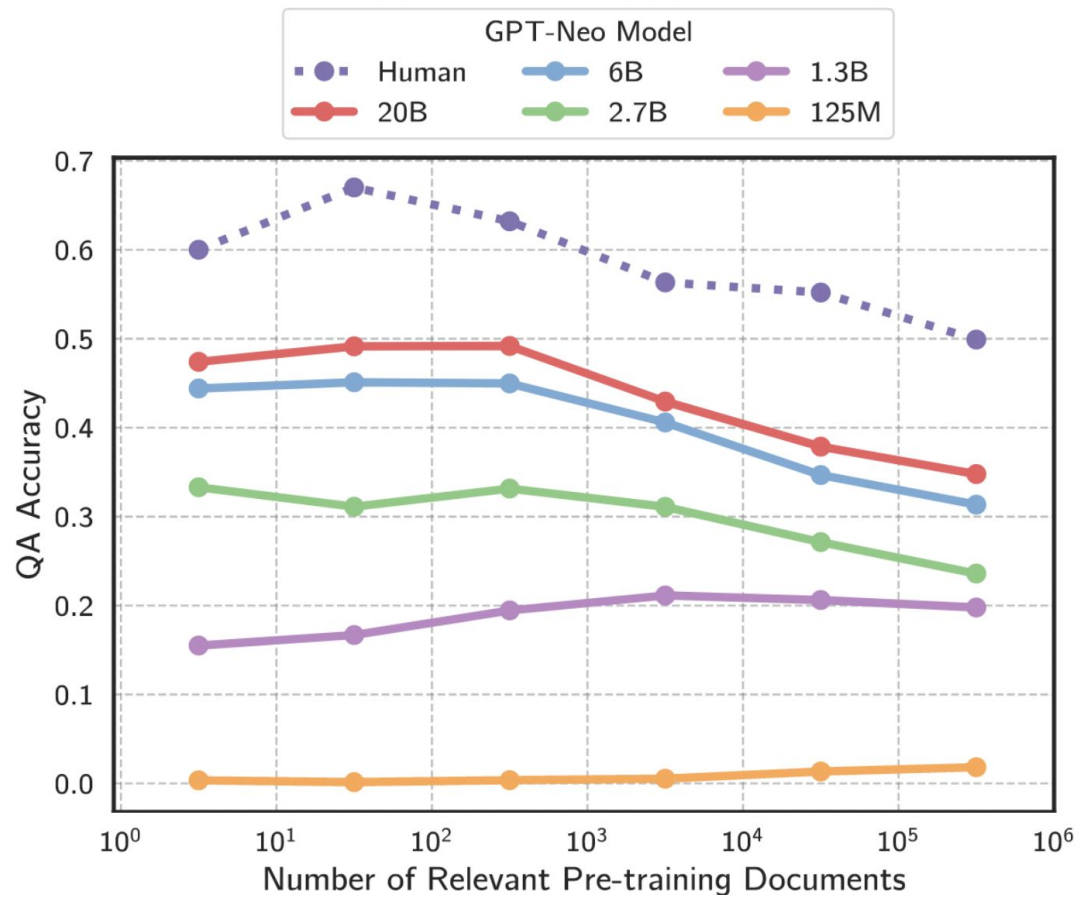
Florence



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

Thanks.

Please give me feedback:

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