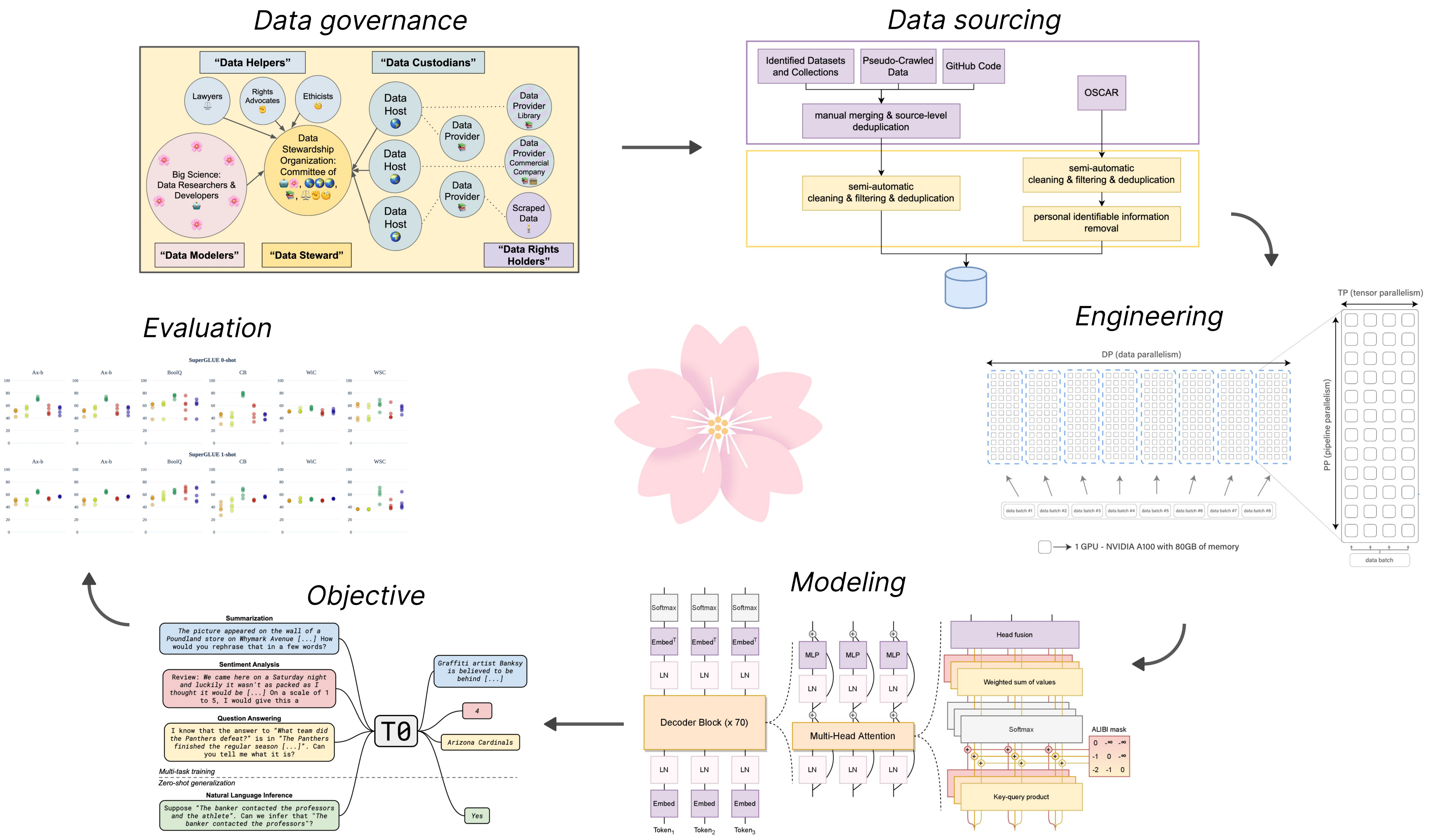
Building Better Language Models: *Insights from BigScience* 🌸

Colin Raffel



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

Review: *We came here on a Saturday night and luckily it wasn't as packed as I thought it would be [...]* On a scale of 1 to 5, I would give this a

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?

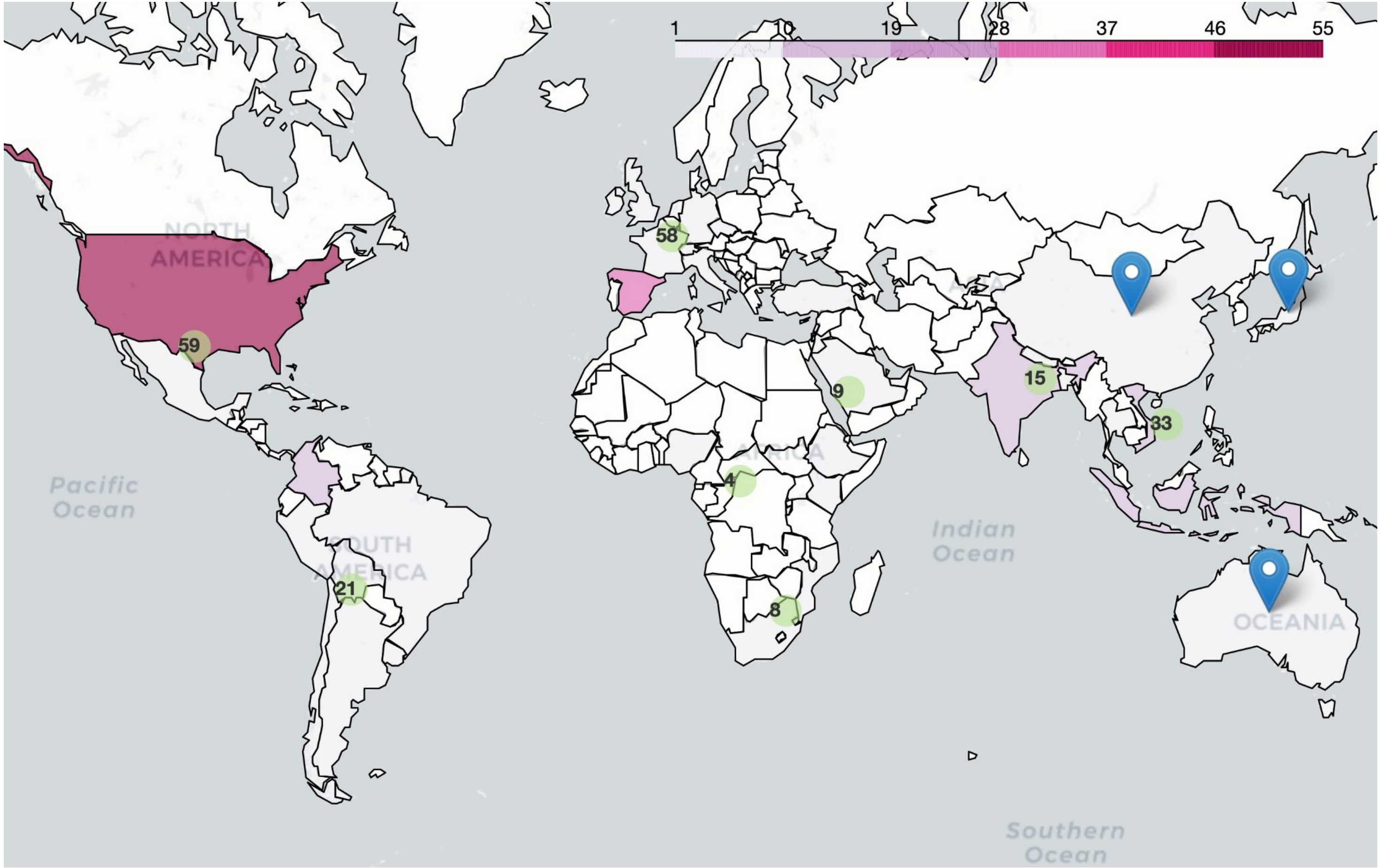
Suppose "*The banker contacted the professors and the athlete*". Can we infer that "*The banker contacted the professors*"?

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

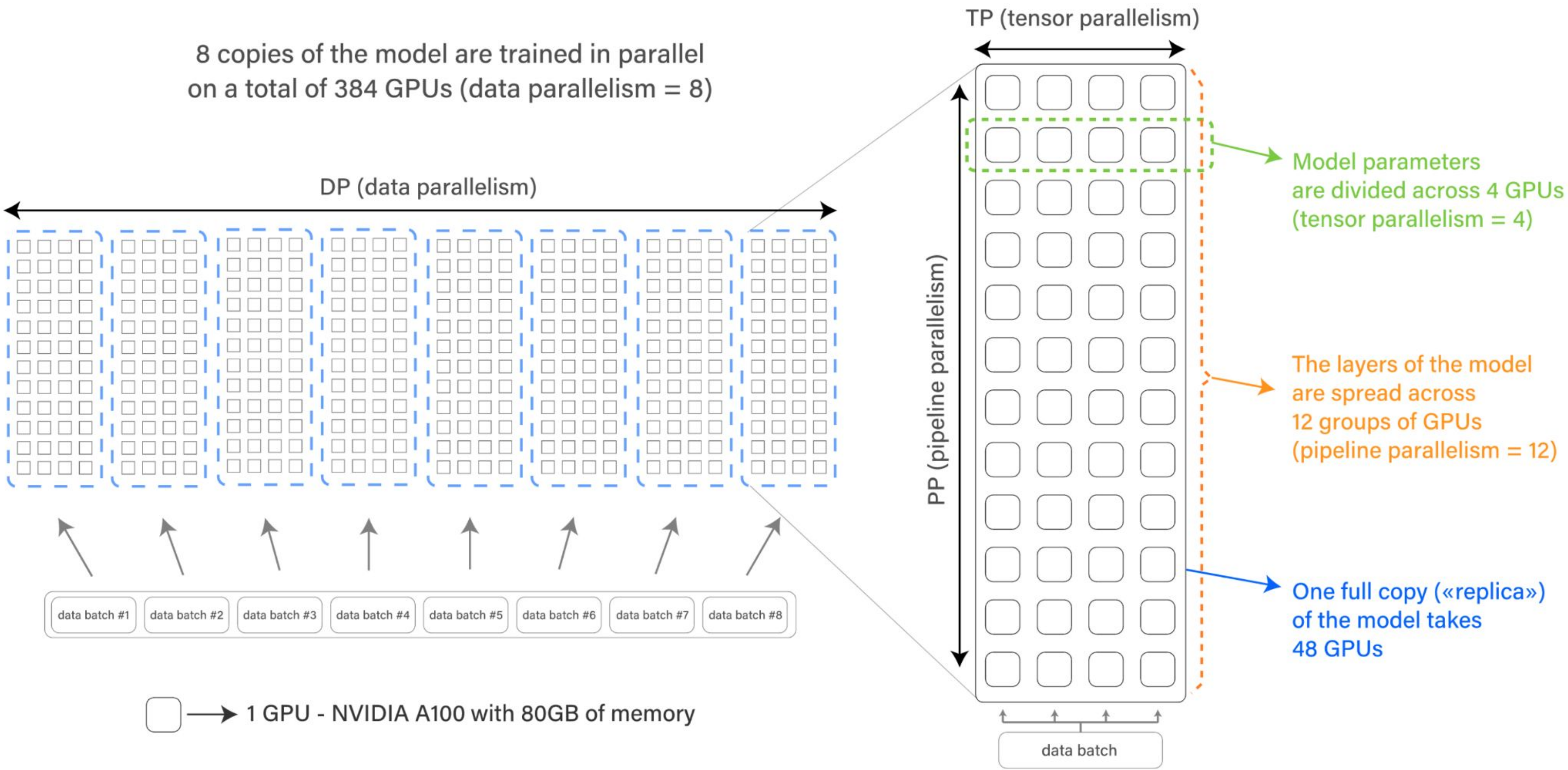
4

Arizona Cardinals

Yes



From https://huggingface.co/spaces/bigscience/SourcingCatalog



8 copies of the model are trained in parallel on a total of 384 GPUs (data parallelism = 8)

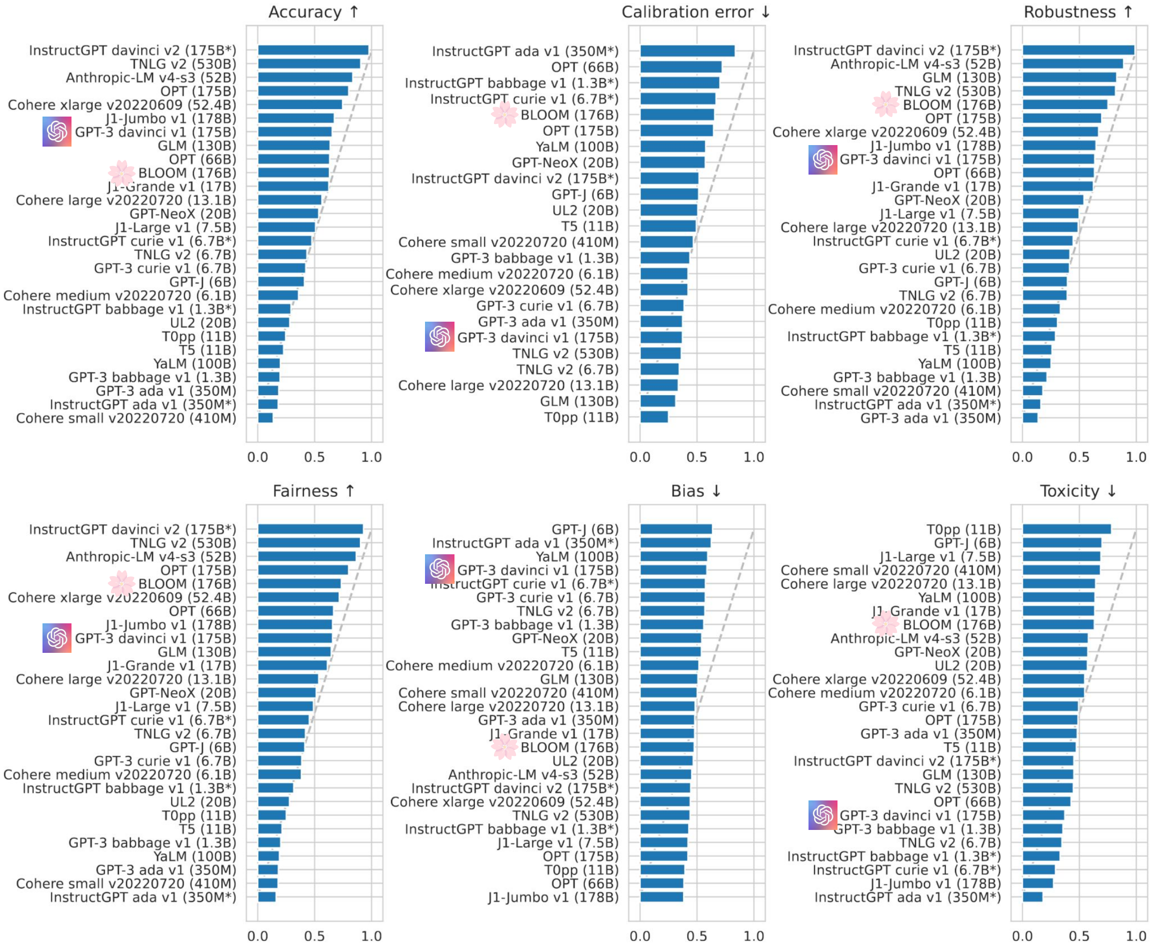
1 GPU - NVIDIA A100 with 80GB of memory

Model parameters are divided across 4 GPUs (tensor parallelism = 4)

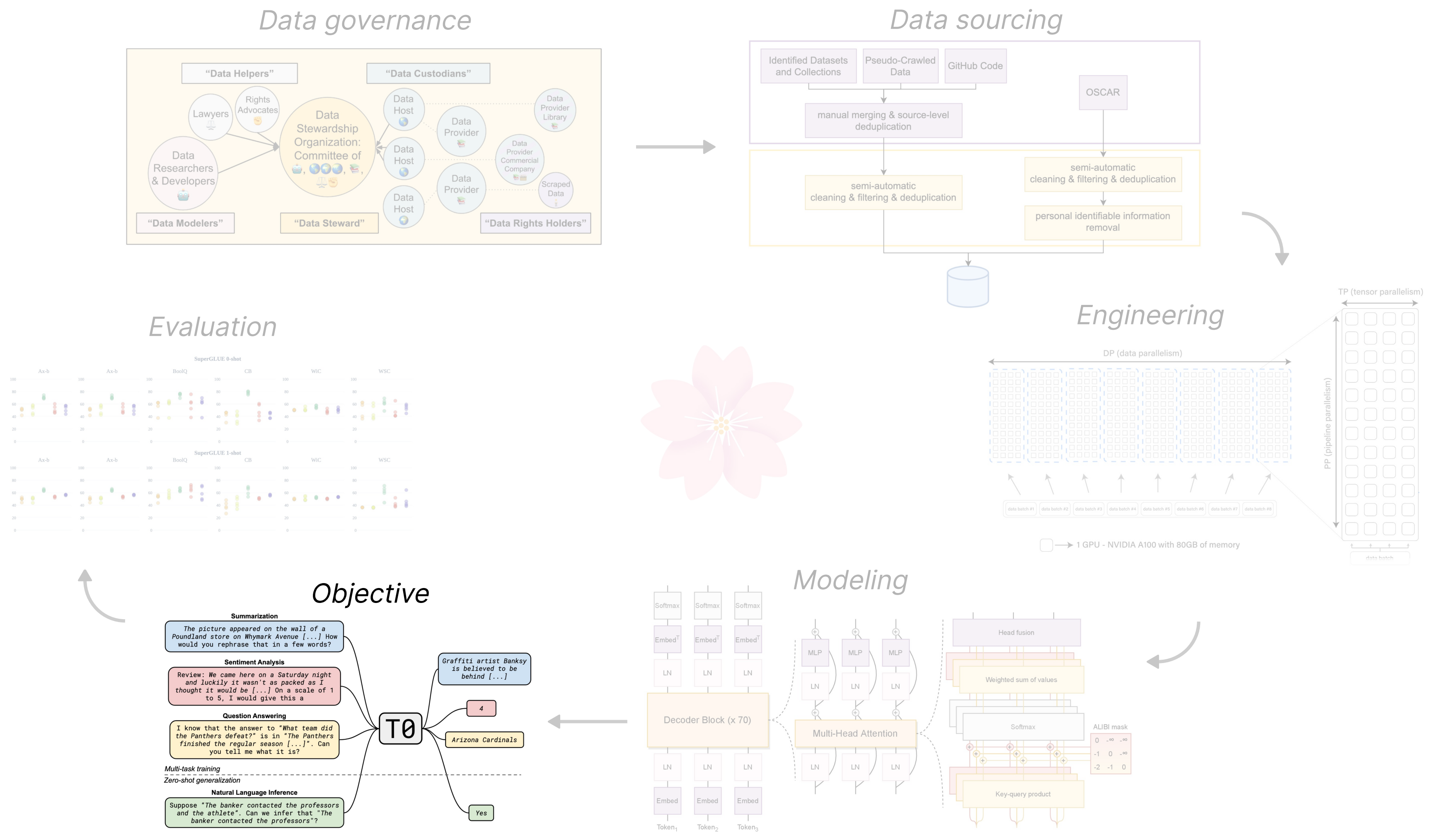
The layers of the model are spread across 12 groups of GPUs (pipeline parallelism = 12)

One full copy («replica») of the model takes 48 GPUs

From "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model" by Le Scao et al.



From "Holistic Evaluation of Language Models" by Liang et al.



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

Review: We came here on a Saturday night and luckily it wasn't as packed as I thought it would be [...] On a scale of 1 to 5, I would give this a

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?

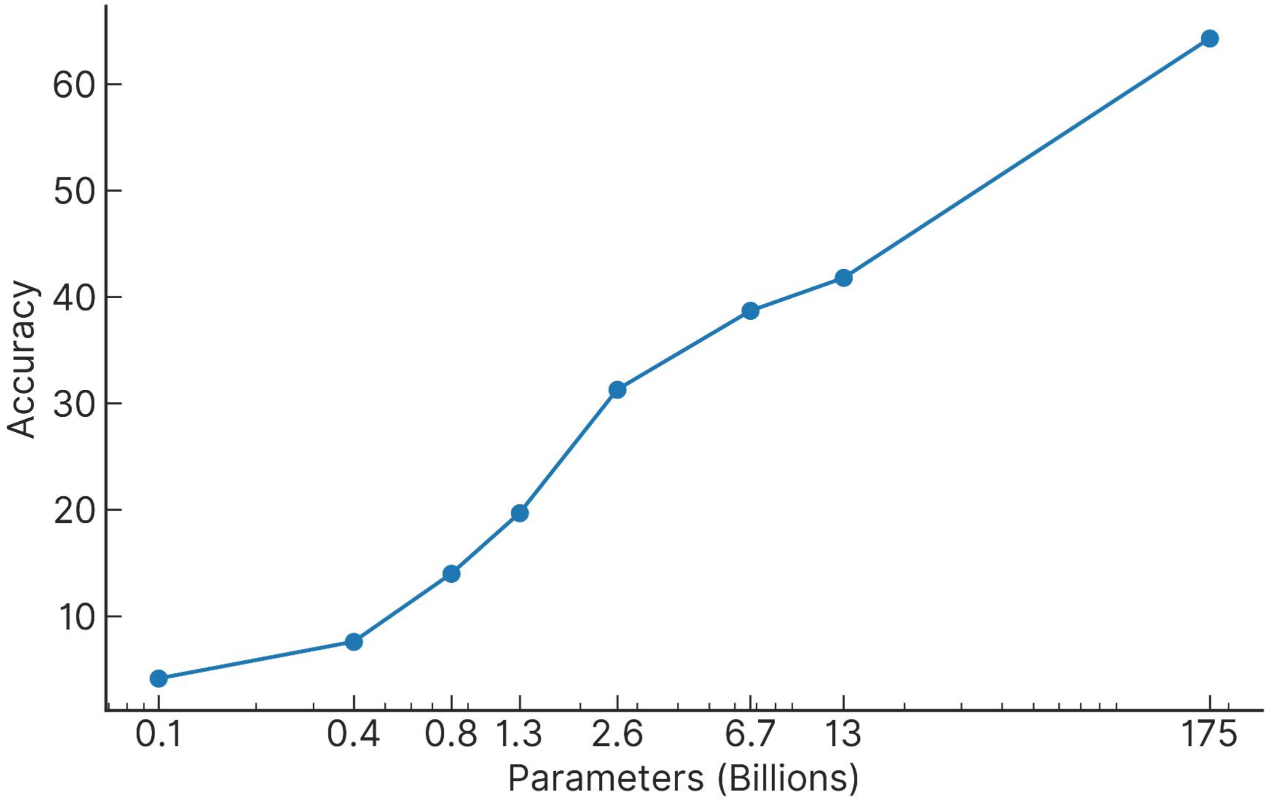
Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

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

4

Arizona Cardinals

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

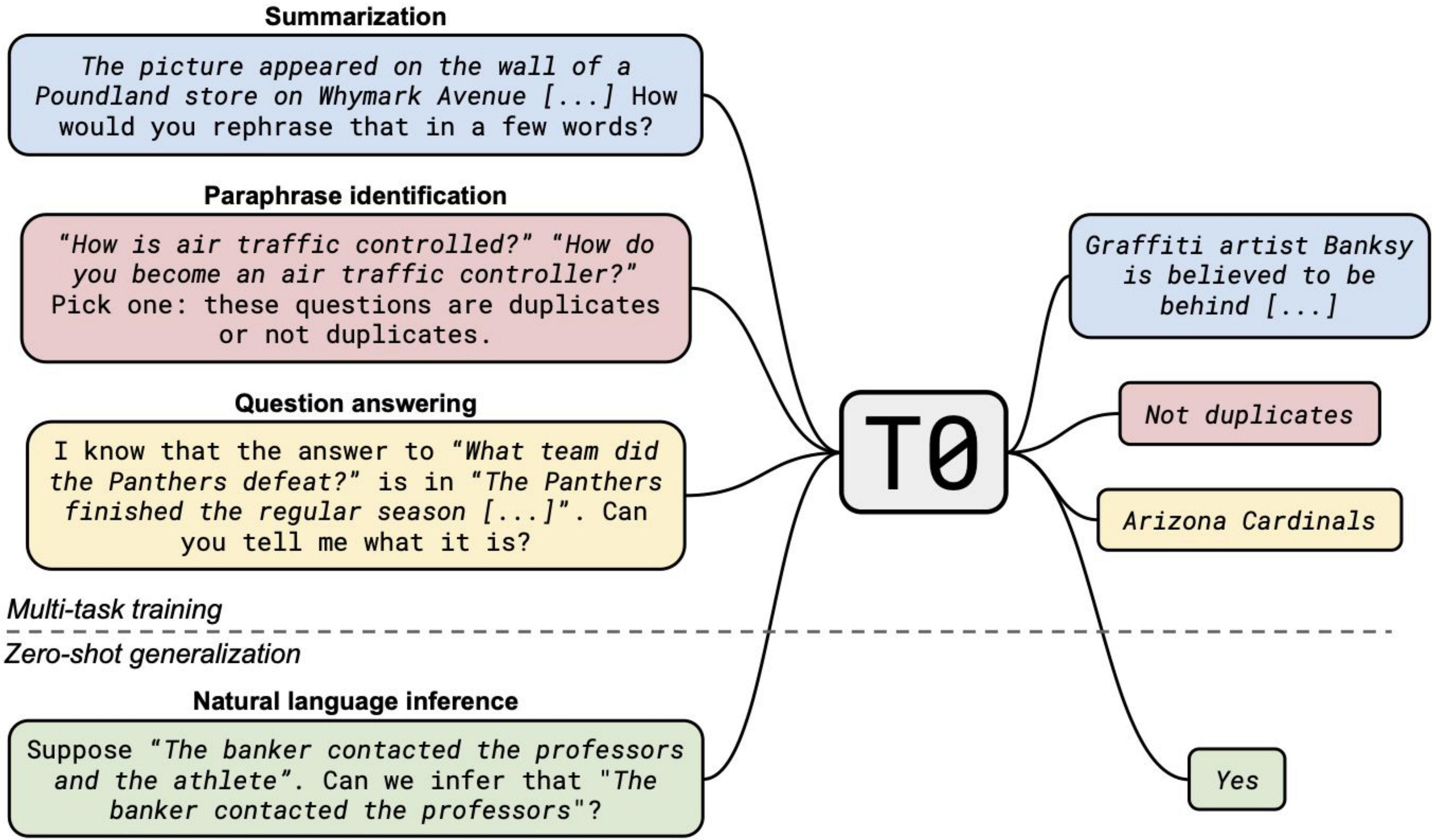
https://blog.nytsoi.net/tag/reddit

... 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

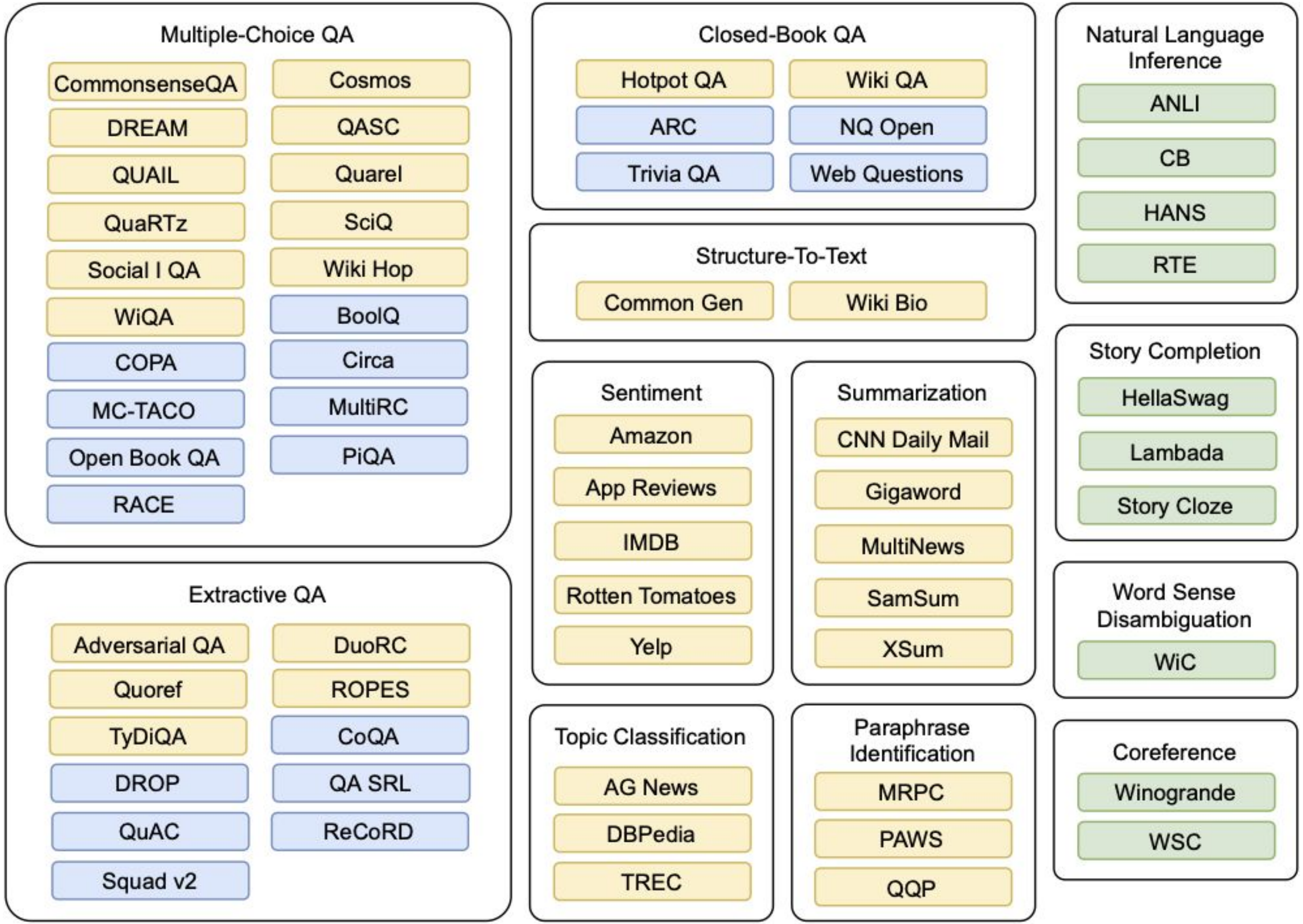
Graffiti artist Banksy is believed to be behind [...]

Not duplicates

Arizona Cardinals

Yes

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



Multiple-Choice QA

- CommonsenseQA
- Cosmos
- DREAM
- QASC
- QUAIL
- Quarel
- QuaRTz
- SciQ
- Social I QA
- Wiki Hop
- WiQA
- BoolQ
- COPA
- Circa
- MC-TACO
- MultiRC
- Open Book QA
- PiQA
- RACE

Extractive QA

- Adversarial QA
- DuoRC
- Quoref
- ROPES
- TyDiQA
- CoQA
- DROP
- QA SRL
- QuAC
- ReCoRD
- Squad v2

Closed-Book QA

- Hotpot QA
- Wiki QA
- ARC
- NQ Open
- Trivia QA
- 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

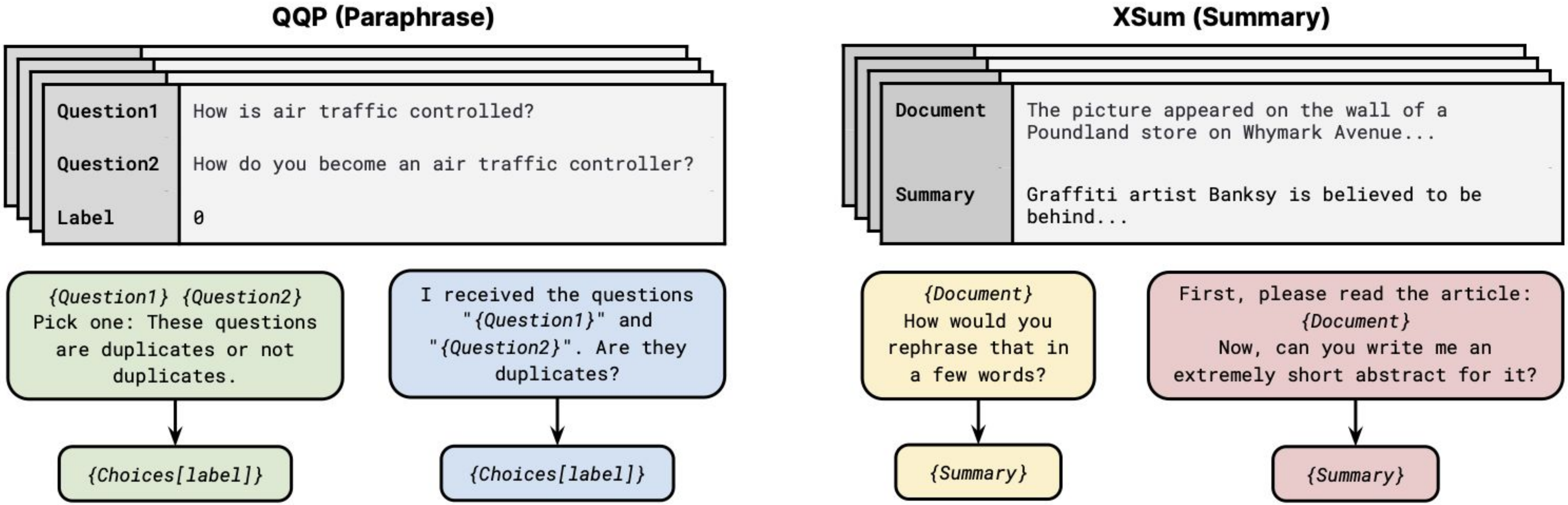
Word Sense Disambiguation

- 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...

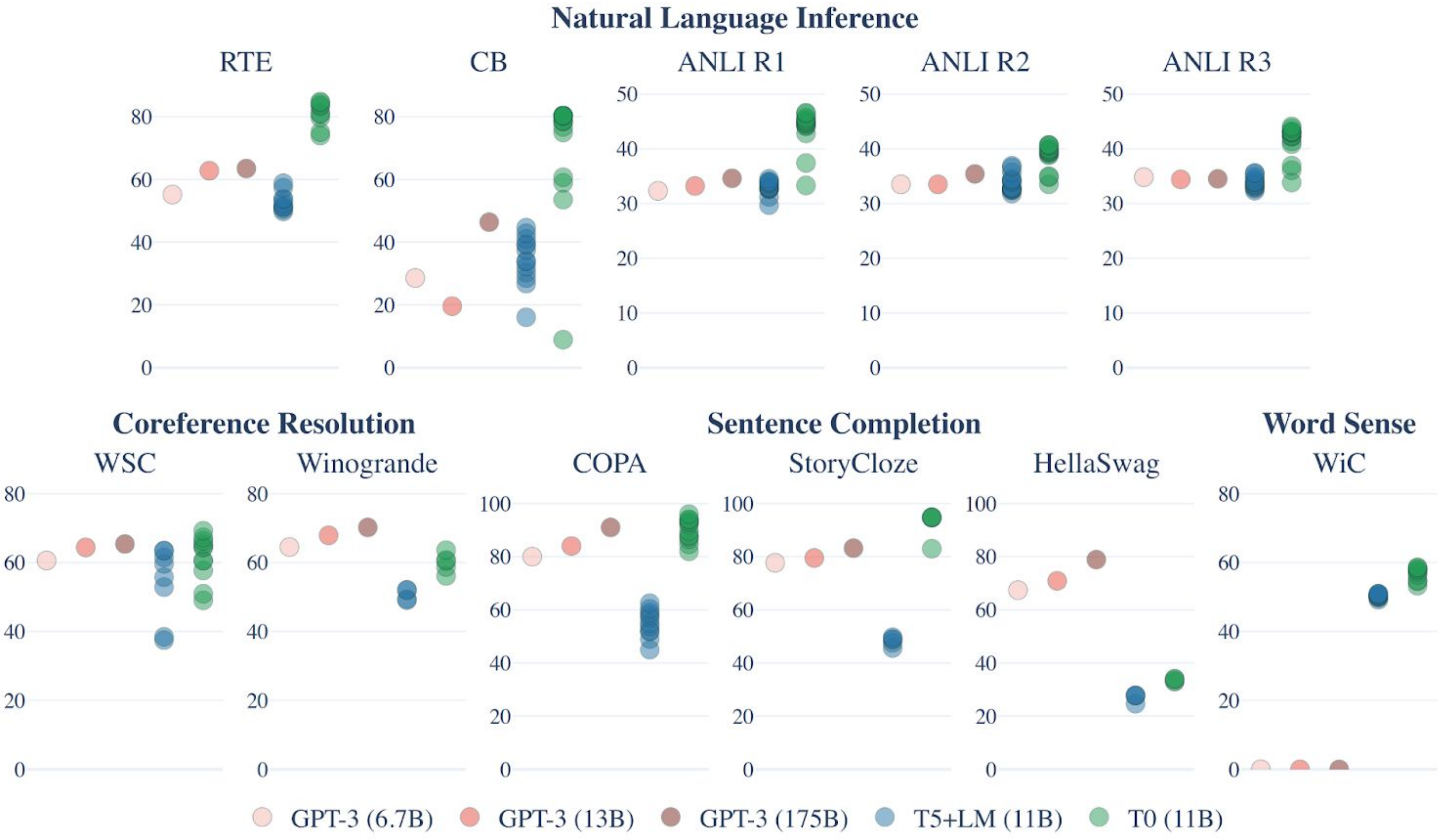
{Document}
How would you rephrase that in a few words?

{Summary}

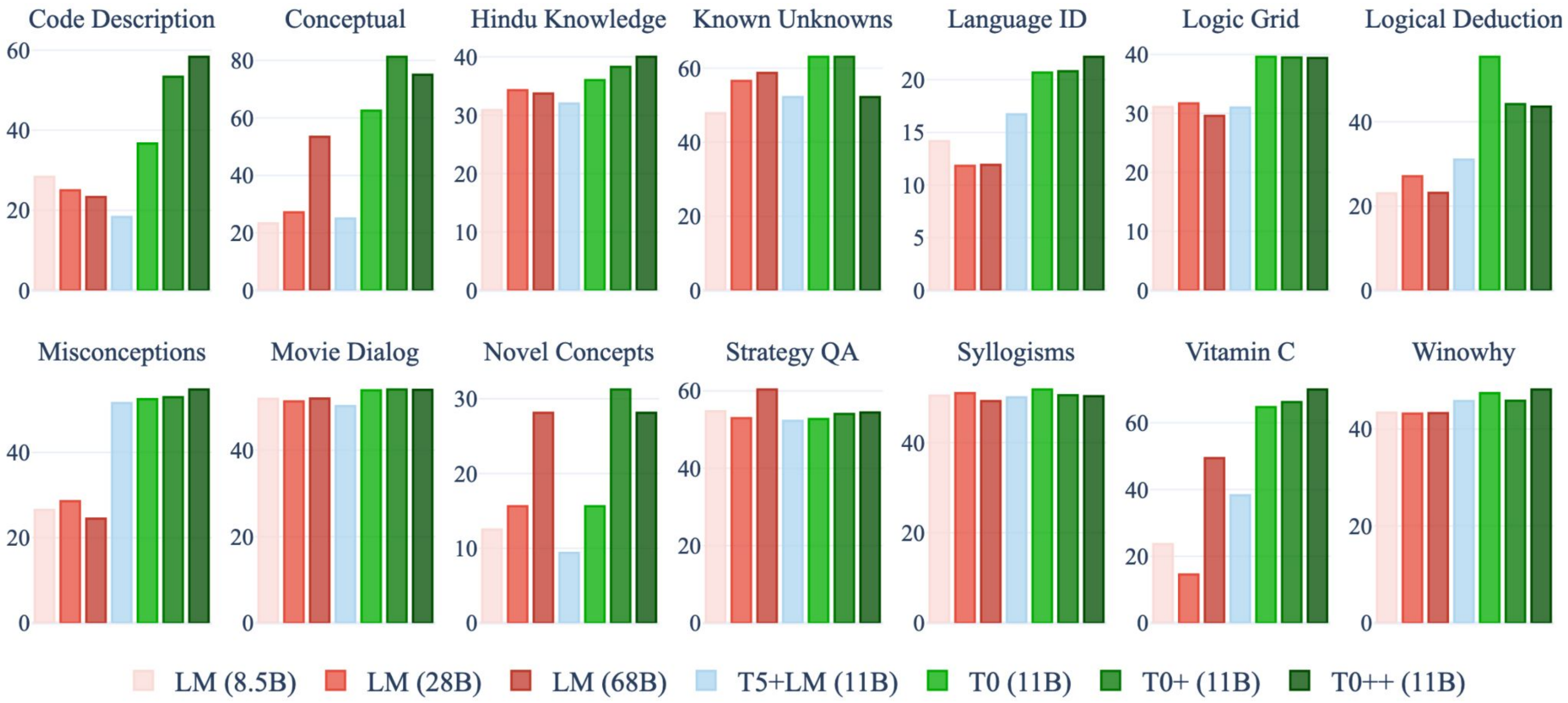
First, please read the article:
{Document}
Now, can you write me an extremely short abstract for it?

{Summary}

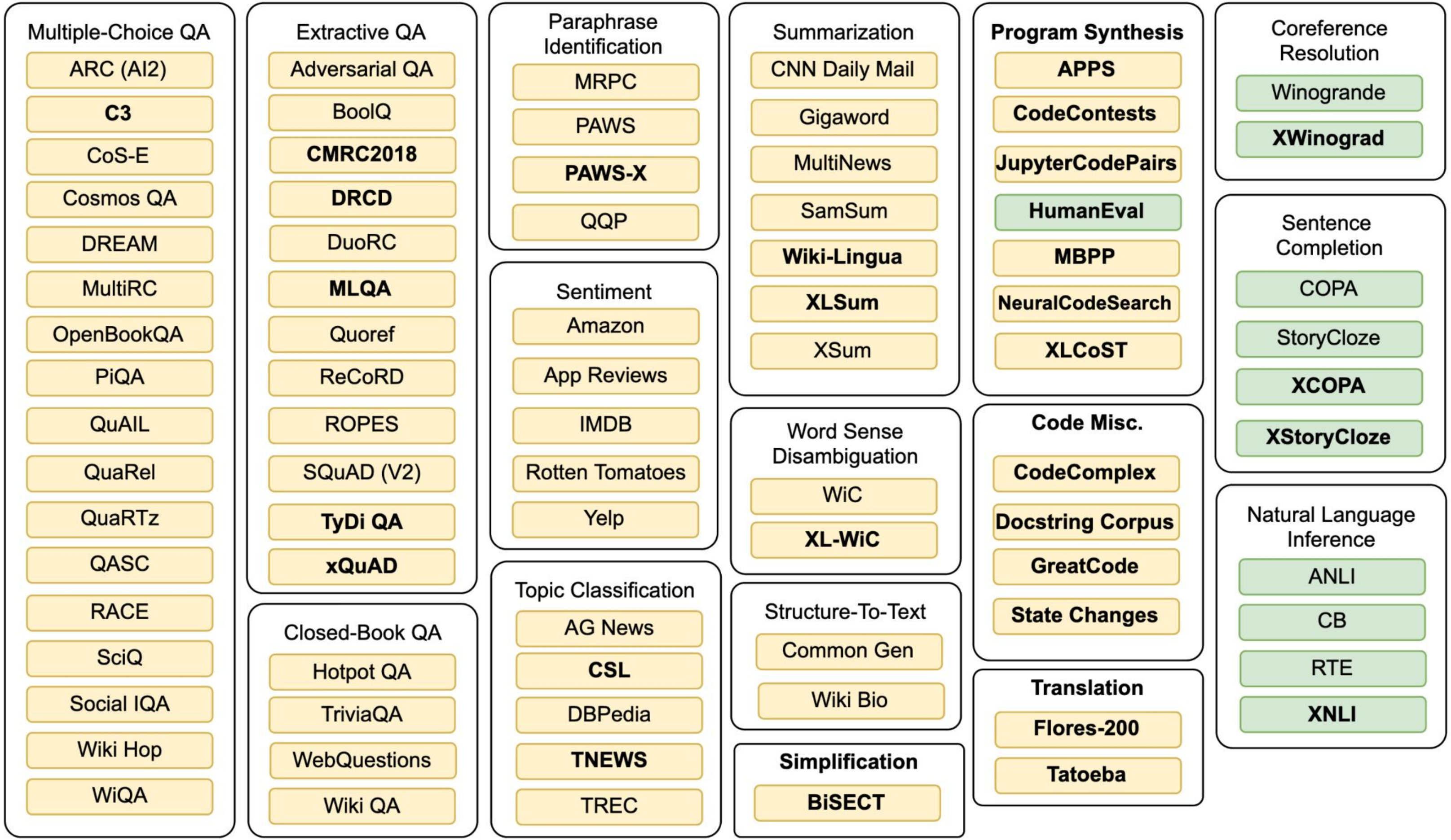
From "Multitask Prompted Training Enables Zero-Shot Task Generalization" by Sanh et al.



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



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



Multiple-Choice QA
- ARC (AI2)
- **C3**
- CoS-E
- Cosmos QA
- DREAM
- MultiRC
- OpenBookQA
- PiQA
- QuAIL
- QuaRel
- QuaRTz
- QASC
- RACE
- SciQ
- Social IQA
- Wiki Hop
- WiQA

Extractive QA
- Adversarial QA
- BoolQ
- **CMRC2018**
- **DRCD**
- DuoRC
- **MLQA**
- Quoref
- ReCoRD
- ROPES
- SQuAD (V2)
- **TyDi QA**
- **xQuAD**

Closed-Book QA
- Hotpot QA
- TriviaQA
- WebQuestions
- Wiki QA

Paraphrase Identification
- MRPC
- PAWS
- **PAWS-X**
- QQP

Sentiment
- Amazon
- App Reviews
- IMDB
- Rotten Tomatoes
- Yelp

Topic Classification
- AG News
- **CSL**
- DBPedia
- **TNEWS**
- TREC

Summarization
- CNN Daily Mail
- Gigaword
- MultiNews
- SamSum
- **Wiki-Lingua**
- **XLSum**
- XSum

Word Sense Disambiguation
- WiC
- **XL-WiC**

Structure-To-Text
- Common Gen
- Wiki Bio

Simplification
- **BiSECT**

Program Synthesis
- **APPS**
- **CodeContests**
- **JupyterCodePairs**
- **HumanEval**
- **MBPP**
- **NeuralCodeSearch**
- **XLCoST**

Code Misc.
- **CodeComplex**
- **Docstring Corpus**
- **GreatCode**
- **State Changes**

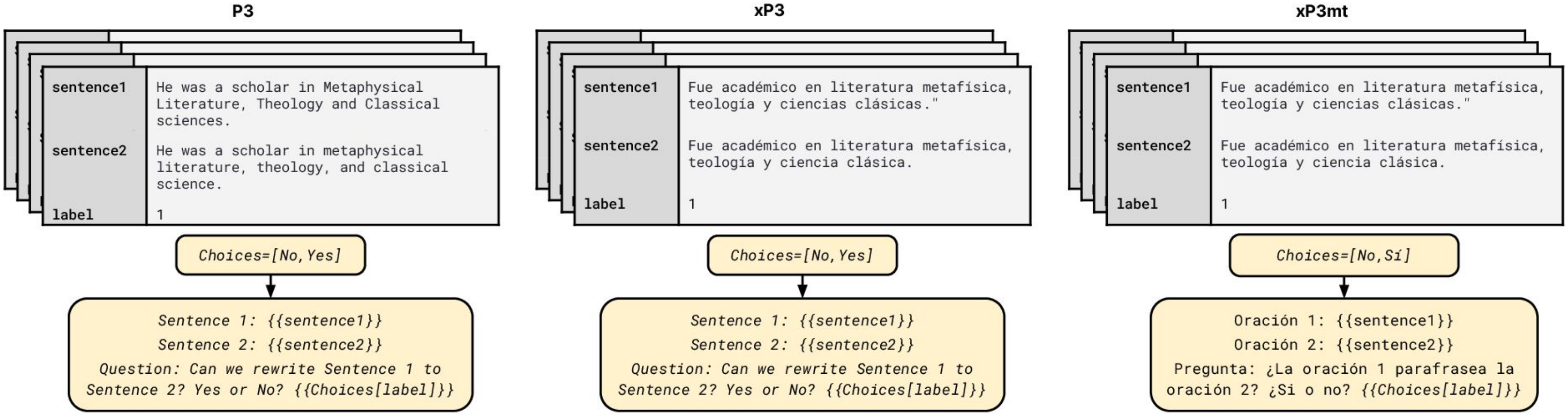
Translation
- **Flores-200**
- **Tatoeba**

Coreference Resolution
- Winogrande
- **XWinograd**

Sentence Completion
- COPA
- StoryCloze
- **XCOPA**
- **XStoryCloze**

Natural Language Inference
- ANLI
- CB
- RTE
- **XNLI**

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



P3

sentence1	He was a scholar in Metaphysical Literature, Theology and Classical sciences.
sentence2	He was a scholar in metaphysical literature, theology, and classical science.
label	1

Choices=[No,Yes]

Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Can we rewrite Sentence 1 to Sentence 2? Yes or No? {{Choices[label]}}

xP3

sentence1	Fue académico en literatura metafísica, teología y ciencias clásicas."
sentence2	Fue académico en literatura metafísica, teología y ciencia clásica.
label	1

Choices=[No,Yes]

Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Can we rewrite Sentence 1 to Sentence 2? Yes or No? {{Choices[label]}}

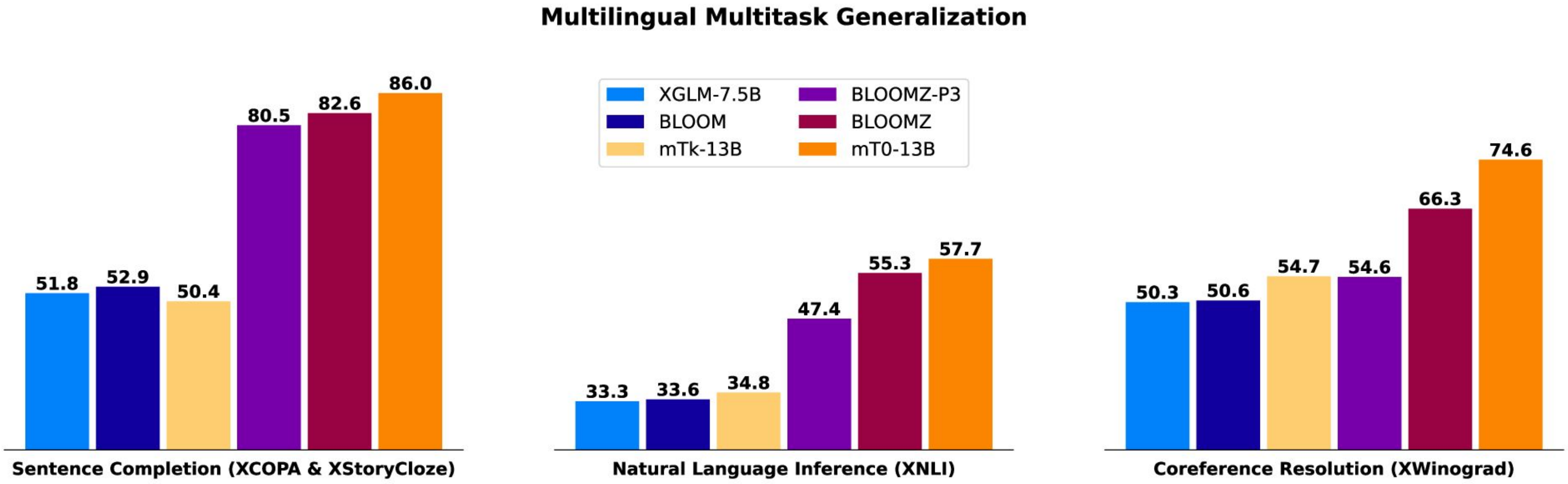
xP3mt

sentence1	Fue académico en literatura metafísica, teología y ciencias clásicas."
sentence2	Fue académico en literatura metafísica, teología y ciencia clásica.
label	1

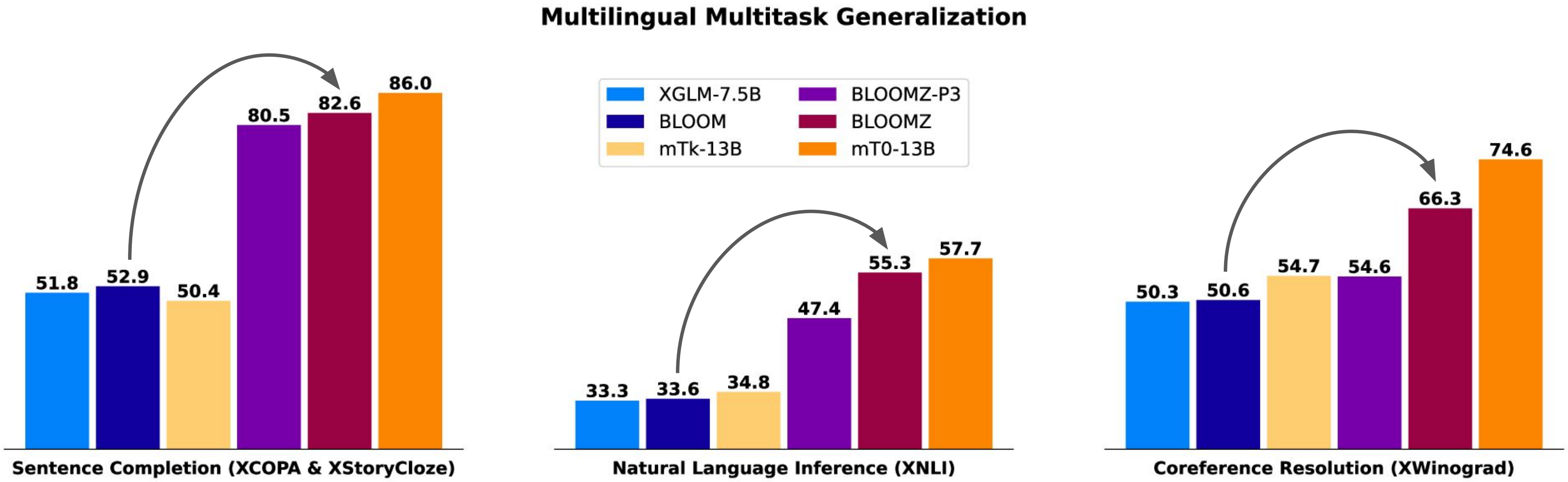
Choices=[No,Sí]

Oración 1: {{sentence1}}
Oración 2: {{sentence2}}
Pregunta: ¿La oración 1 parafrasea la oración 2? ¿Si o no? *{{Choices[label]}}*

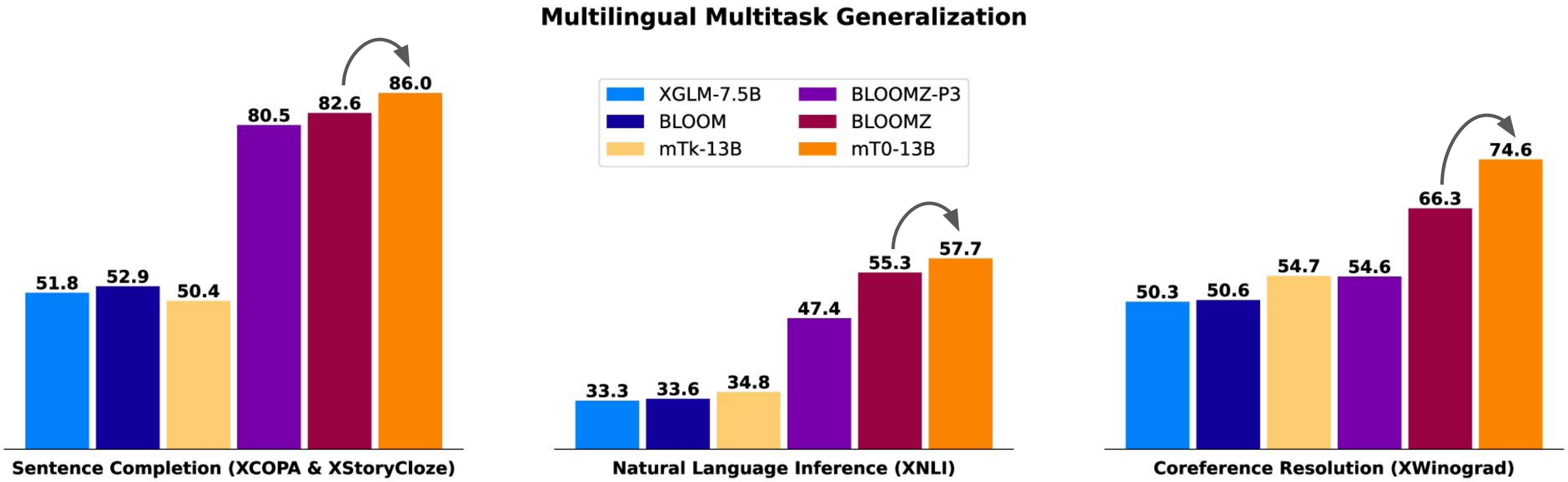
From "Crosslingual Generalization through Multitask Finetuning" by Muennighoff et al.



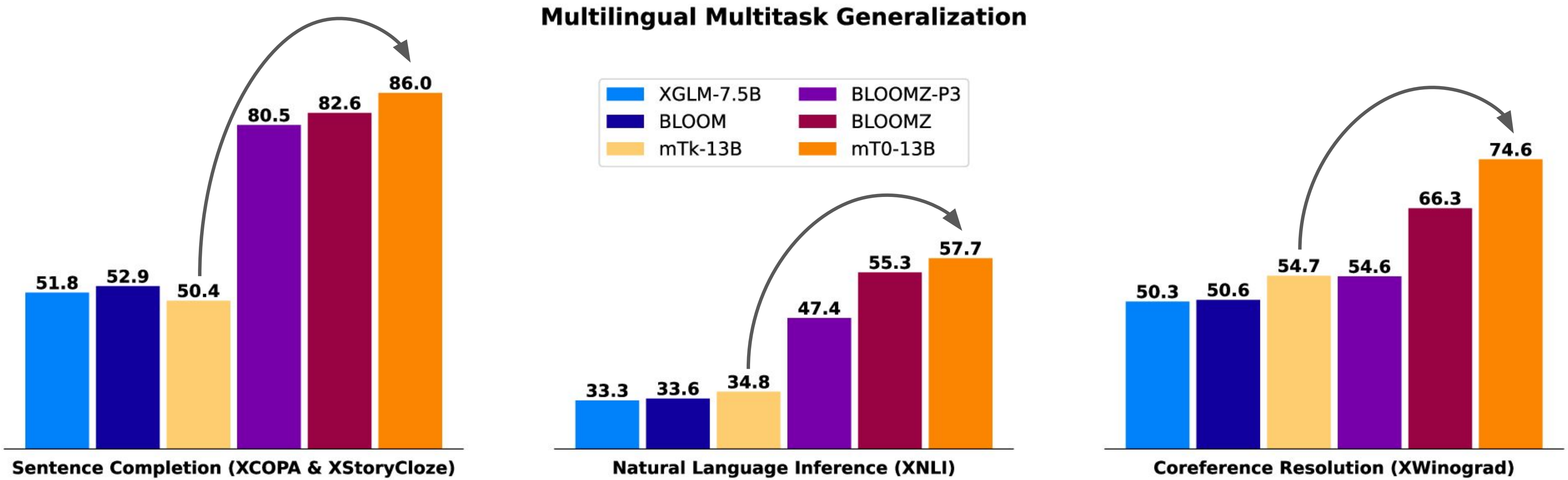
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.



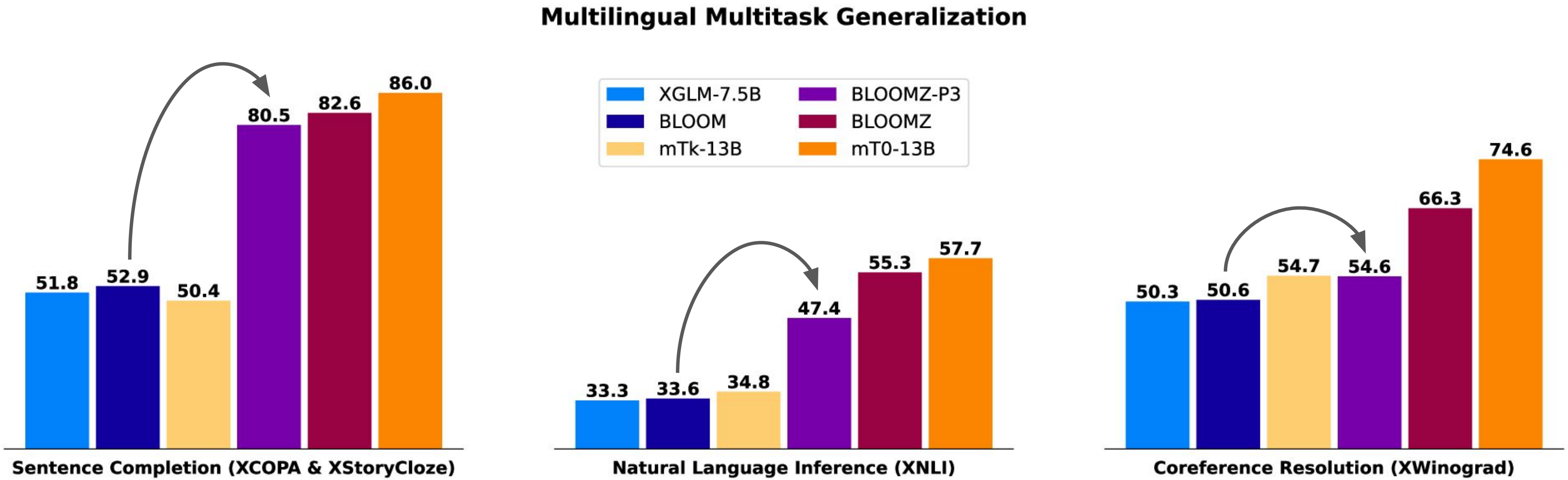
From "Crosslingual Generalization through Multitask Finetuning" by Muennighoff et al.



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

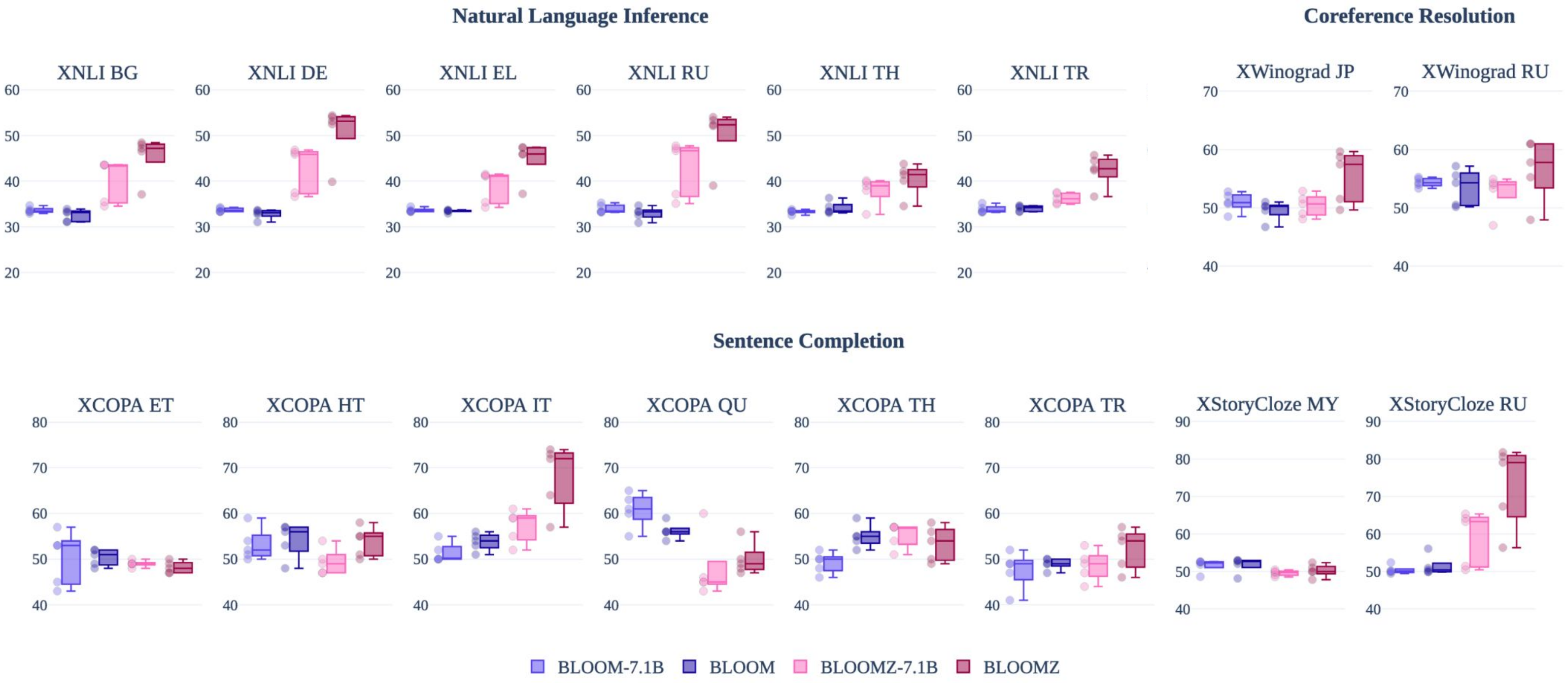


From "Crosslingual Generalization through Multitask Finetuning" by Muennighoff et al.

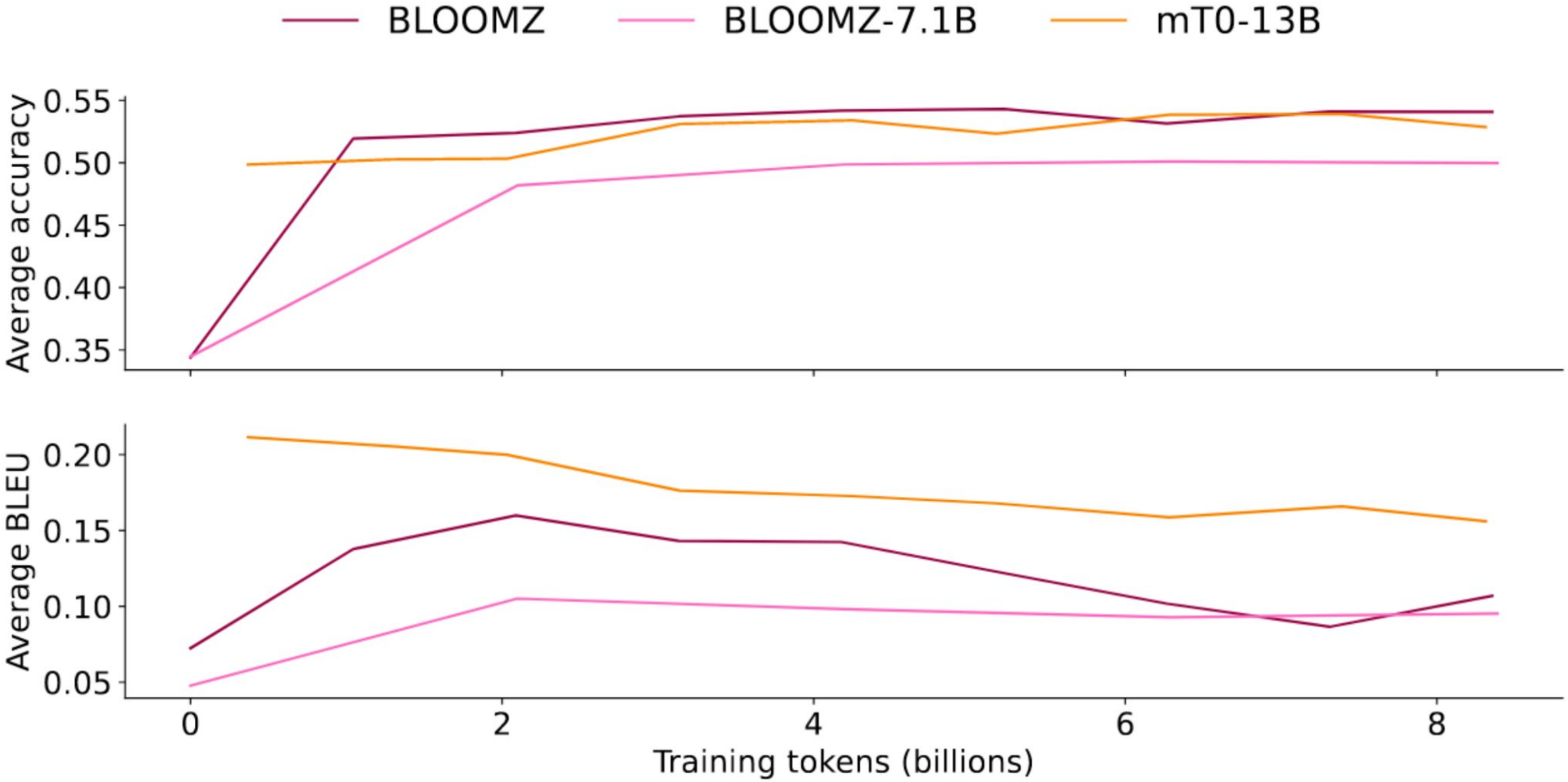


From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.

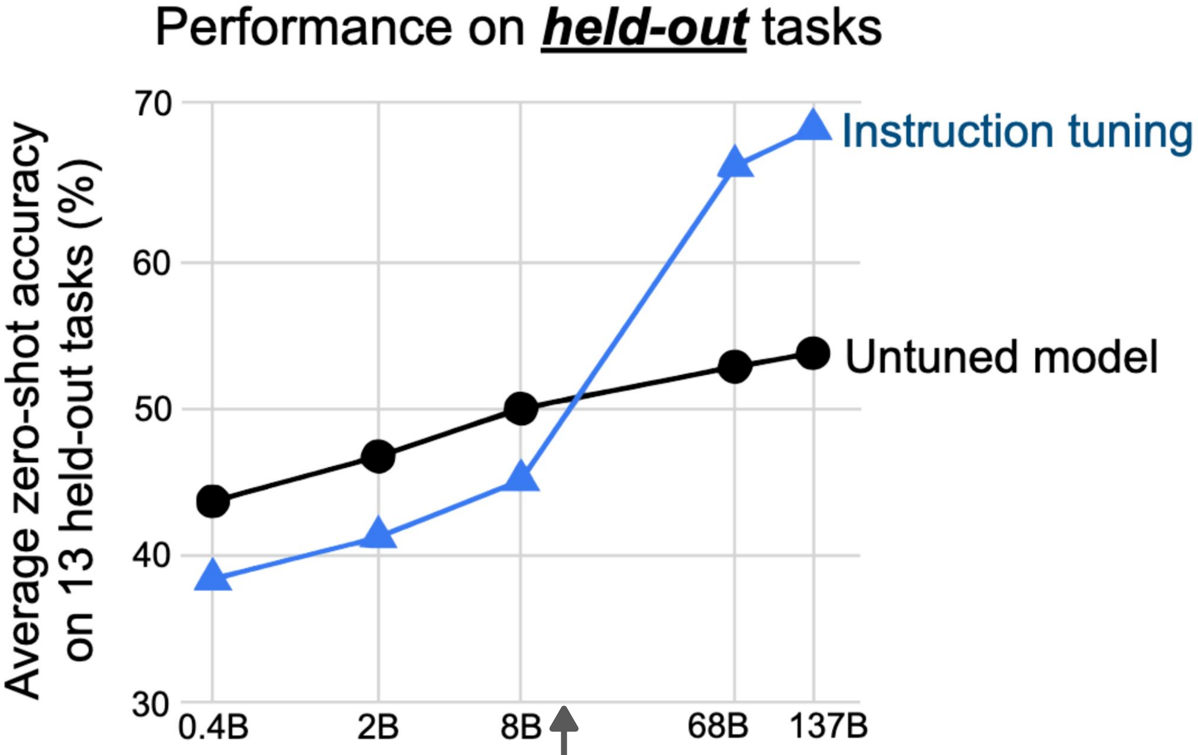
Performance on languages that were never intentionally trained on



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

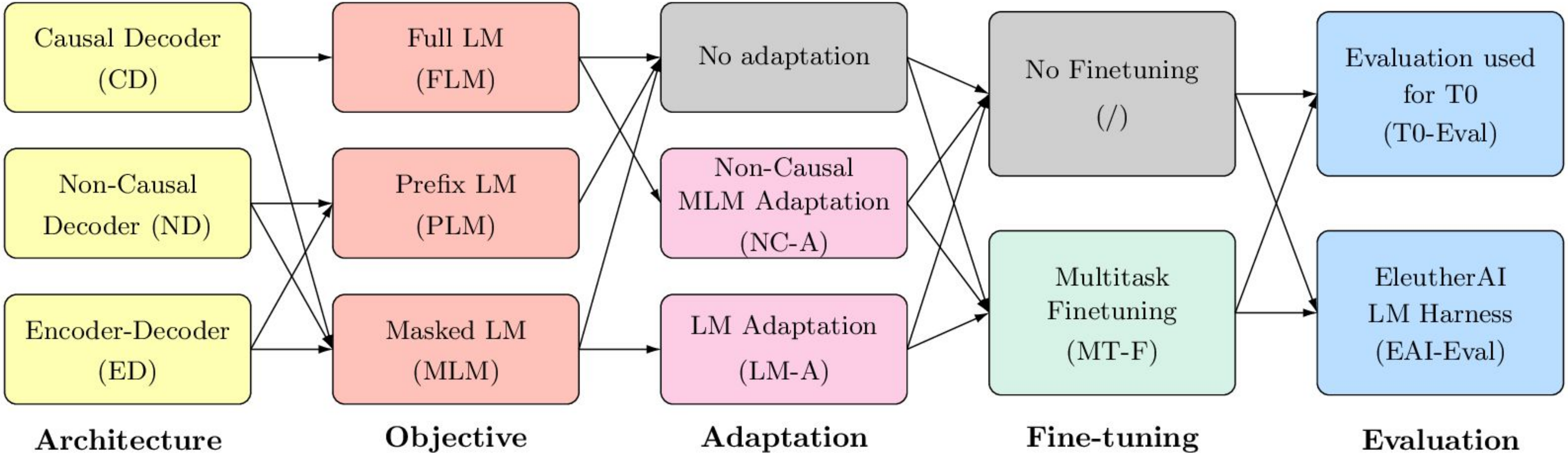


From "Crosslingual Generalization through Multitask Finetuning" by Muennighoff et al.

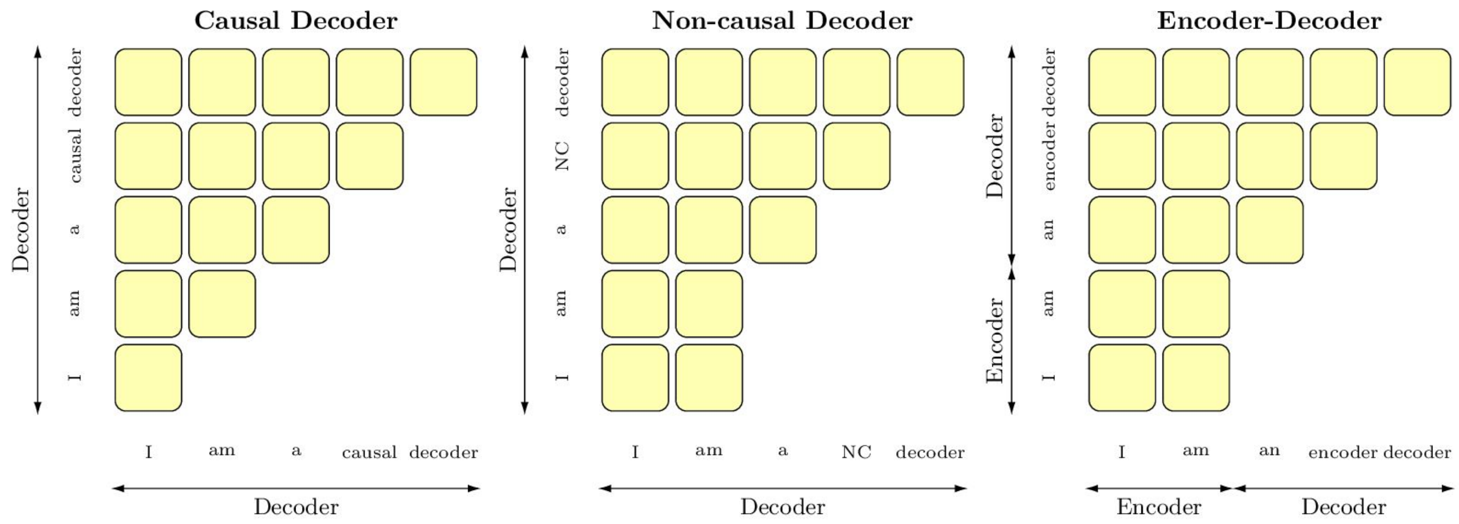


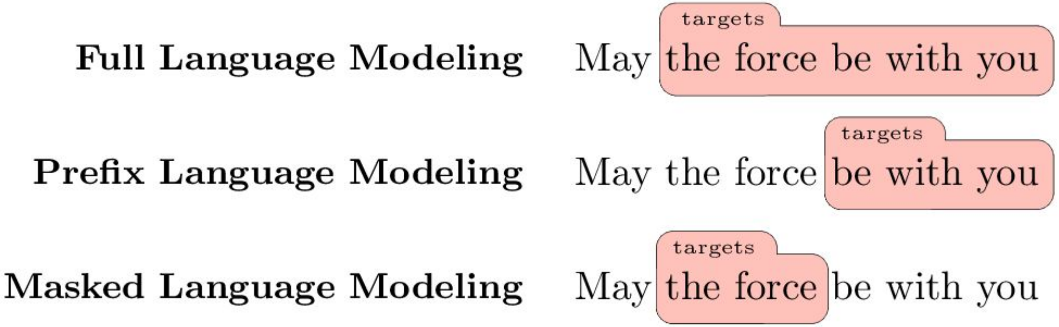
T0 = 11B parameters

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.



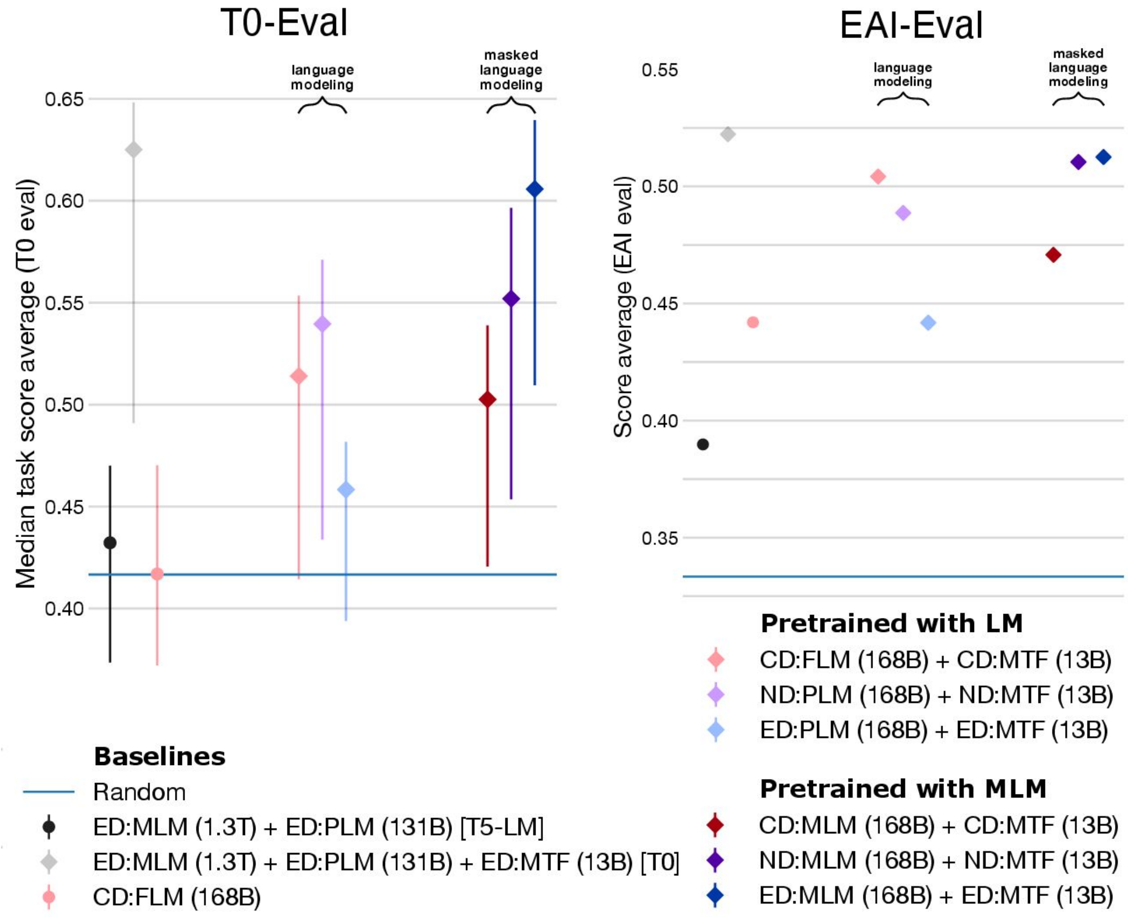


Full Language Modeling May the force be with you
targets

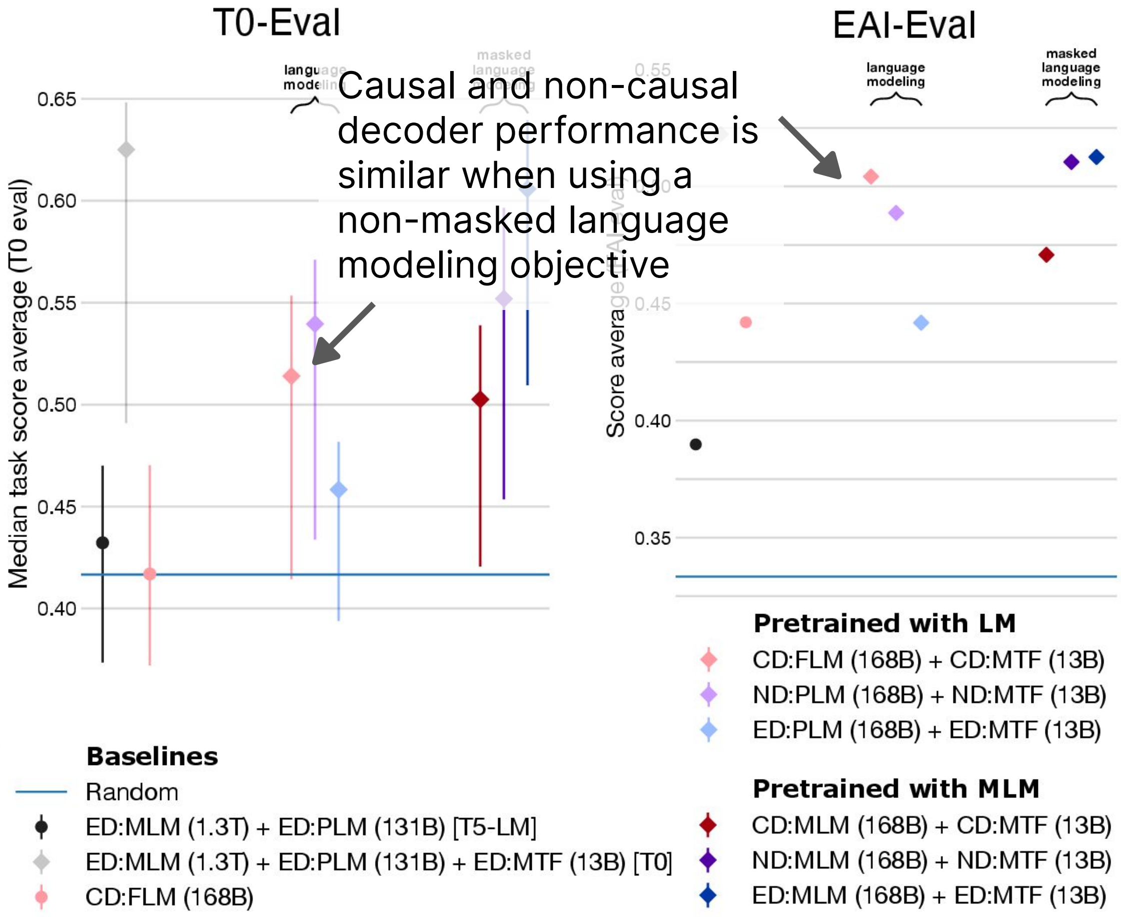
Prefix Language Modeling May the force be with you
targets

Masked Language Modeling May the force be with you
targets

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

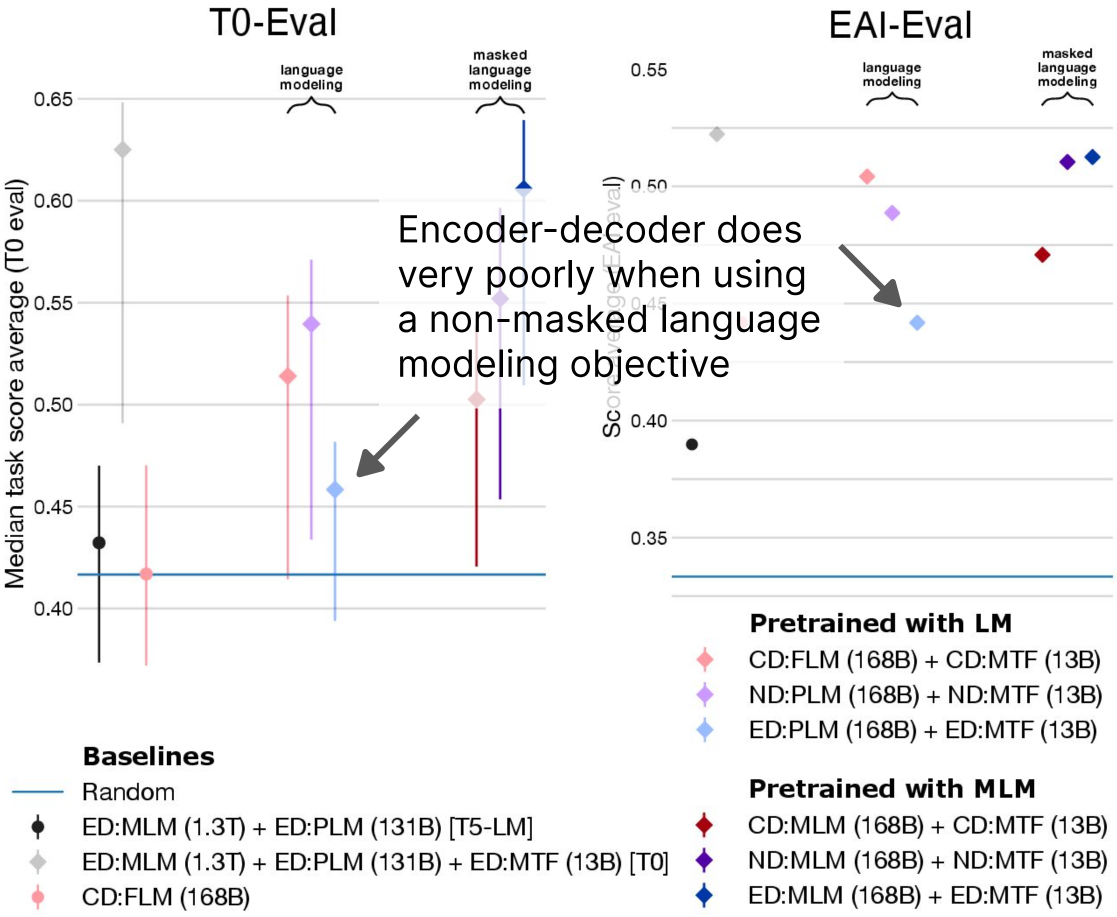


From "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?" by Wang et al.



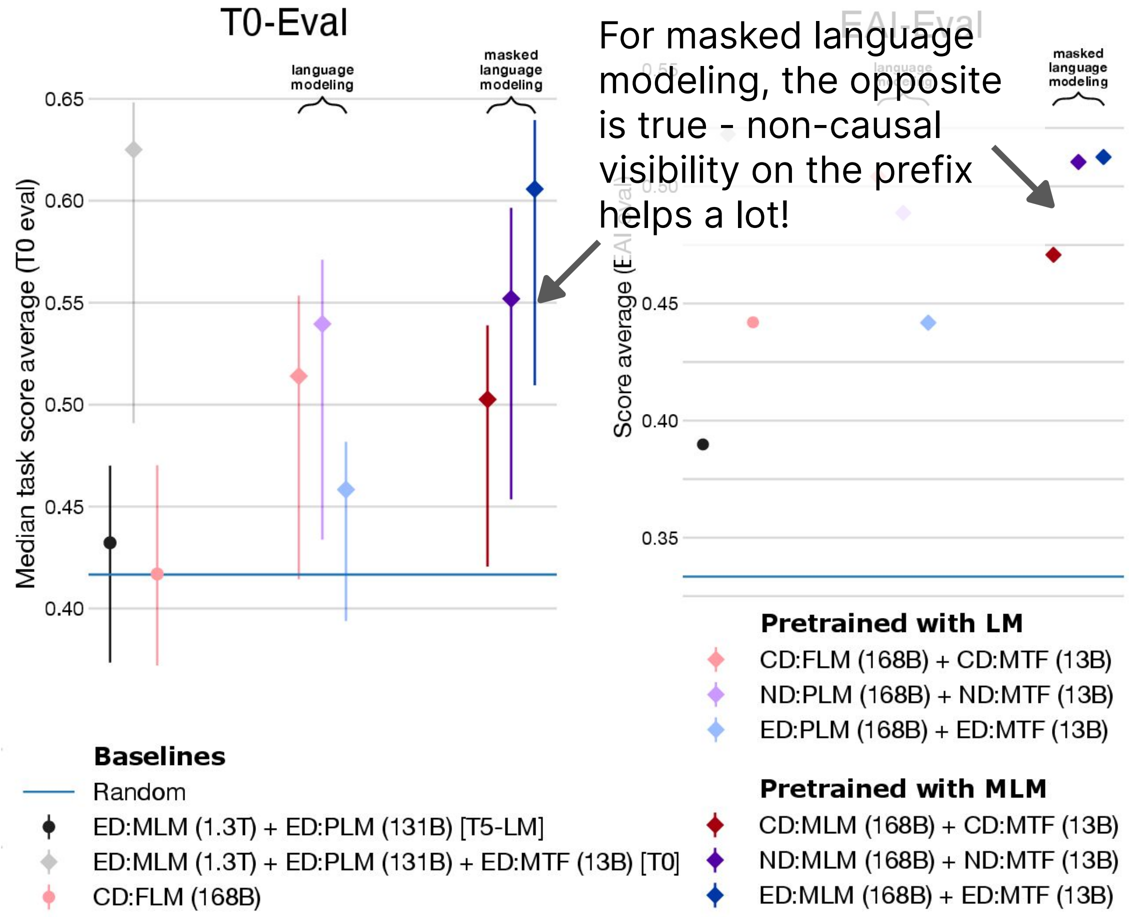
Causal and non-causal decoder performance is similar when using a non-masked language modeling objective

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



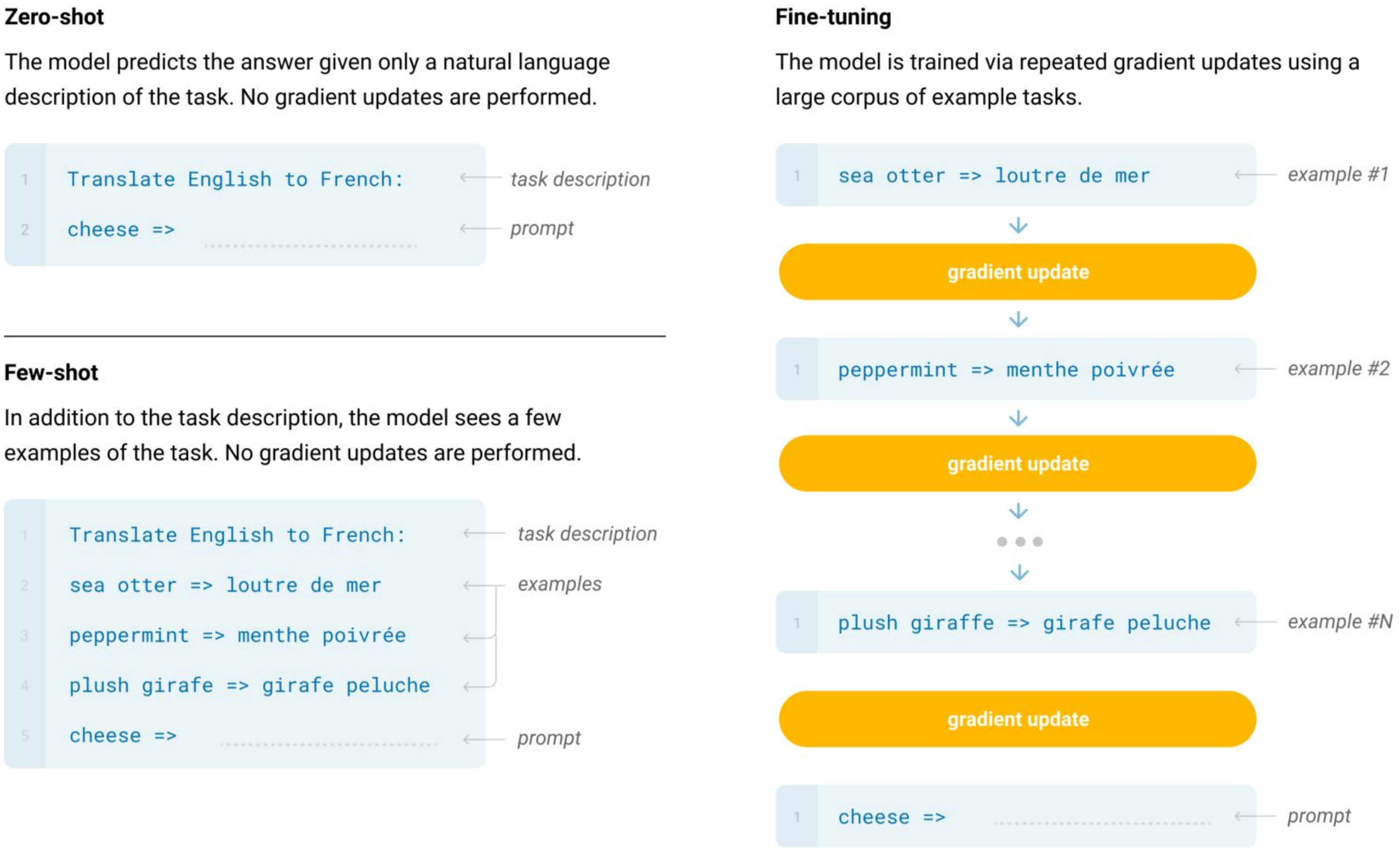
Encoder-decoder does very poorly when using a non-masked language modeling objective

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



For masked language modeling, the opposite is true - non-causal visibility on the prefix helps a lot!

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



Zero-shot

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

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

Few-shot

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

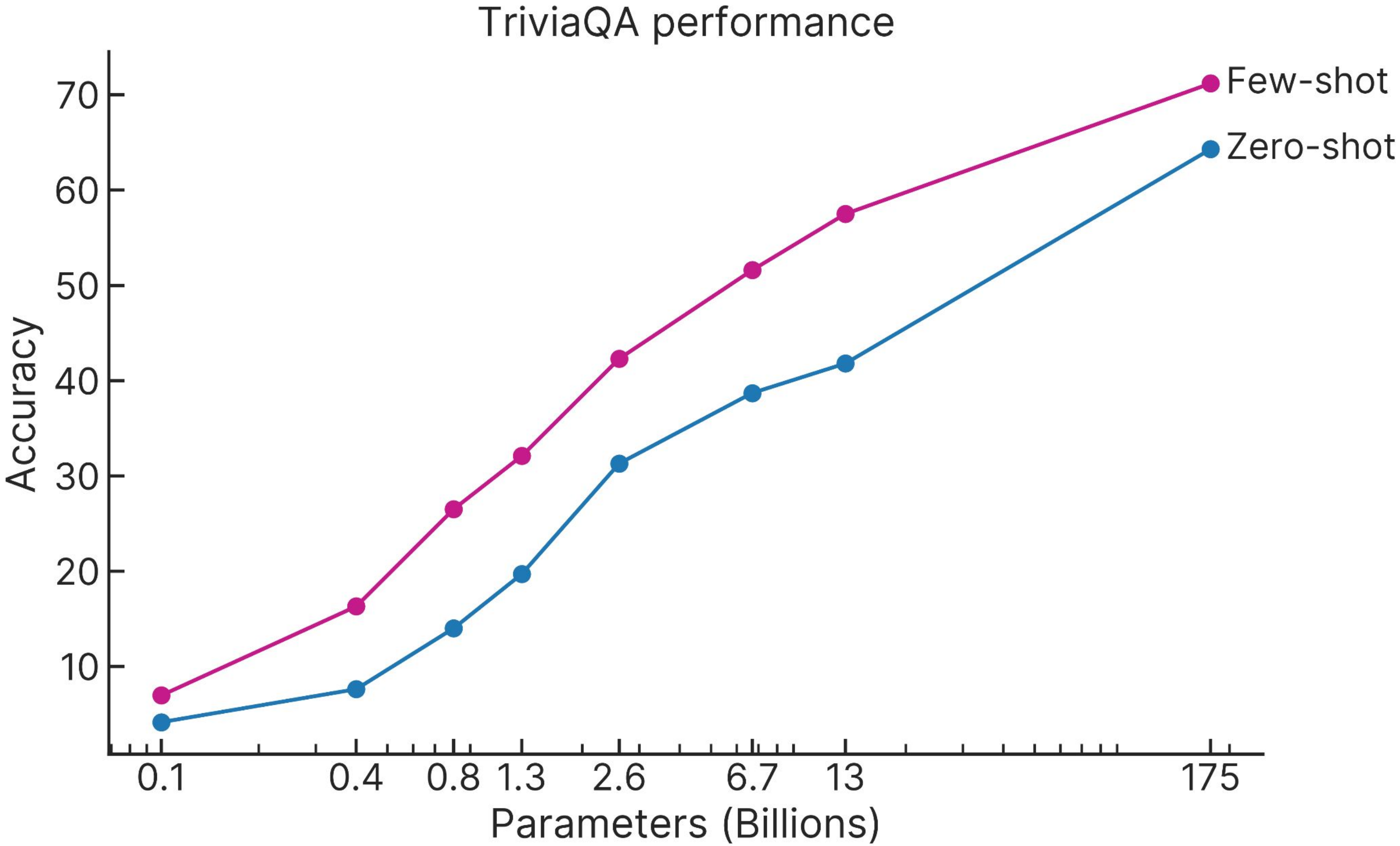
```
Translate English to French:      ← task description
sea otter => loutre de mer        ← examples
peppermint => menthe poivrée      ←
plush girafe => girafe peluche    ←
cheese => ........................ ← prompt
```

Fine-tuning

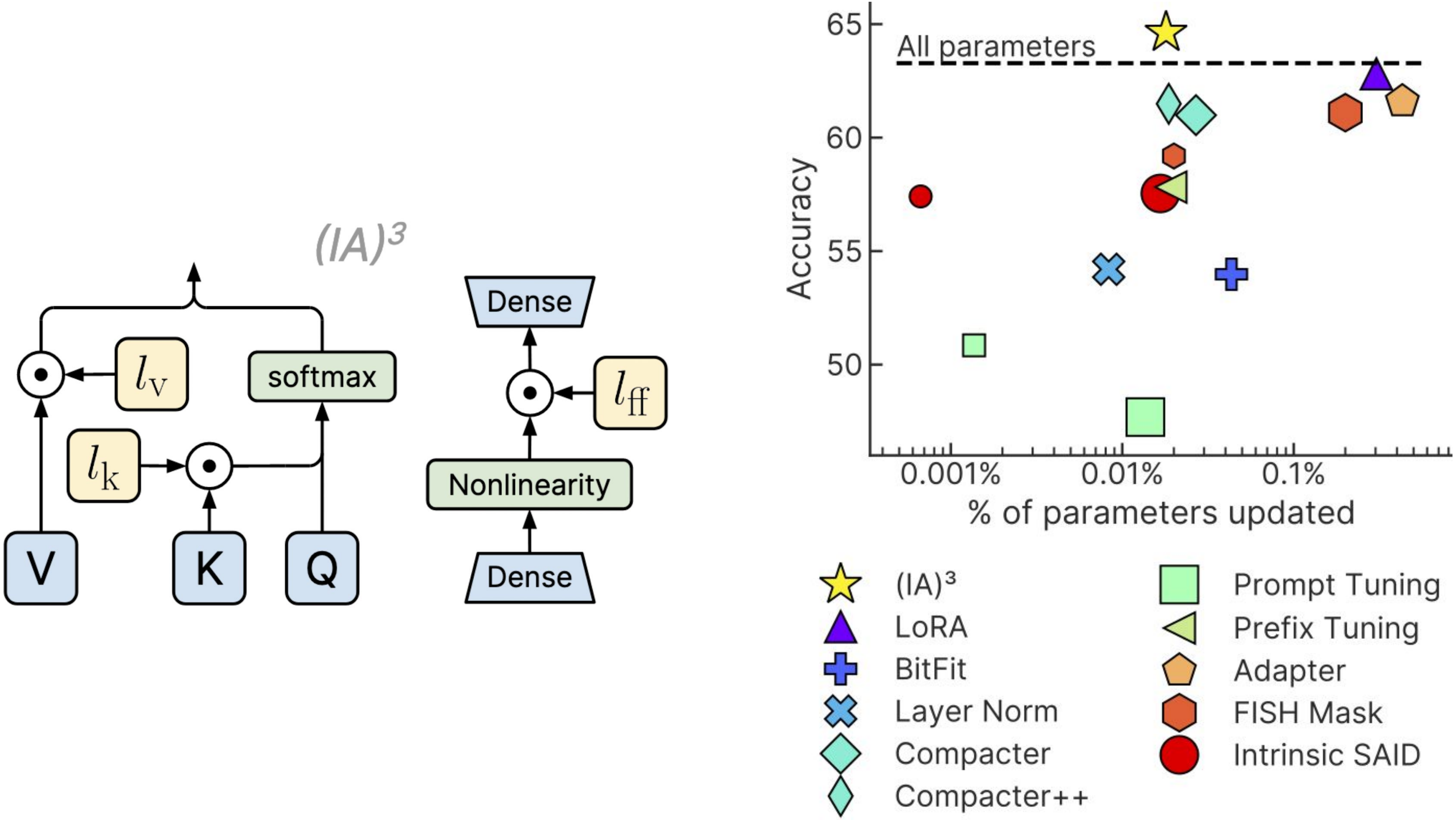
The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer               ← example #1
↓
gradient update
↓
peppermint => menthe poivrée             ← example #2
↓
gradient update
↓
• • •
↓
plush giraffe => girafe peluche          ← example #N

gradient update

cheese => ........................        ← prompt
```

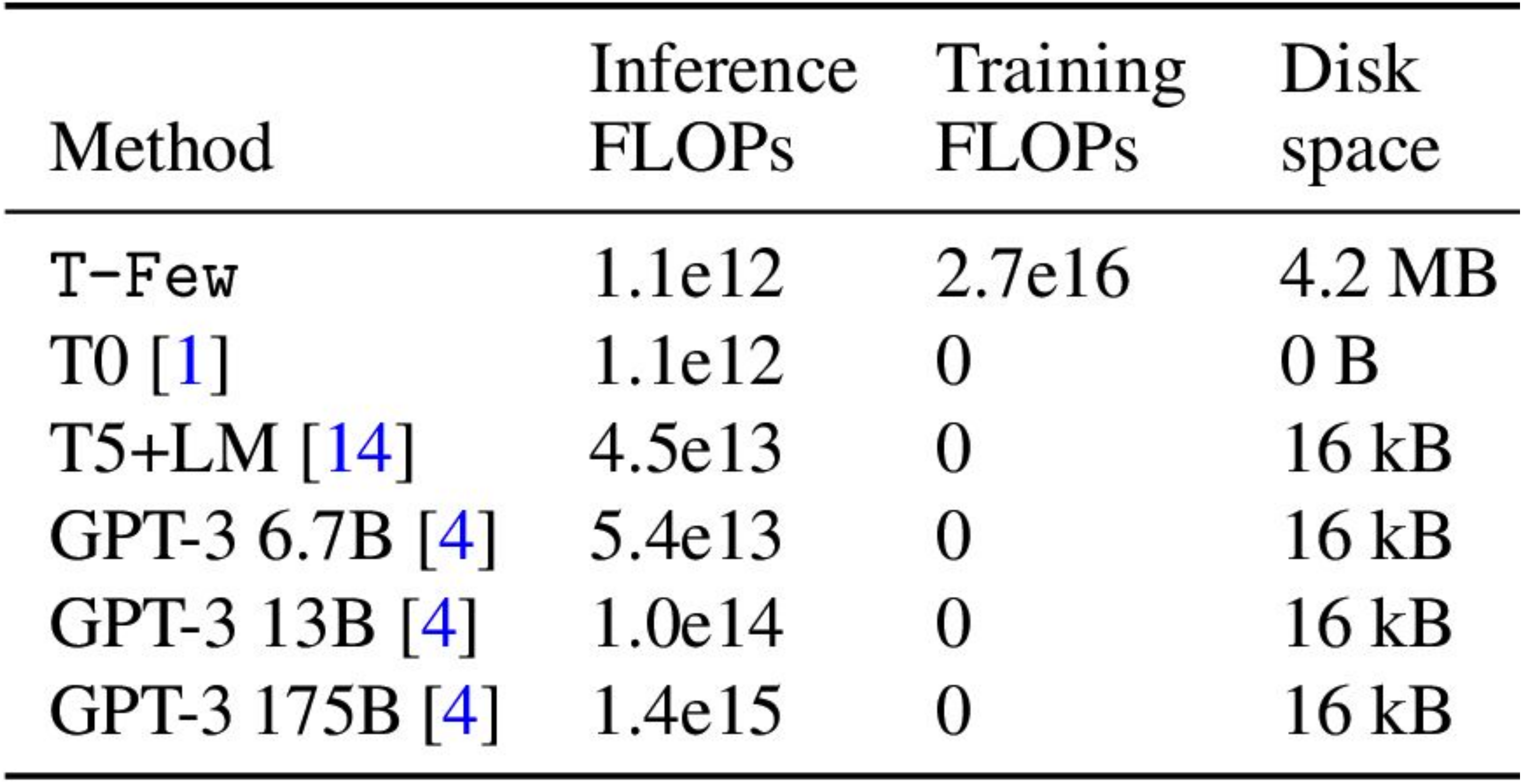
From “Language Models are Few-Shot Learners” by Brown et al.



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

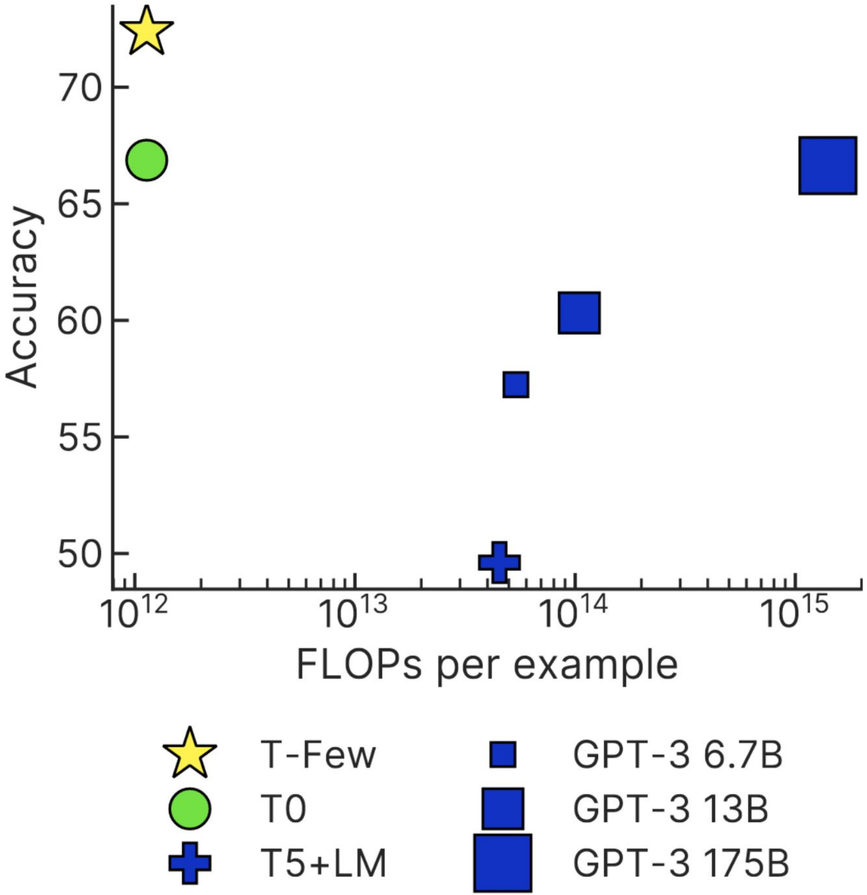


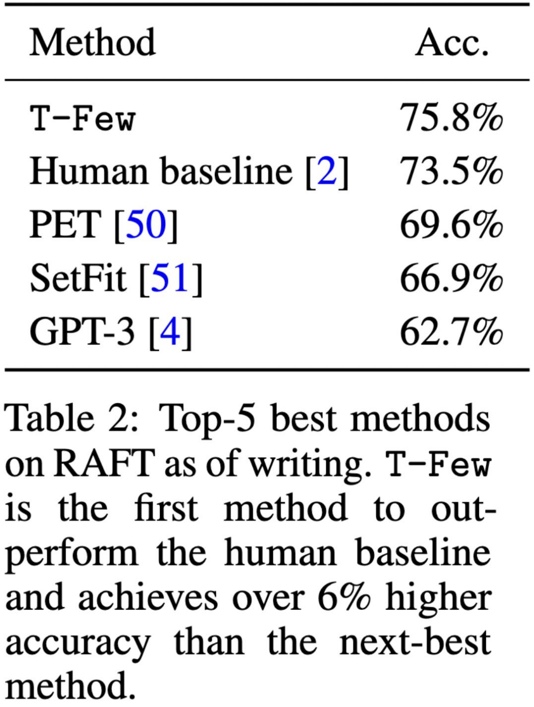
From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022



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

From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022





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.

From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022

References

Multitask Prompted Training Enables Zero-Shot Task Generalization

Crosslingual Generalization through Multitask Finetuning

What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

Please give me feedback:

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

Thanks!