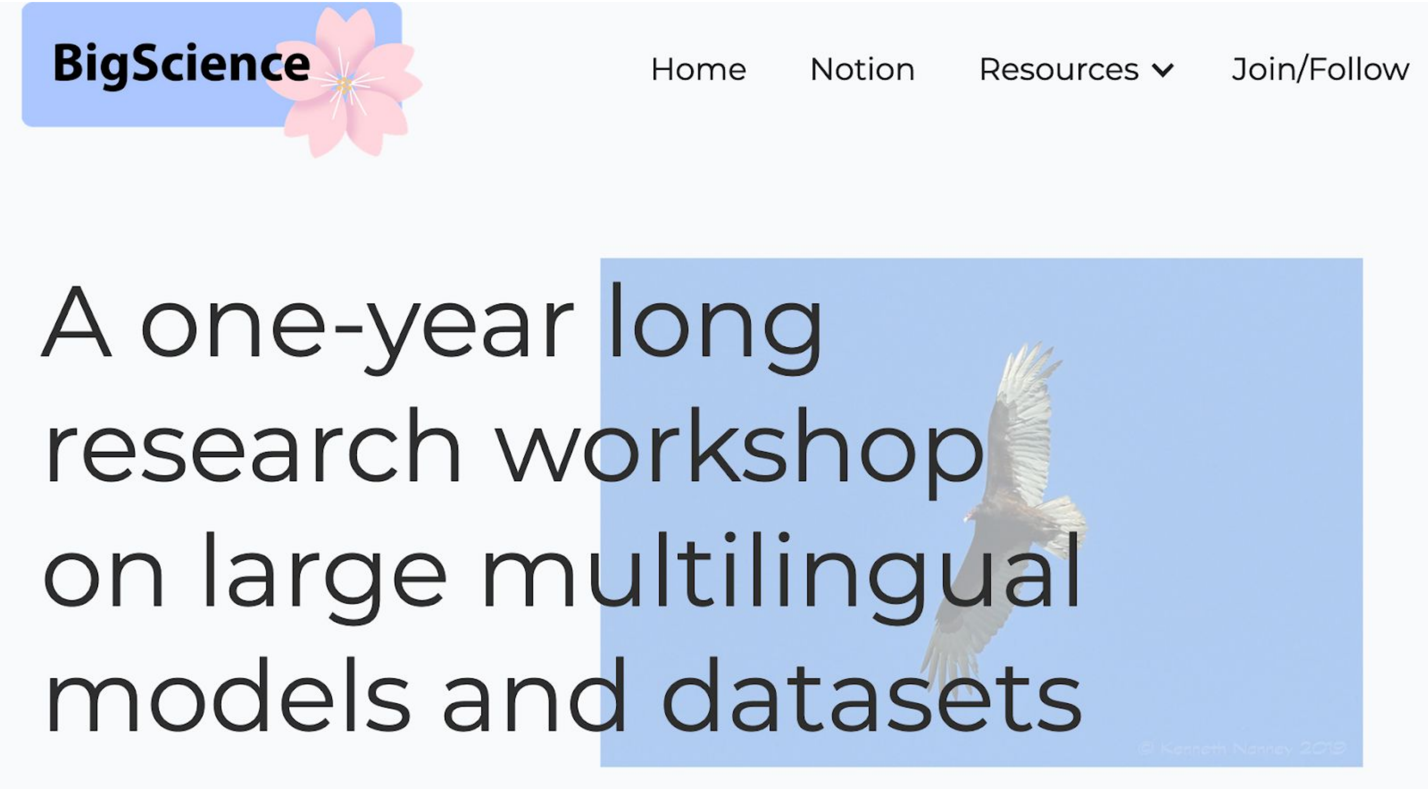
T0:
Multitask Prompt Training

Sasha Rush /w

BigScience

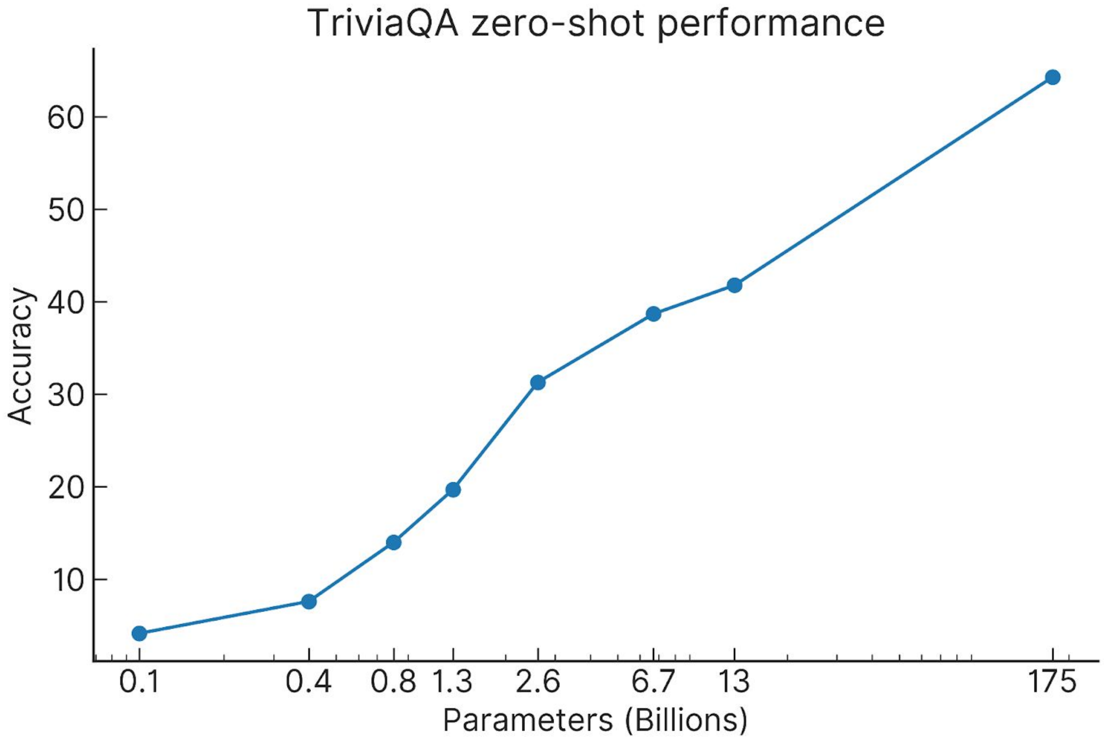






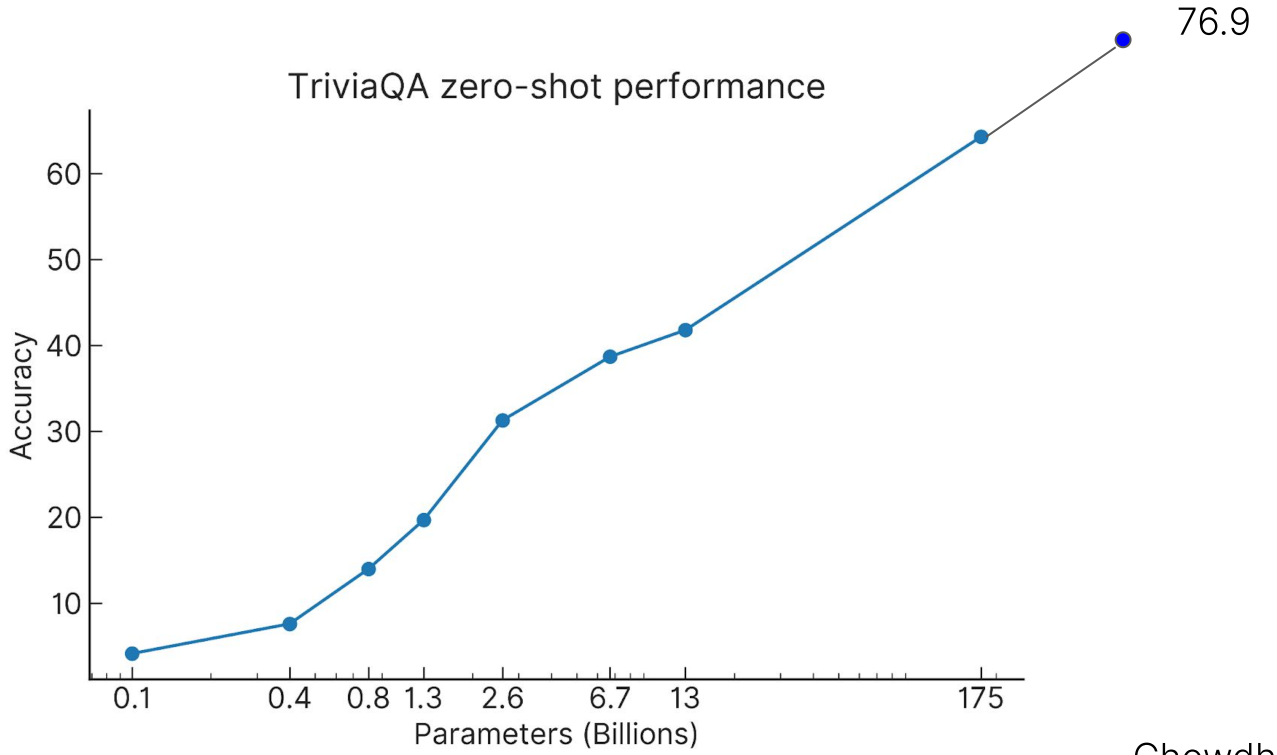
https://bigscience.huggingface.co/

Language Models are Few-Shot Learners



Brown et al. 2000

PaLM: Scaling Language Modeling with Pathways



Chowdhery et al. 2022

Zero-Shot

Q: *'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist?*

A: ....

Prompt Template

Q: {Question}

A: {Answer}

Few-Shot

Q: Which President of the Philippines was deposed in 1986?

A: Marcos

Q: Who was president of the USA at the outbreak of World War I?

A: Wilson

Q: *‘Nude Descending A Staircase’ is perhaps the most famous painting by which 20th century artist?*

A: ...

Today's Talk

- "Multitask prompted training enables zero-shot task generalization"

- Punchline ->

  Training on many NLP tasks improves generalization to new unseen tasks.

- Artifact ->

  T0 - A smaller model with strong zero-shot prompting abilities

Sanh et al. (2022)

Outline

- Preliminary Work
  - Datasets
  - How many data points is a prompt worth
- T0
- Context: BigScience

Preliminary Work: Datasets

(Lhoest et al, 2021)

Datasets: Tour of the library

```python
from datasets import load_dataset

dataset = load_dataset("boolq")

# Each dataset has a features schema and metadata.

print(dataset.features, dataset.info)

# Any slice of data points can be accessed directly without loading the full dataset into memory.

dataset["train"][start:end]

# Processing can be applied to every data point in a batched and parallel fashion using standard li-
# braries such as NumPy or Torch.

tokenized = dataset.map(tokenize, num_proc=32)
```

Datasets: Internals

Apache Arrow:

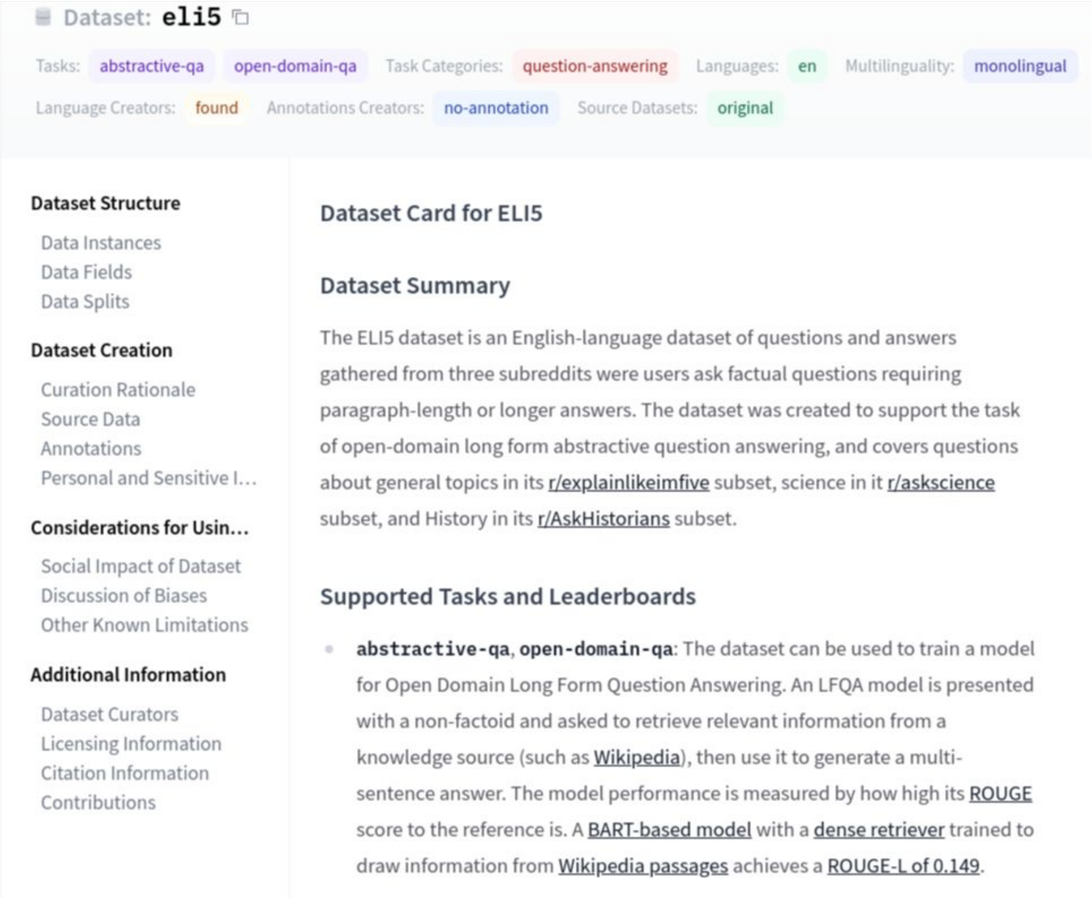
- language-independent columnar memory format
- memory-mapping to load terabytes of data without using RAM
- zero-copy reads for fast data access without serialization overhead
    - <1ms latency even on billion-scale datasets
    - end-to-end zero-copy to deep-learning frameworks

*jax not fully end-to-end

Dataset cards

- document the datasets
- community-driven
- dynamic
- search by task/lang/etc.

- standardized types
- get feature names
- types across dataset



Datasets: Meta-Datasets

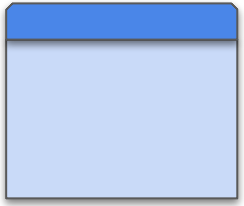
- Benchmarks: LM Evaluation Harness
- Workshops / Shared tasks: GEM
- Robustness evaluation: Robustness Gym

Preliminary Work:
How many data points is a prompt worth?

(Le Scao et al, 2021)

Finetuning with Prompting

1. Start from pre-trained language model



2. *Modify labeled training data to prompted form*

Natural language decathlon (McCann et al. 2018) - GPT2 (Radford et al. 2019) - GPT3 (Brown et al. 2020) - T5 (Raffel et al. 2019) - PET (Schick et al. 2020) - Zero-shot text classification (Puri et al. 2019)

Goals

- Sanity check the use of prompts in training.
- Does training with prompts improve over standard labels?
- How can we measure that difference?

Experimental setup

RoBERTa-Large
Testing on SuperGLUE + MNLI
Best of 4 runs on every data size

- Linear classification head
- Fine-tuned via backpropagation on the predicted class

- Task-adaptation with a prompt (3-4 different prompts per task)
- Fine-tuned via backpropagation on the predicted output token

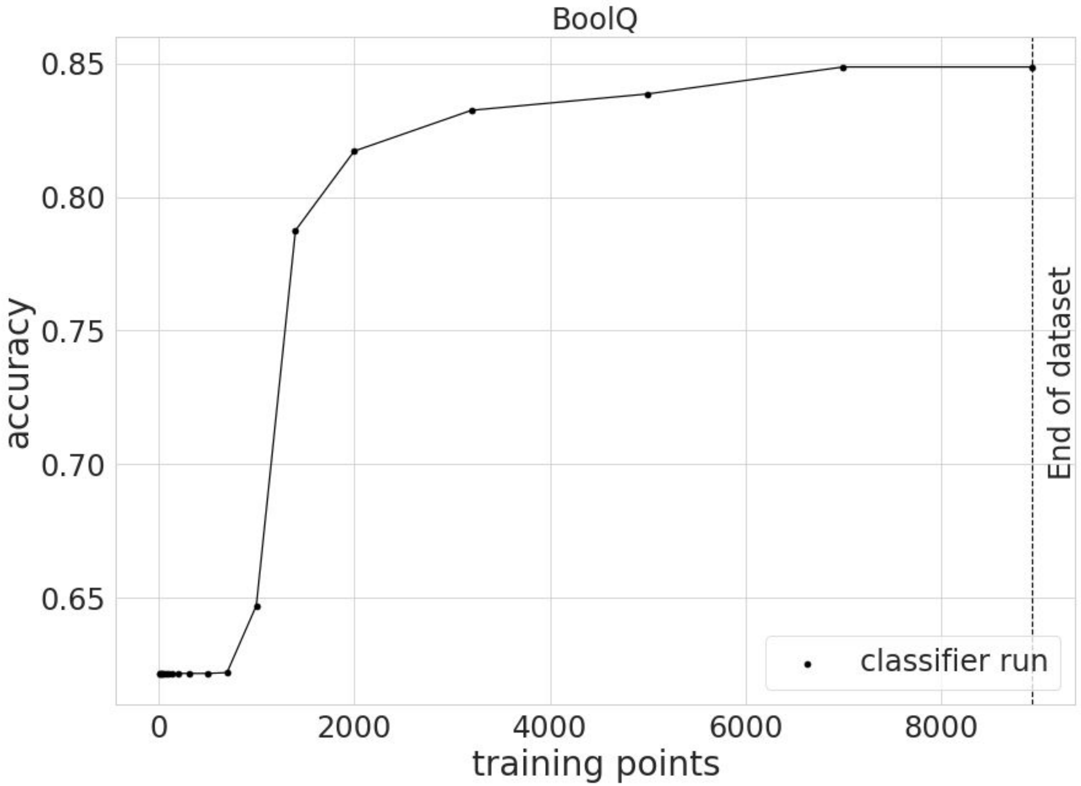
Choice of prompts

Prompts from *It's Not Just Size That Matters* (Schick and Schütze 2020) For BoolQ, for example:

- {passage}. Question: {question}? Answer: ….
- {passage}. Based on the previous passage, {question}?....

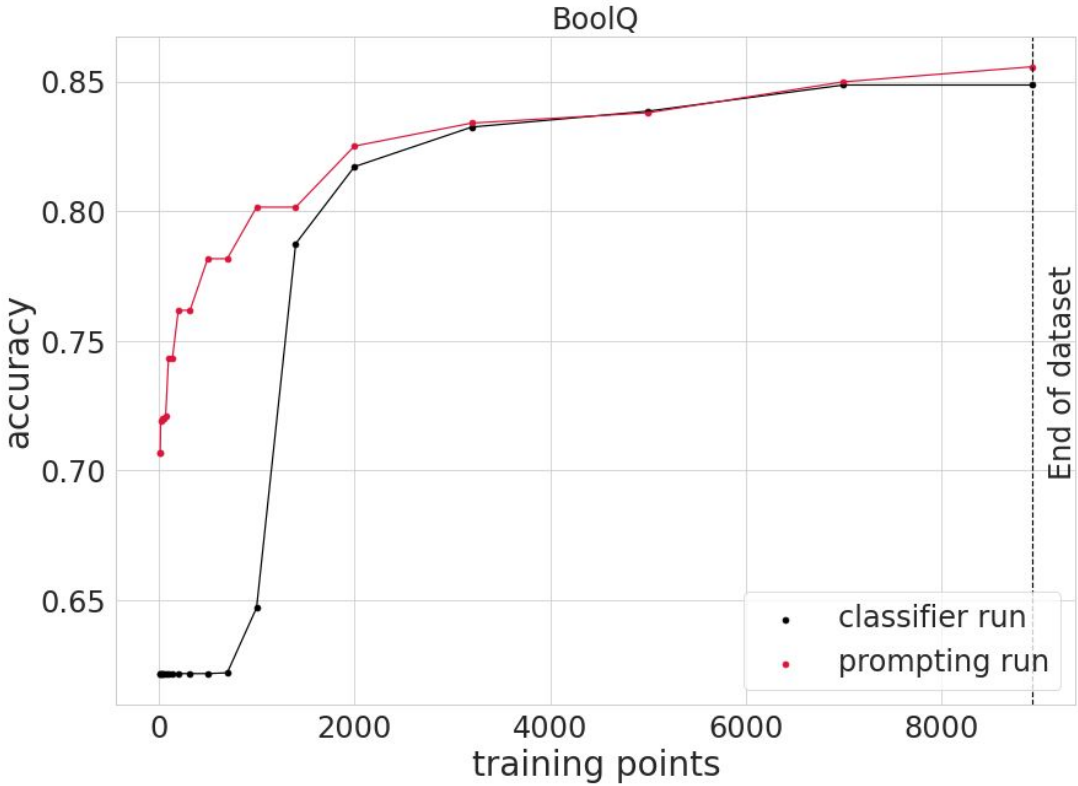
Data Advantage

Performance vs. dataset size on BoolQ for the classifier model.



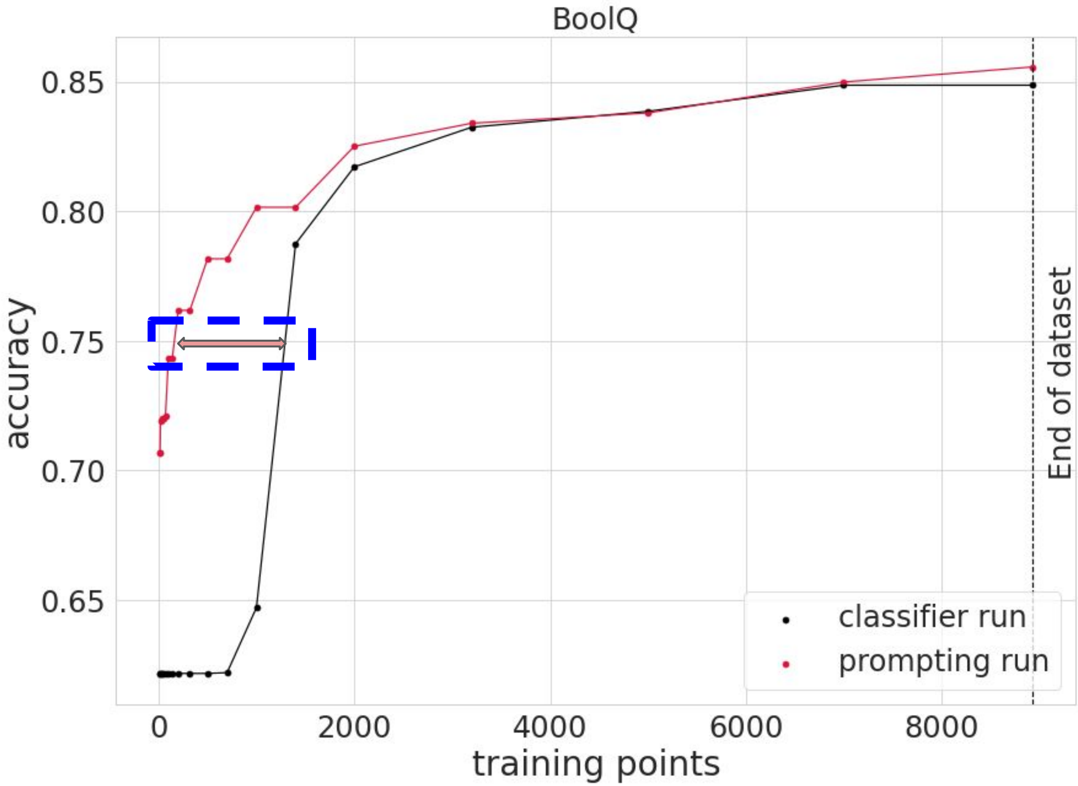
Data Advantage

Performance vs. dataset size on BoolQ for the classifier and prompting models.



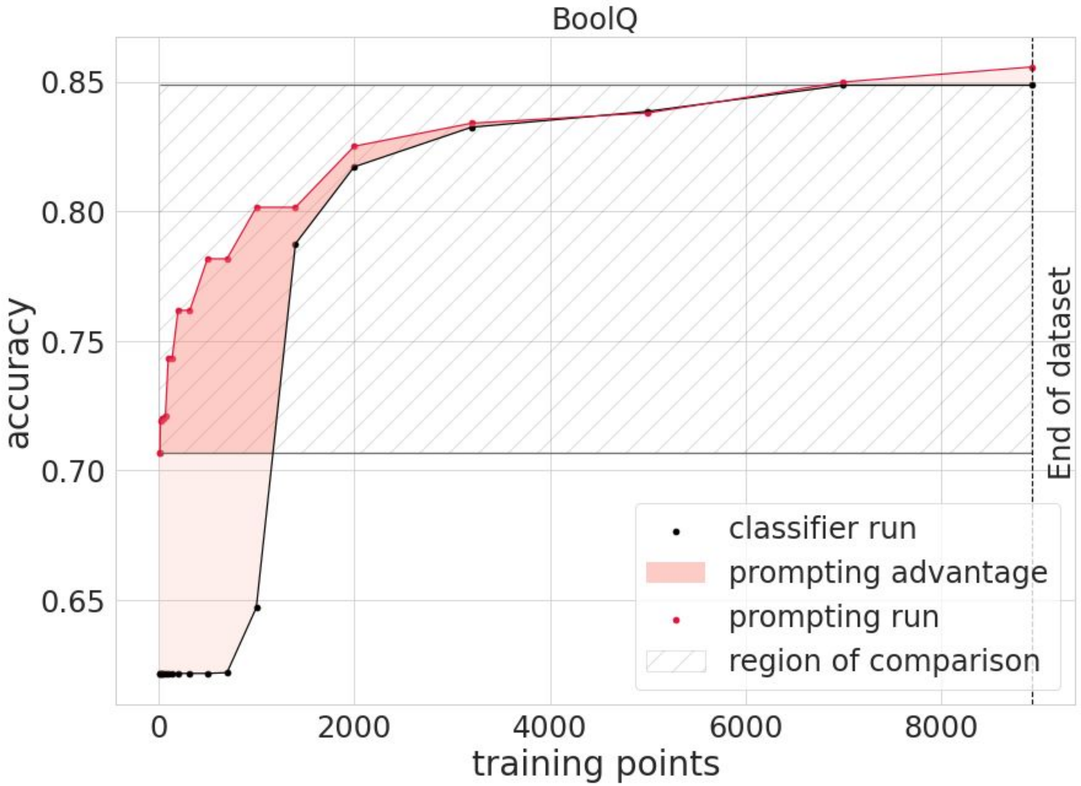
Data Advantage

The prompted model reaches 0.75 accuracy with 1132 data points less than the classifier.

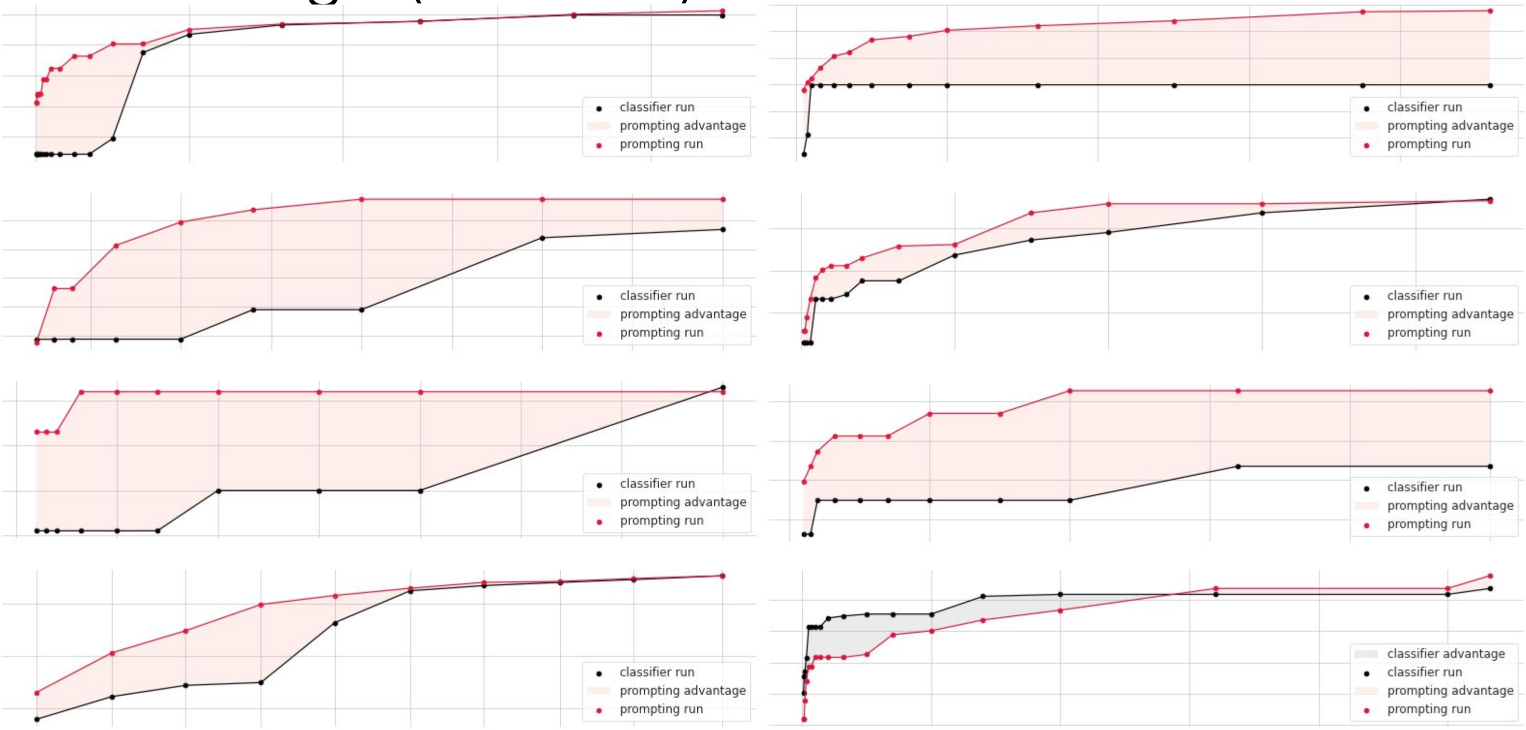


Data Advantage

Over the whole region, the prompted model is 752 data points ahead of the classifier on average.



Data advantage (all tasks)



BoolQ
752±46

MultiRC
384±378

CB
90±2

RTE
282±34

COPA
288±242

WSC
281±137

MNLI
3506±536
(x log scale)

WiC
-424±74

What we know

- Does the model understand the prompt?
  - Probably not. (Webson & Pavlick, 2022)
- Does the prompt need to be human understandable?
  - Not clear, particularly in few-shot versions.
- What can we say?
  - Language is a convenient modality for task encoding.

T0

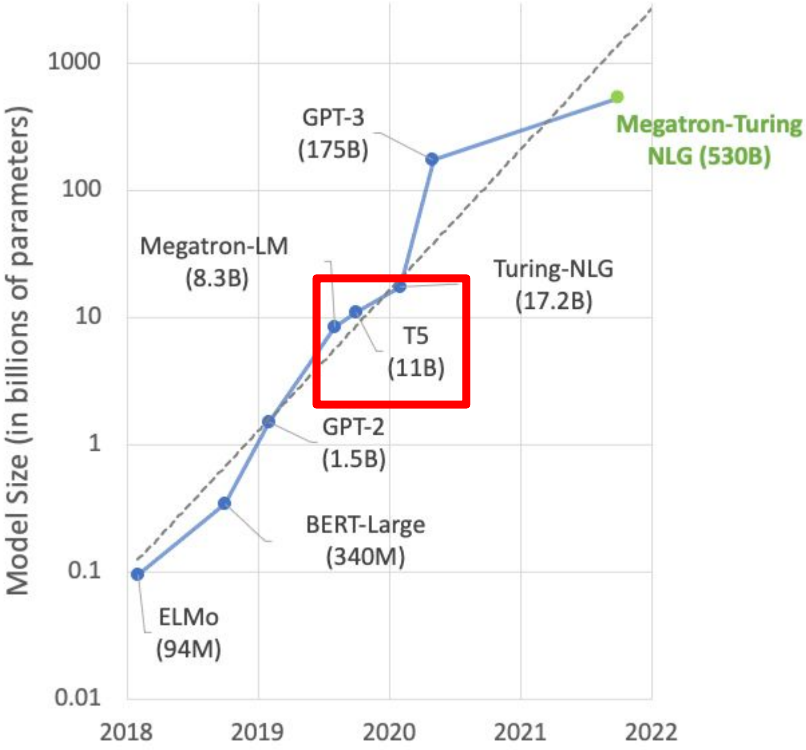
BigScience

(Sanh et al, 2022)

Research Question

- Can we induce zero-shot task transfer through pretraining on prompts?
- Practical benefit → Smaller models with zero-shot ability
- Research → Generic pretraining versus targeted induction.

Review: T5



Text-to-Text Transfer Transformer

T5

T5 - Unsupervised Pretraining Stage

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 ___.

T5

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

For example, we might train a single model on many tasks, but

T5+LM

when reporting performance we are allowed to select a different checkpoint for each …

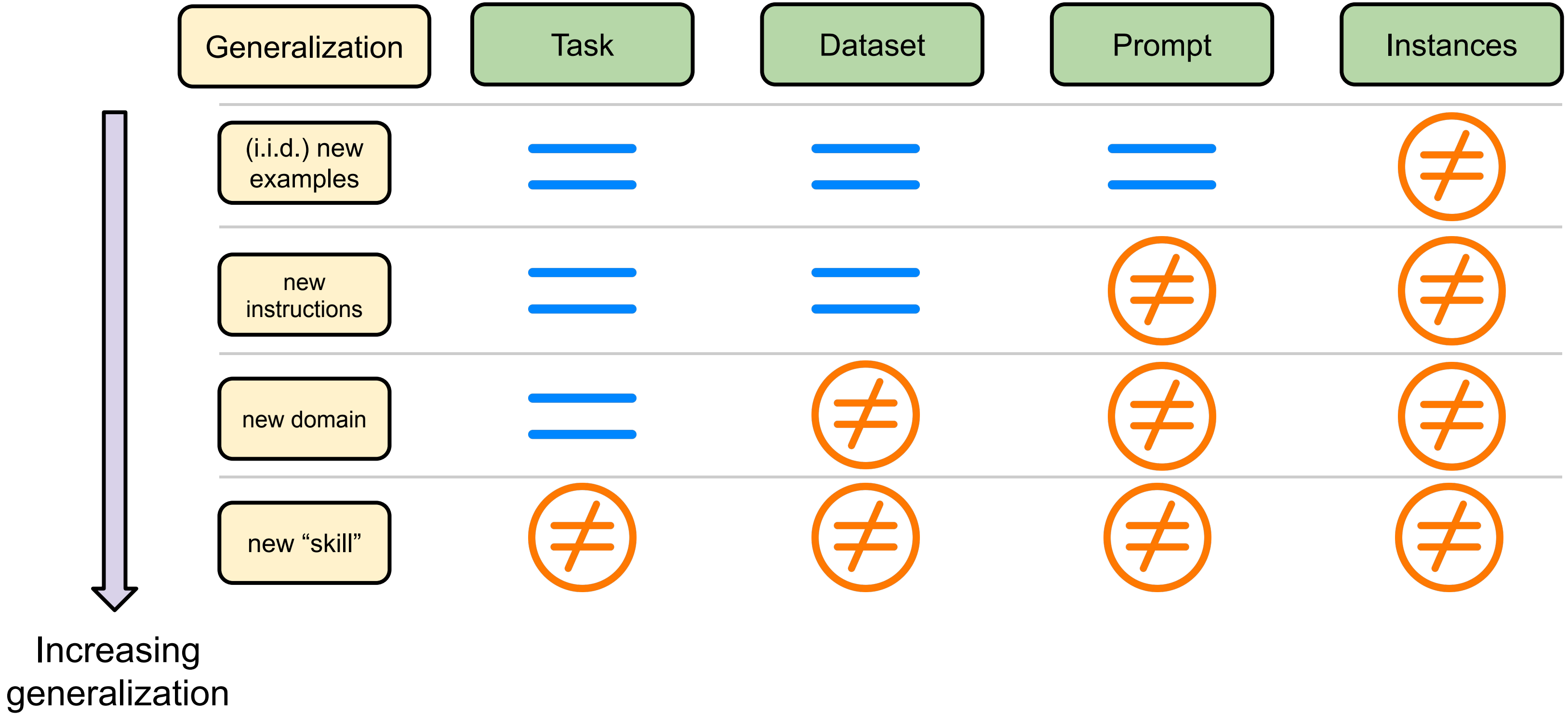
T5 (Raffel et al. 2019), T5+LM (Lester et al. 2021)

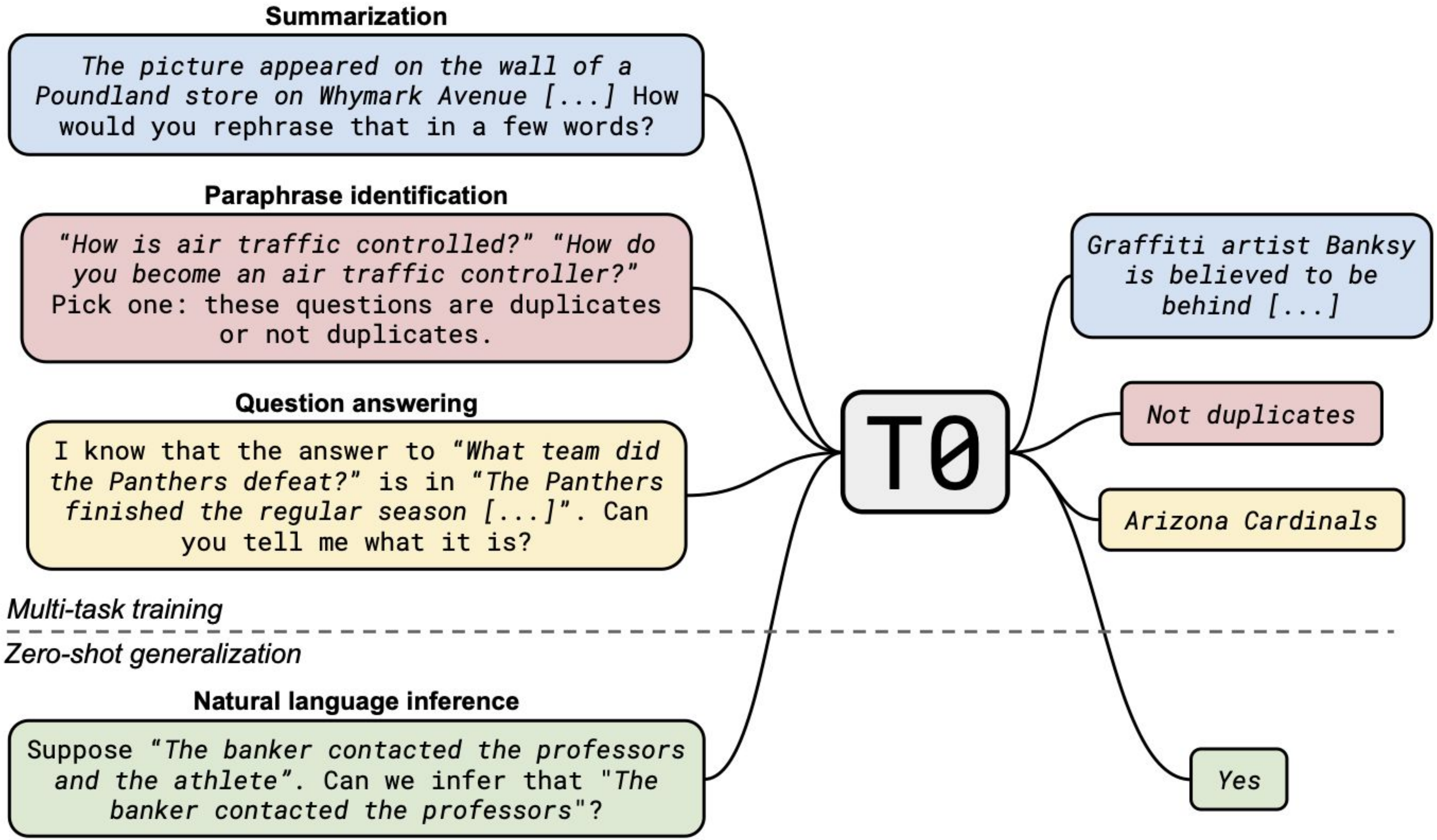
T0 Recipe

- Produce templates for turning a large set of datasets to prompts.

- Pretrain T5 LM on those prompts for a significant amount of time.

- Evaluate model on tasks it has not seen before.


Generalization
Task
Dataset
Prompt
Instances
(i.i.d.) new examples
new instructions
new domain
new “skill”
Increasing generalization


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

PromptSource:
Prompts for Training

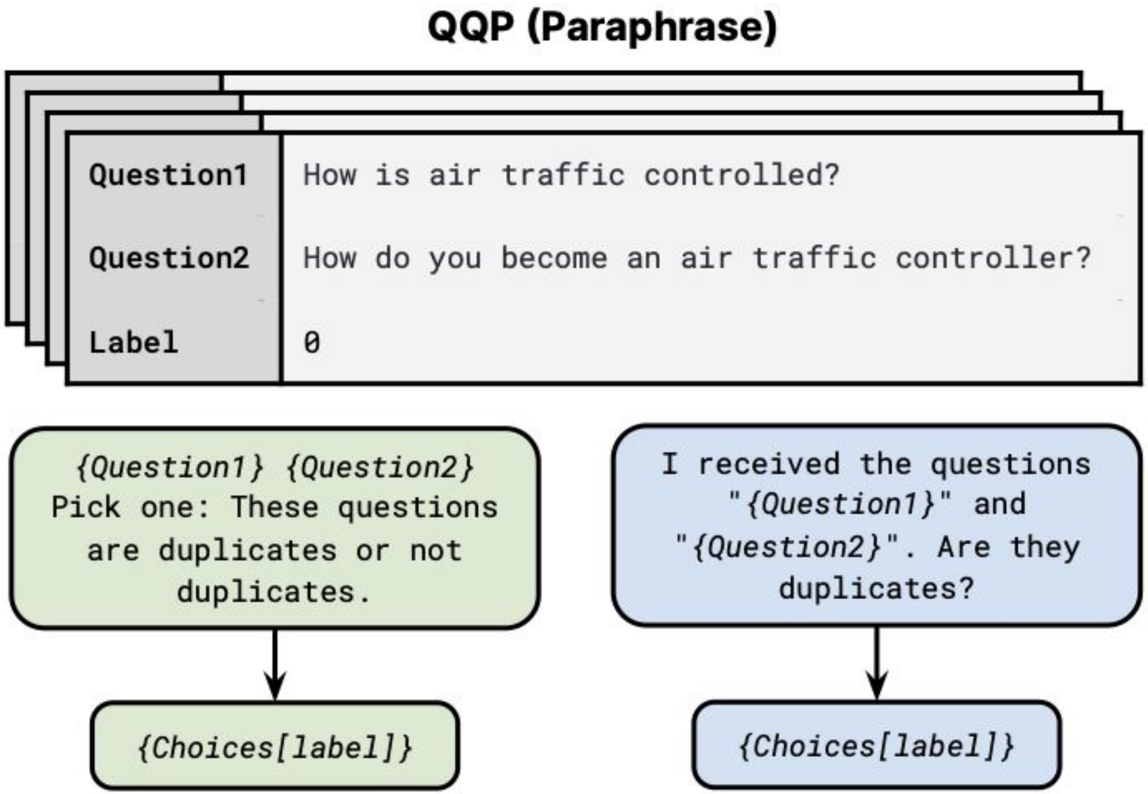
Closed-book question answering
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
who was frank sinatra? a: an american singer, actor, and producer.

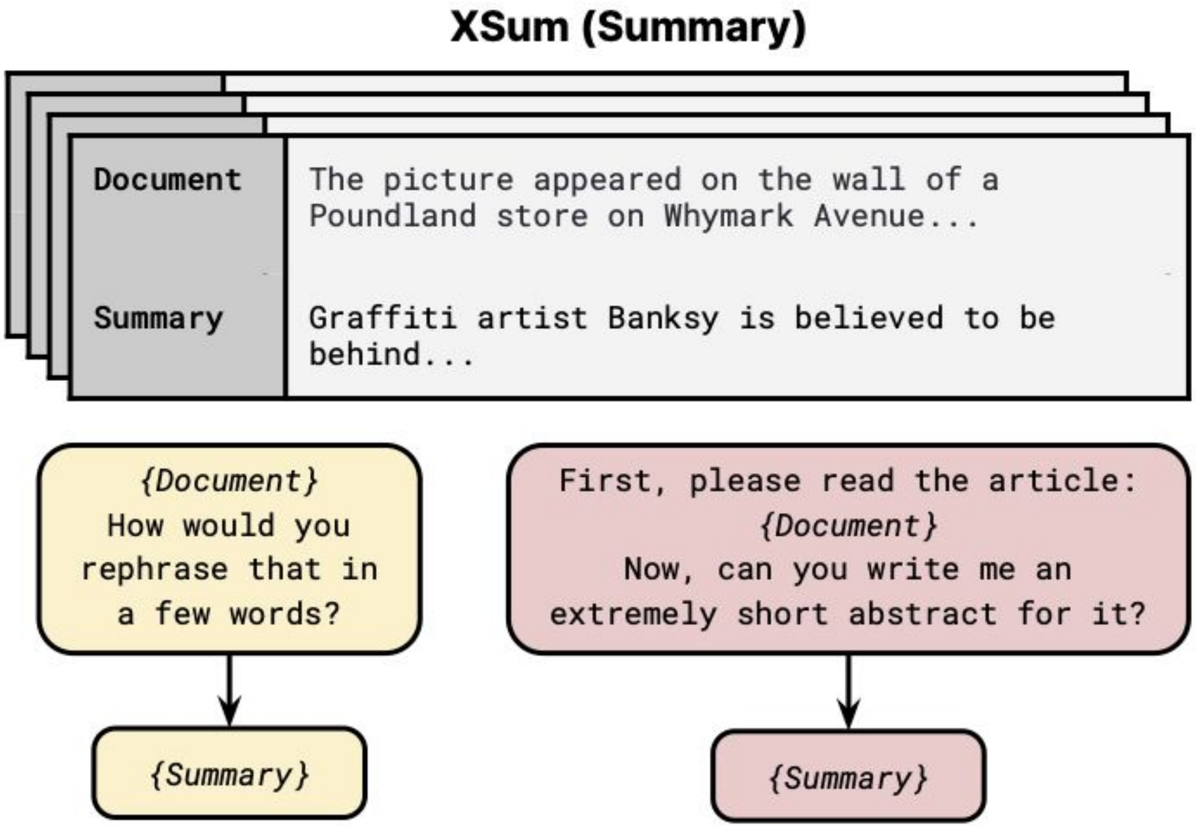
Paraphrase identification
https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same
Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

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.


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}

Prompt Template Language

Jinja template

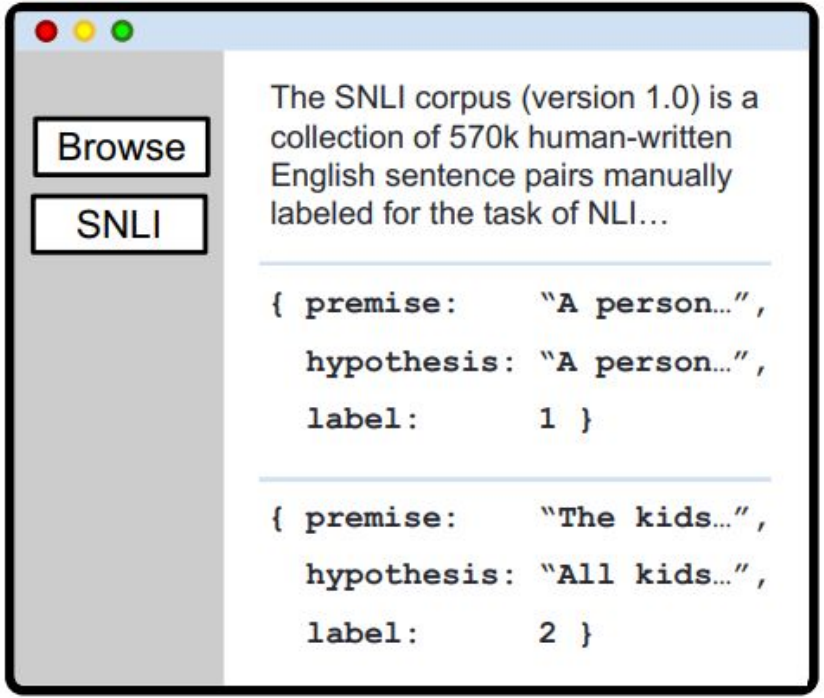
Input template

```
{{premise}}
Question: {{hypothesis}} True, False, or Neither?
```

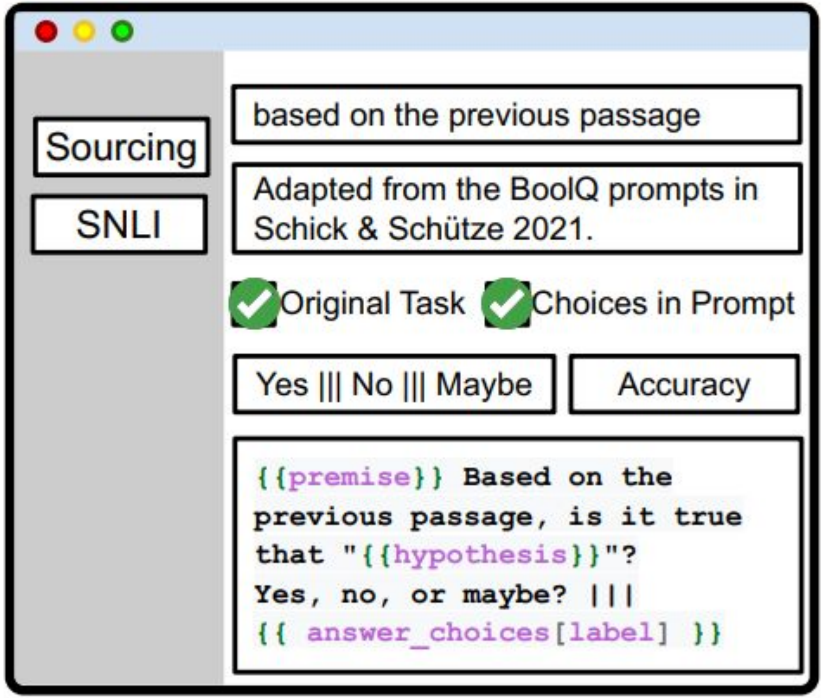
Target template

```
{{ answer_choices[label] }}
```

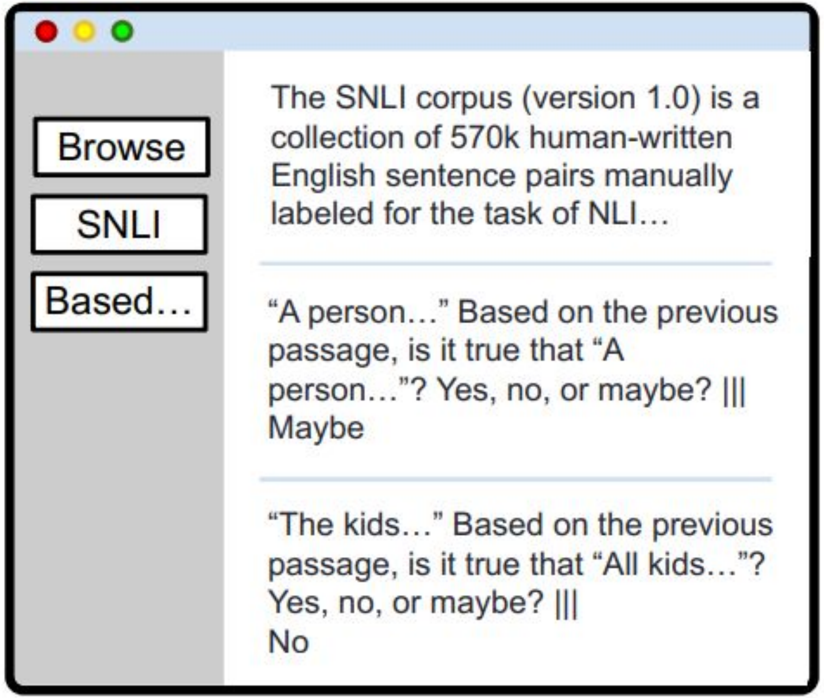
S1: Exploration



S2 + S3 + S4: Creation



S5: Review


Browse
SNLI
Based…
The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for the task of NLI…
"A person…" Based on the previous passage, is it true that "A person…"? Yes, no, or maybe? |||
Maybe
"The kids…" Based on the previous passage, is it true that "All kids…"? Yes, no, or maybe? |||
No

Number of *prompted datasets*: 180

Number of *prompts*: 2085

Prompt Template Language

Answer Choices Key

```
{{choice1}} ||| {{choice2}}
```

Template

```
{{ premise }}

I am hesitating between two options. Help me choose the more likely {% if question == "cause" %} cause: {% else %} effect: {% endif %}
- {{choice1}}
- {{choice2}} ||| {% if label != -1 %}{{ answer_choices[label] }}{%endif%}
```

Save

Prompt + X

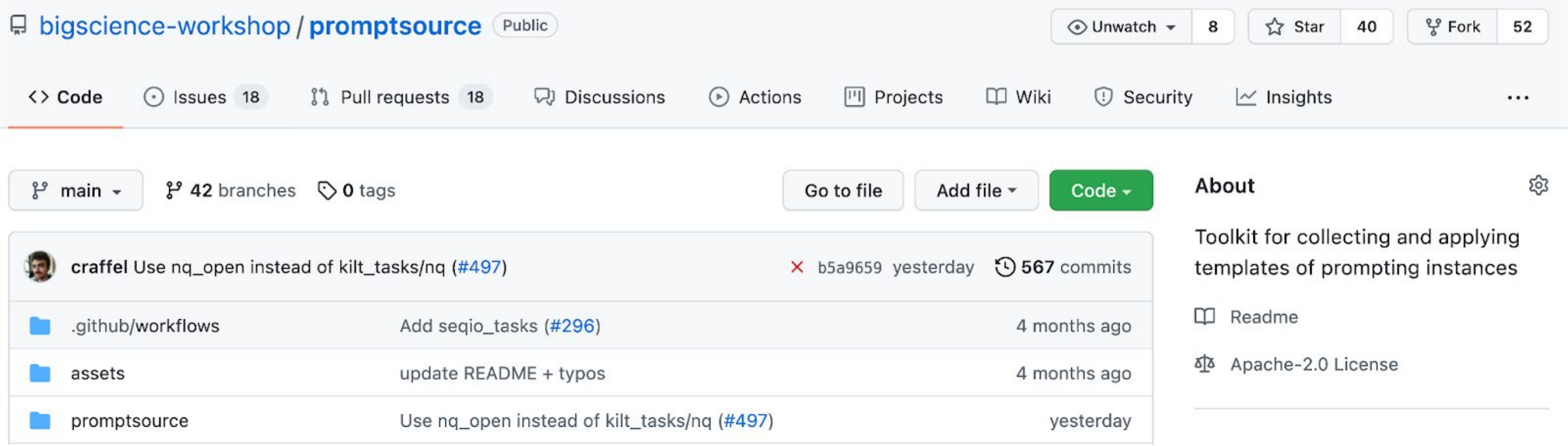
My body cast a shadow over the grass.

I am hesitating between two options. Help me choose the more likely cause:

- The sun was rising.
- The grass was cut.

Y

The sun was rising.



https://github.com/bigscience-workshop/promptsource

Extensions: BigBIO

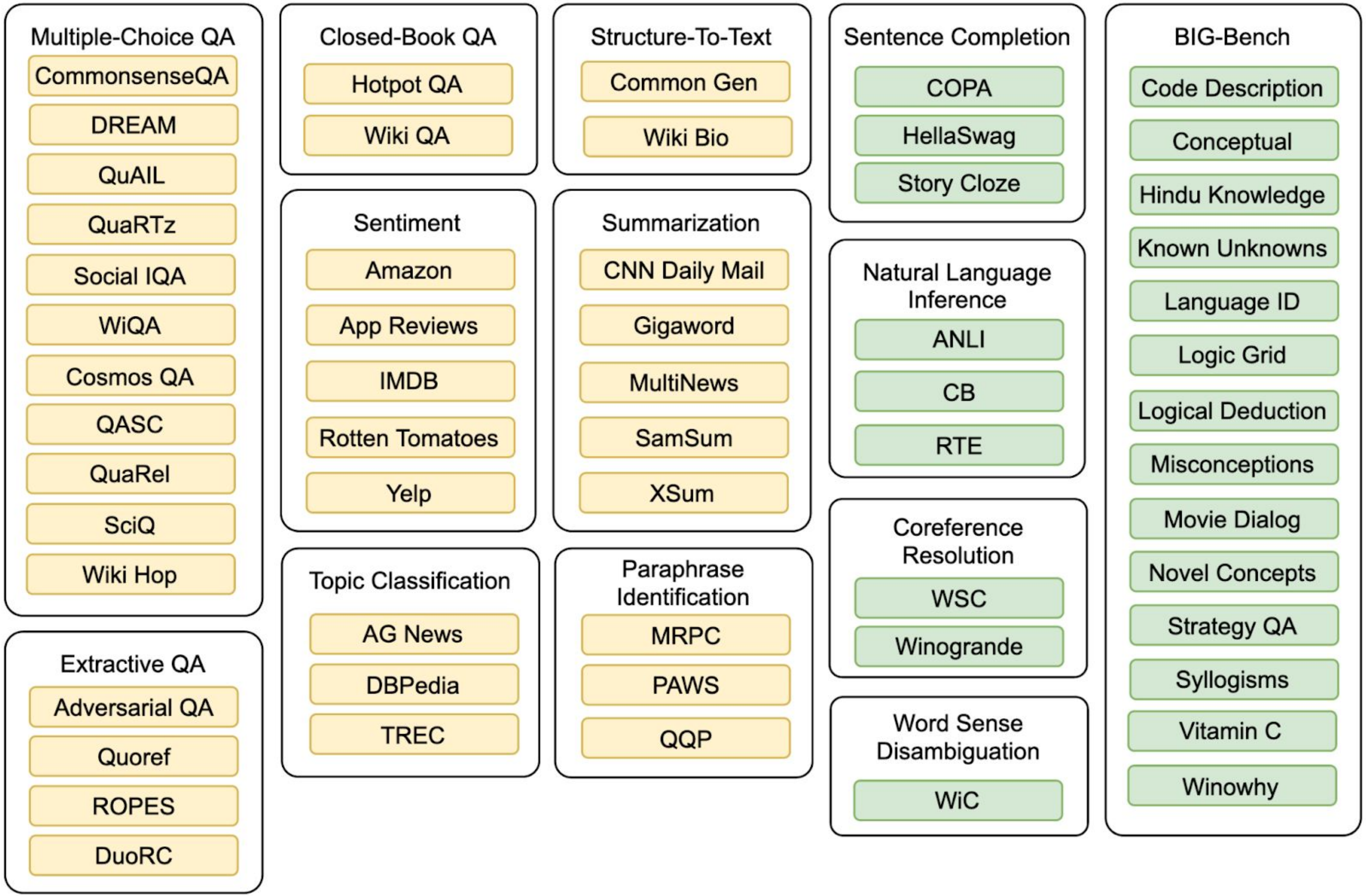
Task Type	Input	Label
RE	Taken together, these results make it clear that @chemical$-bound forms of ORC and @protein$ are likely to be required for productive interactions and pre-RC formation.	`bind`
COREF	We investigated the potential of the @aryl hydrocarbon receptor$ (@AHR$) to suppress NF-kappaB regulated-gene expression, especially acute-phase genes, such as serum amyloid A (Saa).	`coref`
EAE	v-erbA @Gene_expression$ is required to @Negative_regulation$ c-erbA function in erythroid cell differentiation and regulation of the erbA target gene CAII.	cause

Comparison: Natural Instructions v2

- PromptSource was post-hoc instruction generation
- PromptSource has less tasks, but multiple instructions per task
- PromptSource tasks are single language.

Wang et al. 2022

T0 - Experiments


Multiple-Choice QA
CommonsenseQA
DREAM
QuAIL
QuaRTz
Social IQA
WiQA
Cosmos QA
QASC
QuaRel
SciQ
Wiki Hop
Extractive QA
Adversarial QA
Quoref
ROPES
DuoRC
Closed-Book QA
Hotpot QA
Wiki QA
Sentiment
Amazon
App Reviews
IMDB
Rotten Tomatoes
Yelp
Topic Classification
AG News
DBPedia
TREC
Structure-To-Text
Common Gen
Wiki Bio
Summarization
CNN Daily Mail
Gigaword
MultiNews
SamSum
XSum
Paraphrase Identification
MRPC
PAWS
QQP
Sentence Completion
COPA
HellaSwag
Story Cloze
Natural Language Inference
ANLI
CB
RTE
Coreference Resolution
WSC
Winogrande
Word Sense Disambiguation
WiC
BIG-Bench
Code Description
Conceptual
Hindu Knowledge
Known Unknowns
Language ID
Logic Grid
Logical Deduction
Misconceptions
Movie Dialog
Novel Concepts
Strategy QA
Syllogisms
Vitamin C
Winowhy

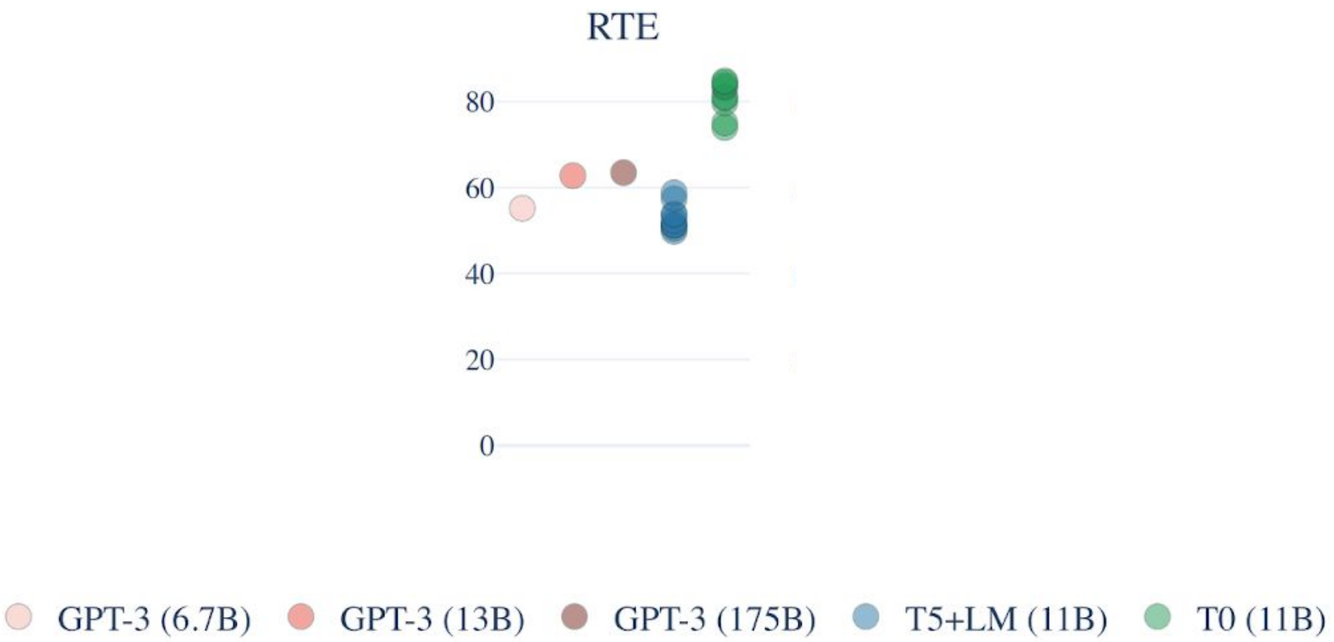
Experimental Details

- Based on T5-LM model , 11B parameters

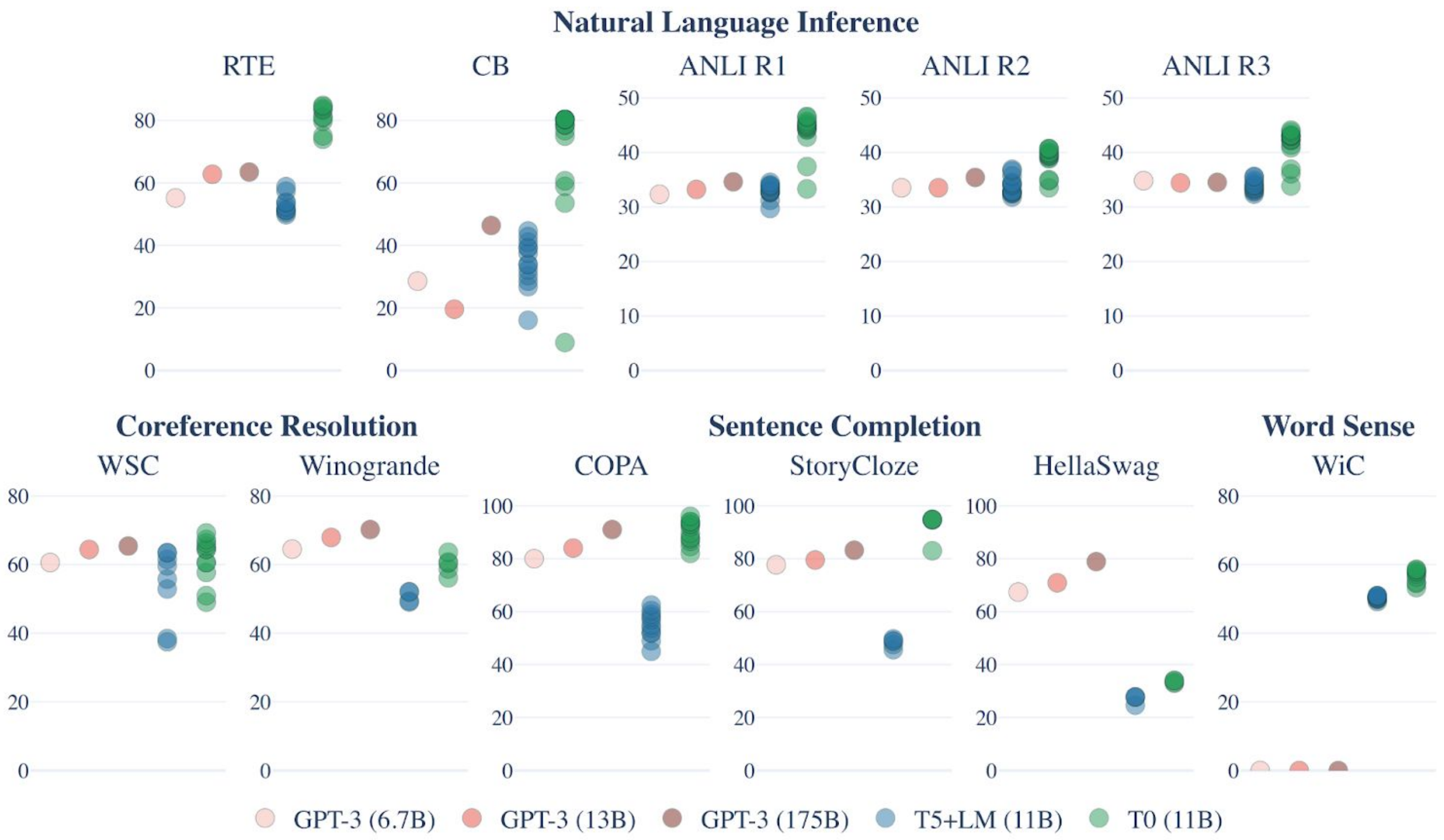
- Comparison to GPT-3 (6.7, 13, 175 B parameters)

- Uniformly sampled from datasets and prompts

- Evaluated on held out task types, across prompts


RTE
80
60
40
20
0
GPT-3 (6.7B)
GPT-3 (13B)
GPT-3 (175B)
T5+LM (11B)
T0 (11B)

Performance on held-out tasks



BIG-Bench

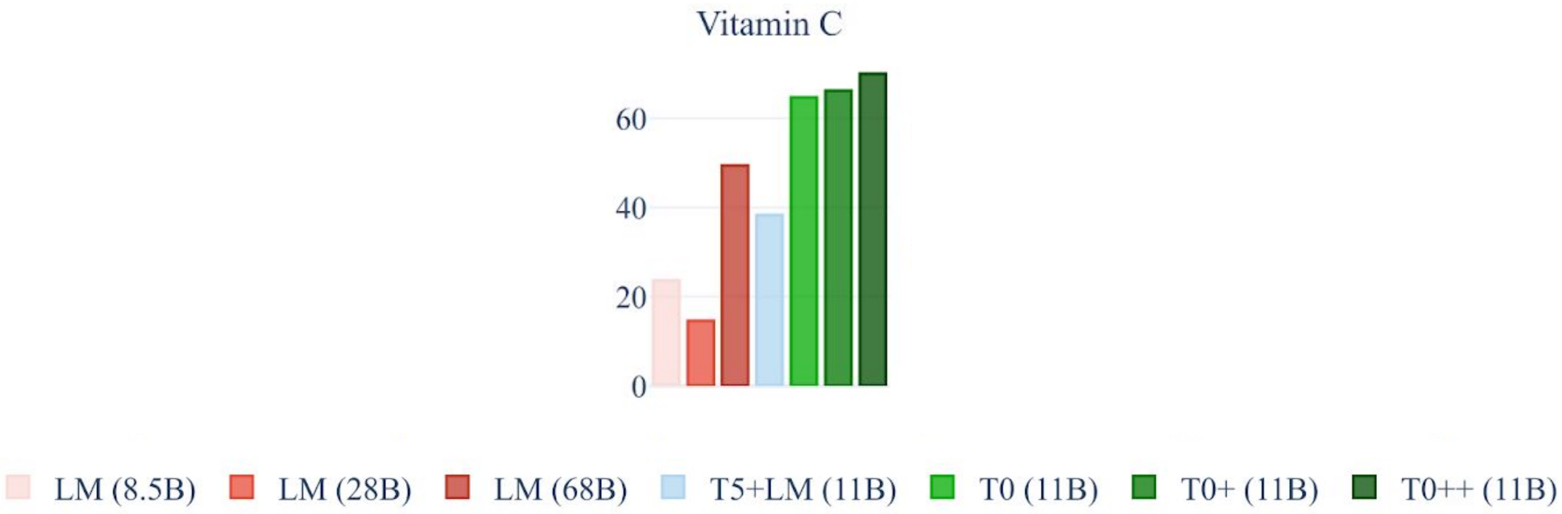
- Evaluation data set meant to test very different tasks

- Comparison with 3 Google LMs (8.5B, 28B, 68B)

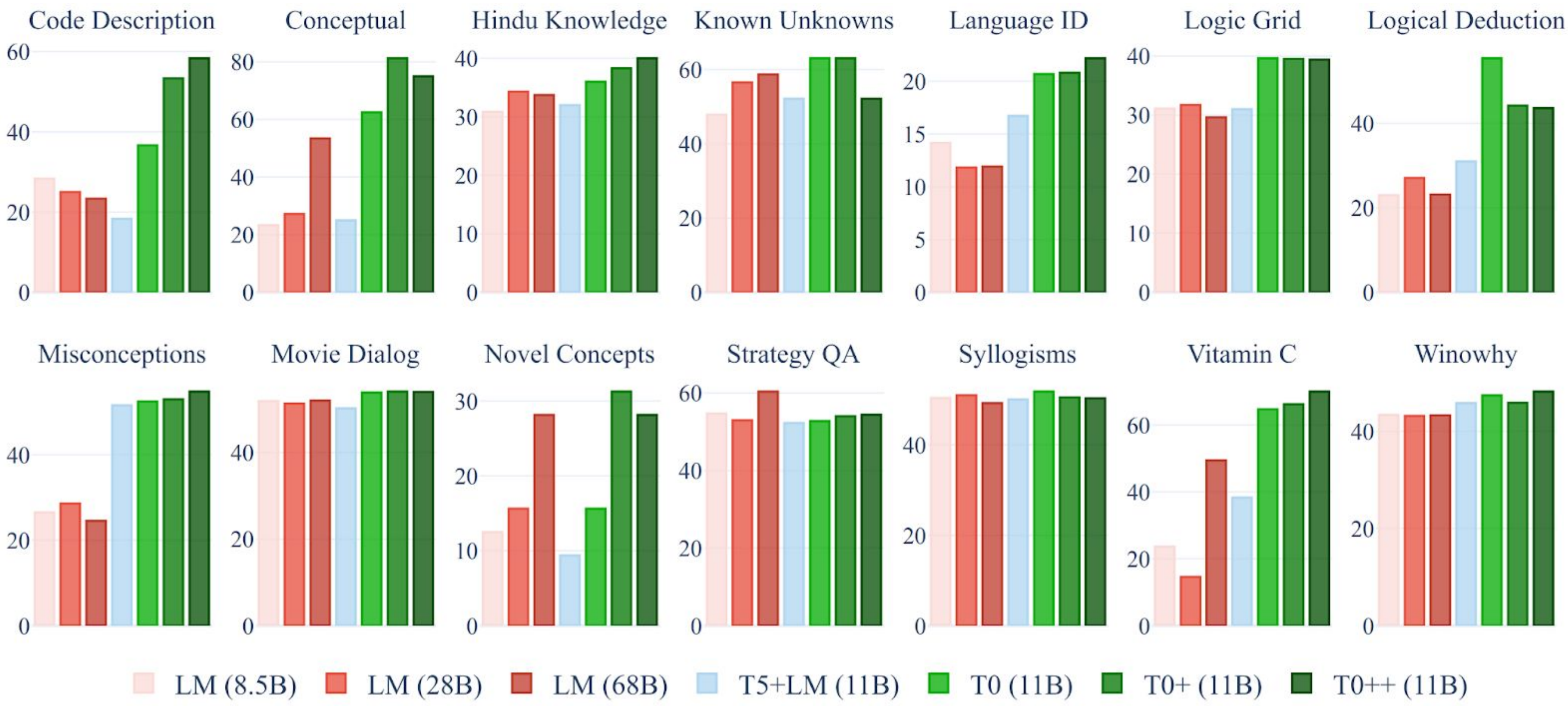
- Three versions of T0 11B trained with different tasks.

BIG-Bench

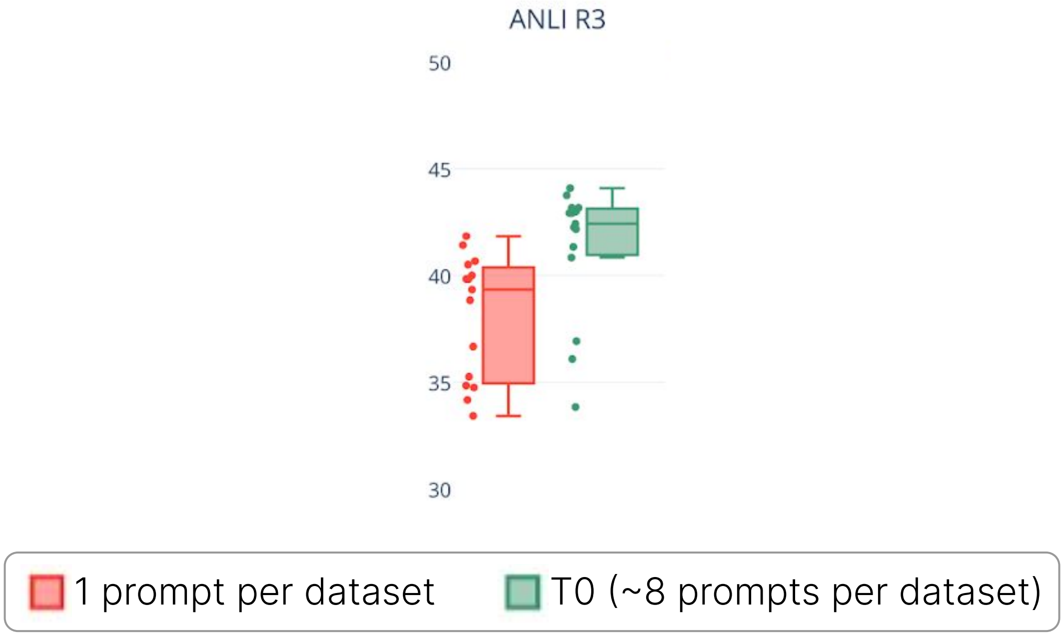
Based only on the information contained in a brief quote from Wikipedia, answer whether the related claim is True, False or Neither. Use Neither when the Wikipedia quote does not provide the necessary information to resolve the question. Input: {claim} ....



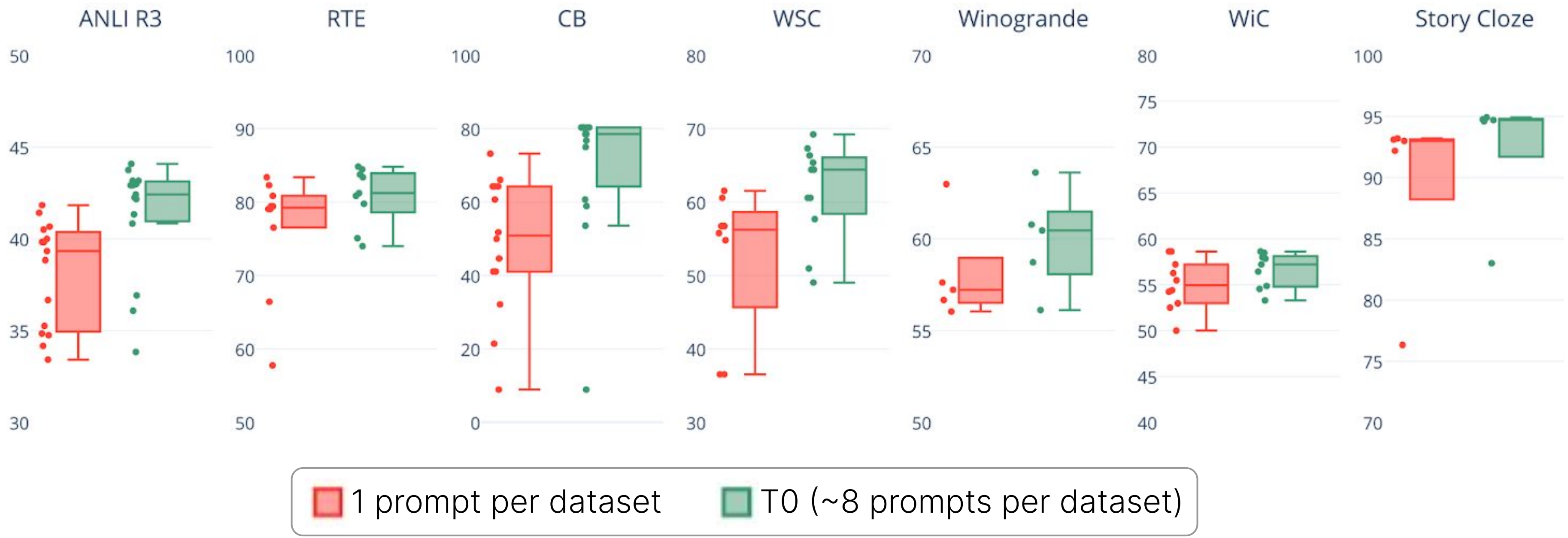
Performance on BIG-Bench subset



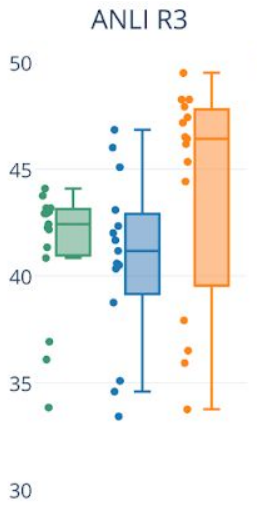
More prompts are better than one



More prompts are better than one

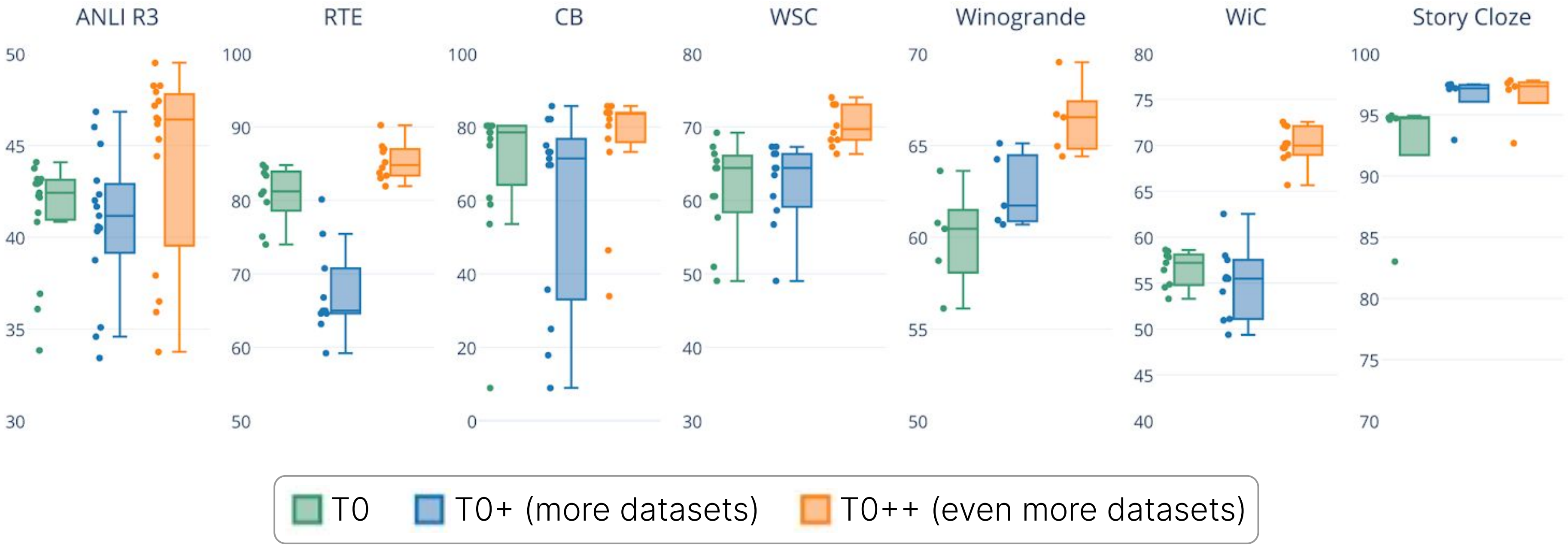


Adding datasets (usually) helps



T0 | T0+ (more datasets) | T0++ (even more datasets)

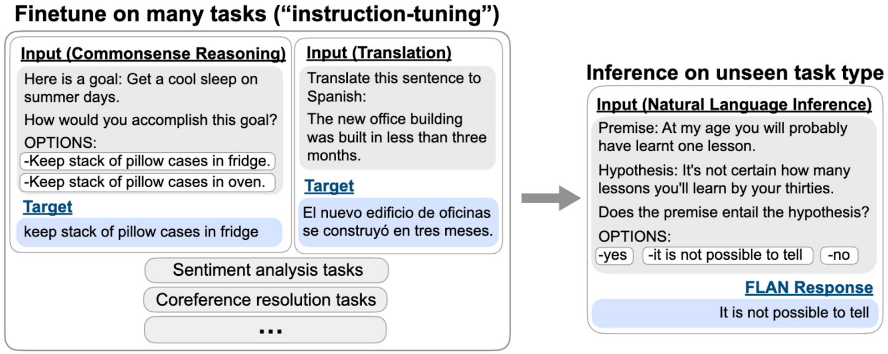
Adding datasets (usually) helps

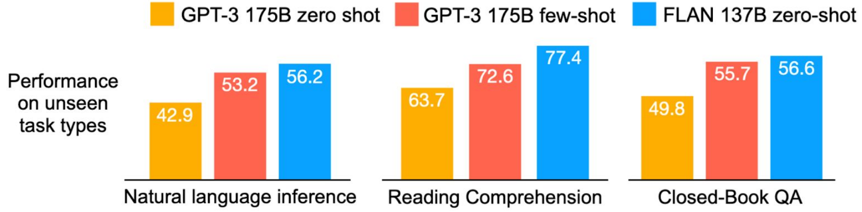


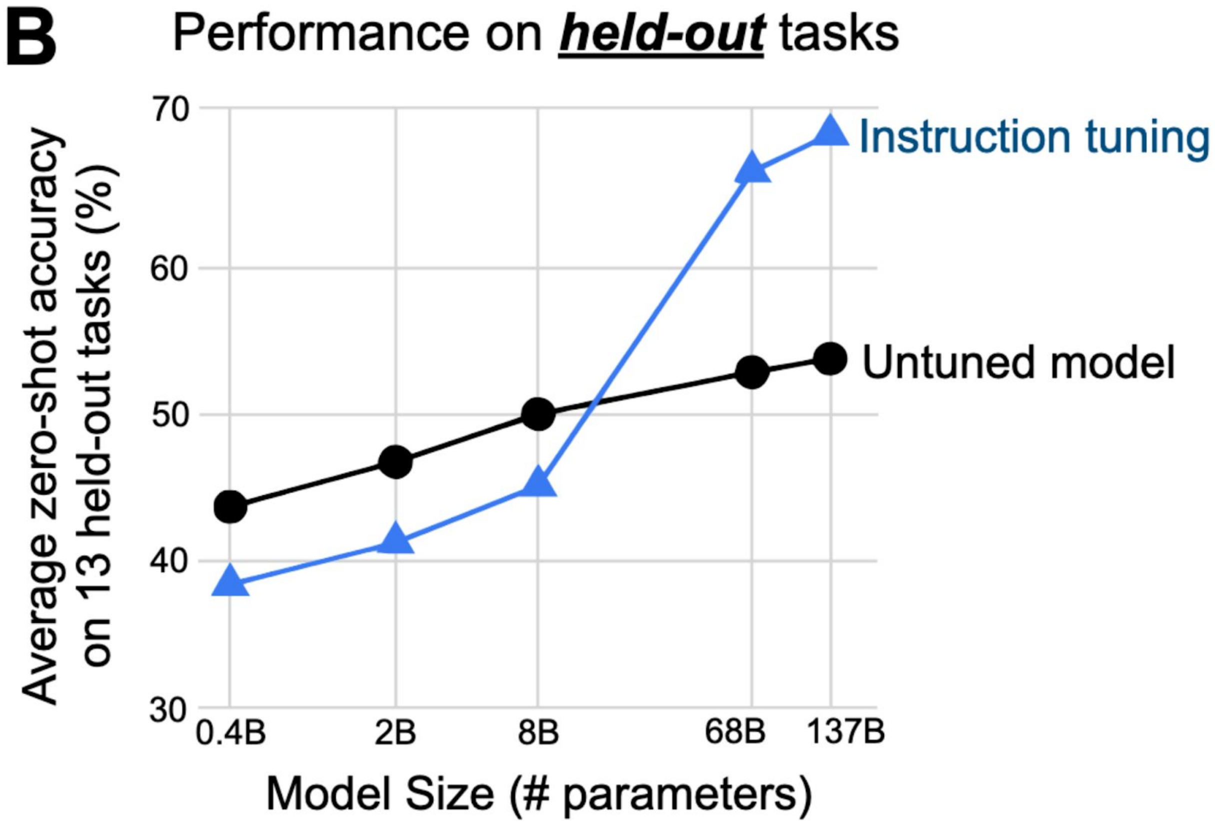
FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

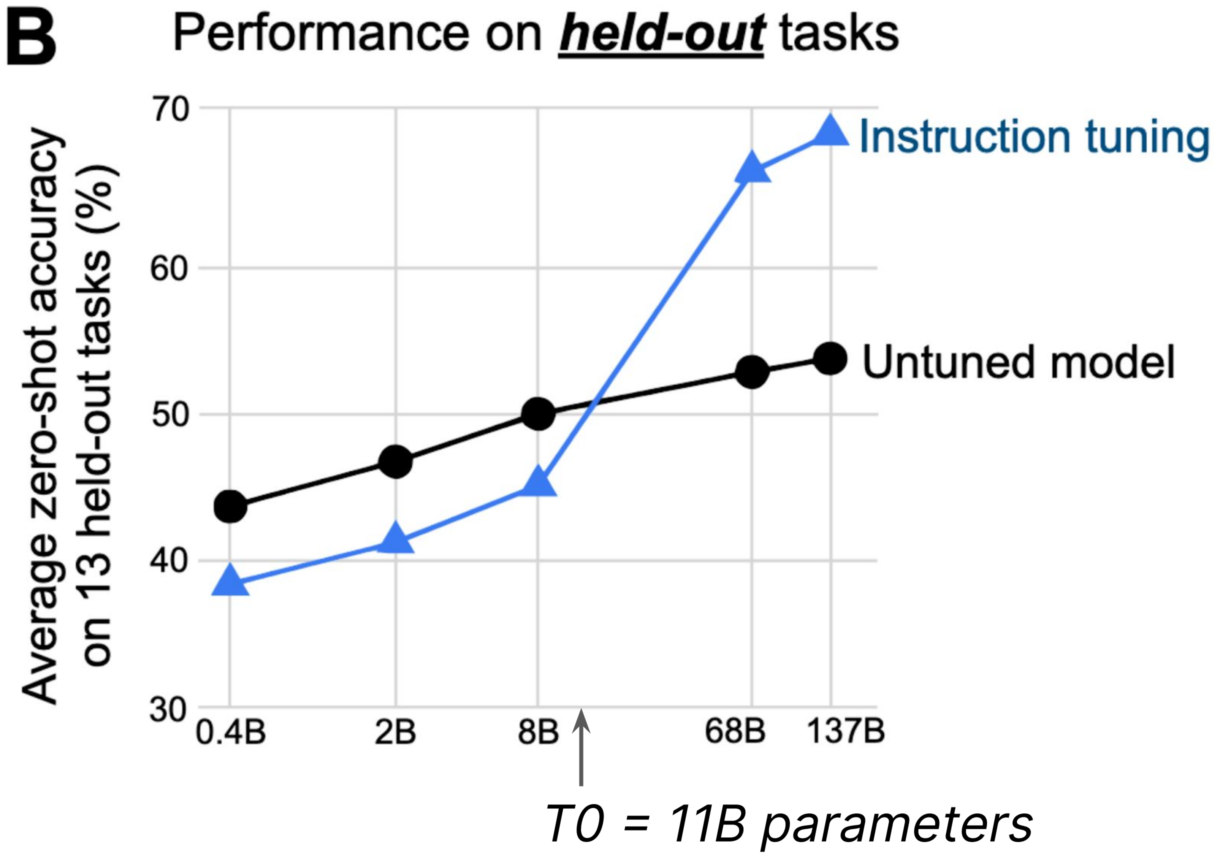
Jason Wei* **Maarten Bosma*** **Vincent Y. Zhao*** **Kelvin Guu*** **Adams Wei Yu**
Brian Lester **Nan Du** **Andrew M. Dai** **Quoc V. Le**

Google Research





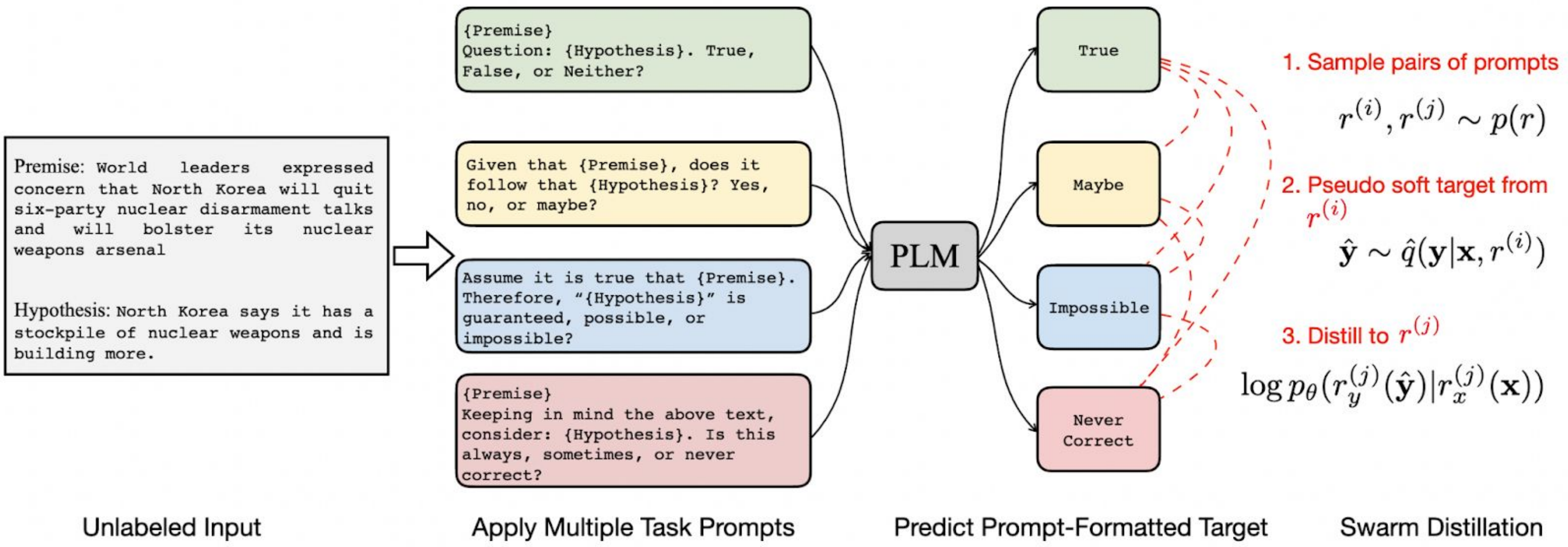

B
Performance on held-out tasks
Average zero-shot accuracy on 13 held-out tasks (%)
70
60
50
40
30
Instruction tuning
Untuned model
0.4B
2B
8B
68B
137B
Model Size (# parameters)


B
Performance on held-out tasks
Average zero-shot accuracy on 13 held-out tasks (%)
70
60
50
40
30
0.4B
2B
8B
68B
137B
Instruction tuning
Untuned model
T0 = 11B parameters

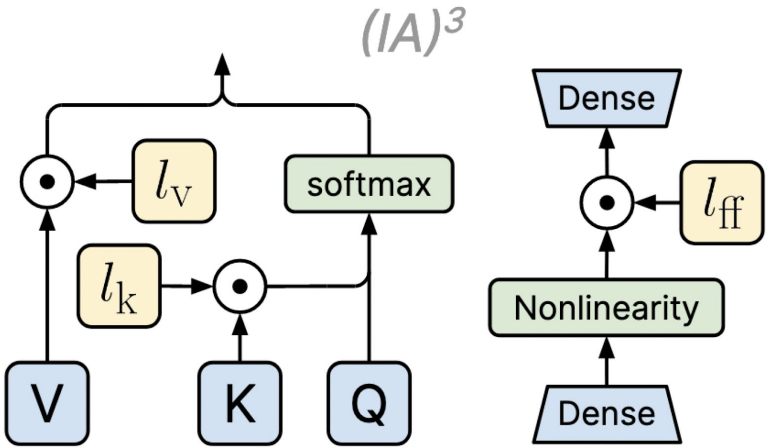
Caveats

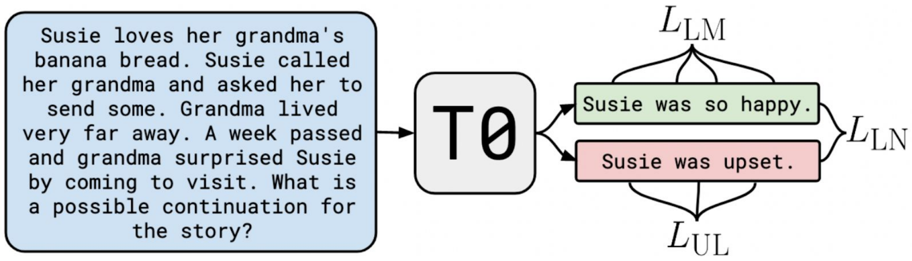
- Task accuracy is dependent on the prompt format / wording
- For each of these tasks numbers are low in an absolute sense (zero-shot)
- Approach does not extend automatically to in-context learning (Natural instructions Wang et al. 2022)
- No evidence (in this work) of prompt understanding in a complex sense

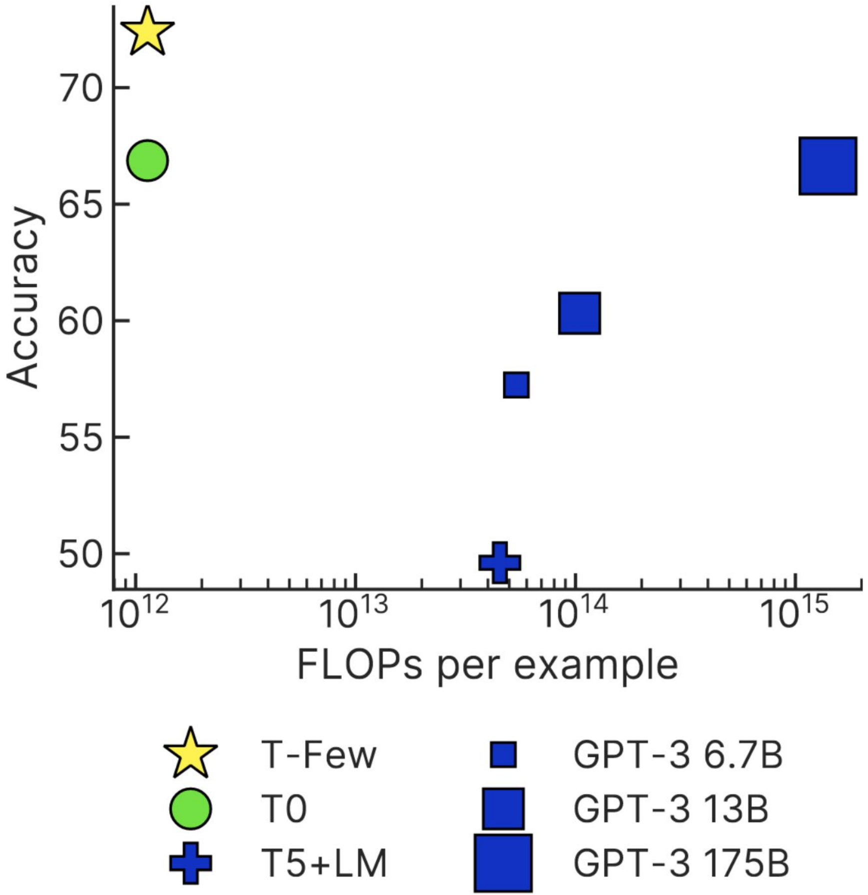
Usage



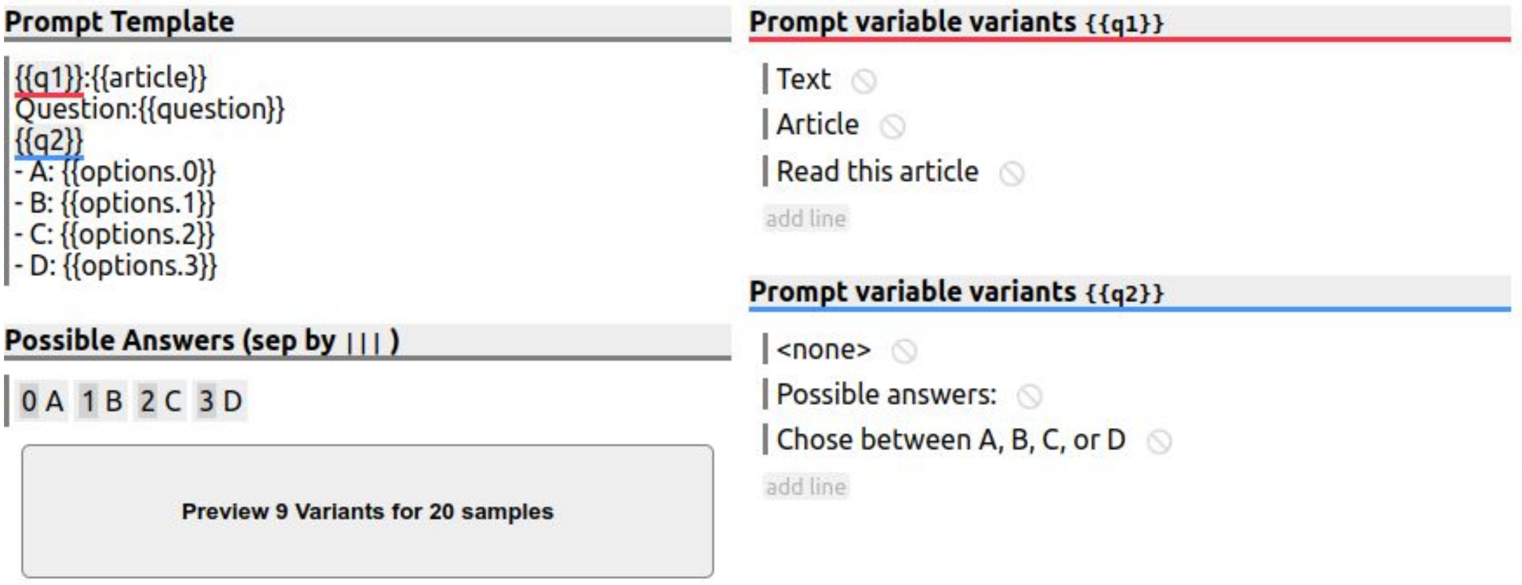
from "Prompt Consistency for Zero-Shot Task Generalization", Zhou et al. 2022







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



http://prompt.vizhub.ai/

From PromptIDE , Strobelt et al. 2022

Epilogue:
BLOOM

Large-scale Public Compute

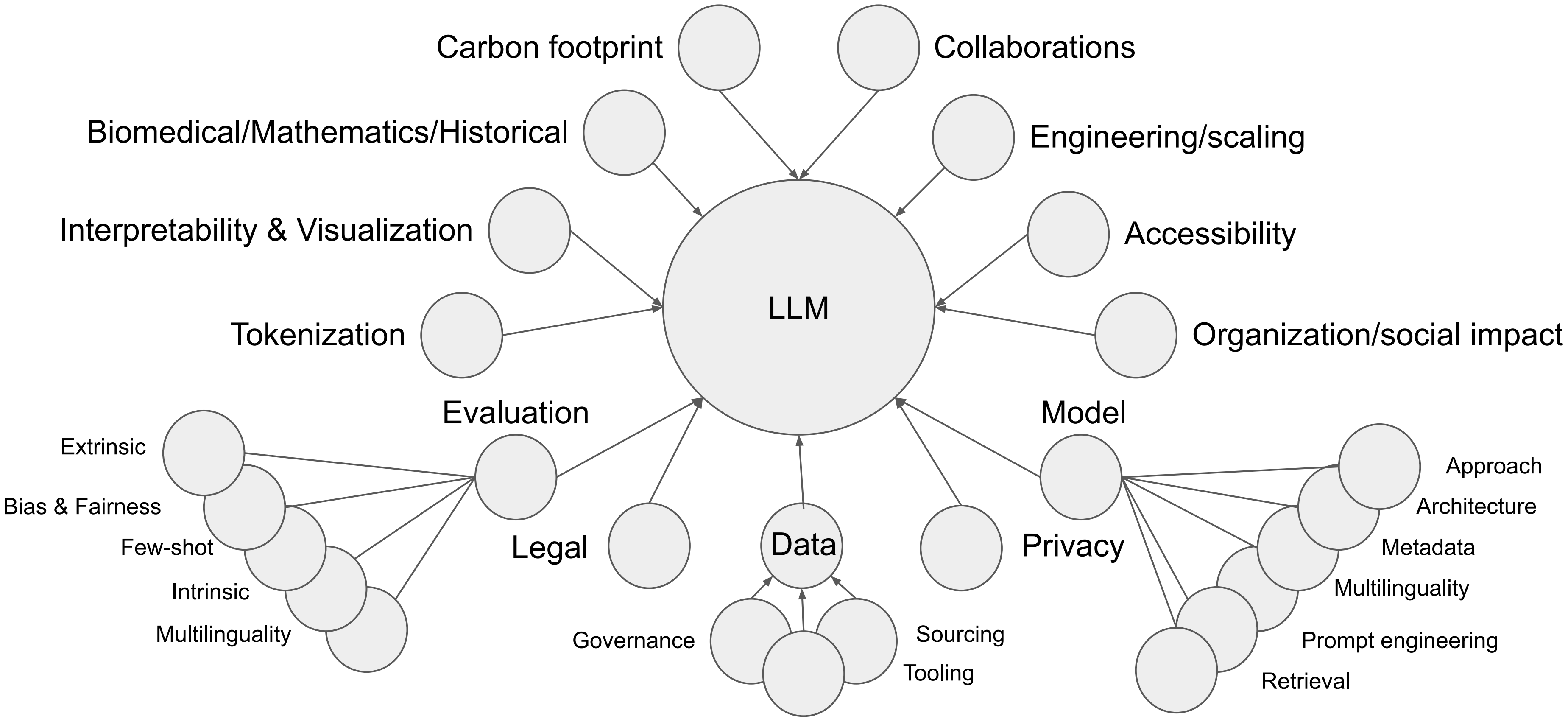
Jean Zay supercomputer at Orsay, France.

Accelerated partition (or GPU partition)

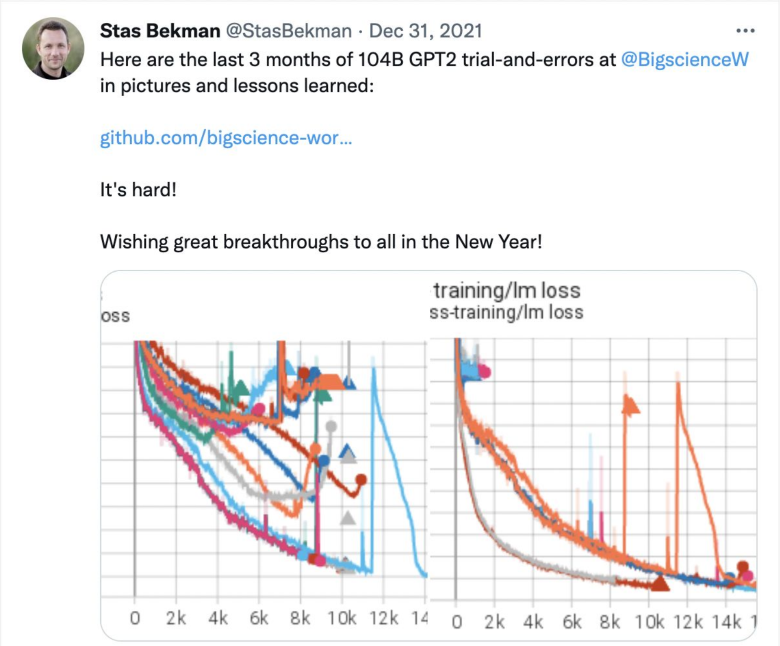
- 261 four-GPU accelerated compute nodes with:
  - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
  - 192 GB of memory per node
  - 4 Nvidia Tesla V100 SXM2 GPUs (32 GB)
- 31 eight-GPU accelerated compute nodes, currently dedicated to the AI community with:
  - 2 Intel Cascade Lake 6226 processors (12 cores at 2.7 GHz), namely 24 cores per node
  - 20 nodes with 384 GB of memory and 11 nodes with 768 GB of memory
  - 8 Nvidia Tesla V100 SXM2 GPUs (32 GB)
- Extension in the summer of 2020, 351 four-GPU accelerated compute nodes with:
  - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
  - 192 GB of memory per node
  - 4 Nvidia Tesla V100 SXM2 GPUs (16 GB)

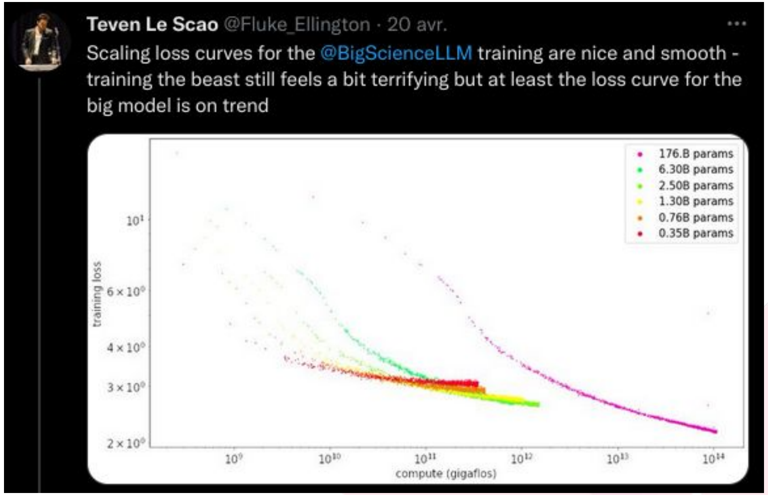
- Cumulated peak performance of 28 Pflop/s with a total of **2696 Nvidia V100 GPUs**
- JZ 3 expands to 3,152 GPUs (V100s and A100s) - use time: 3 months
- Omni-PAth interconnection network 100 Gb/s : 4 links per converged node
- Parallel storage w/capacity of 2.2 PB SSD disks (GridScaler GS18K SSD)

How to Train a Language Model



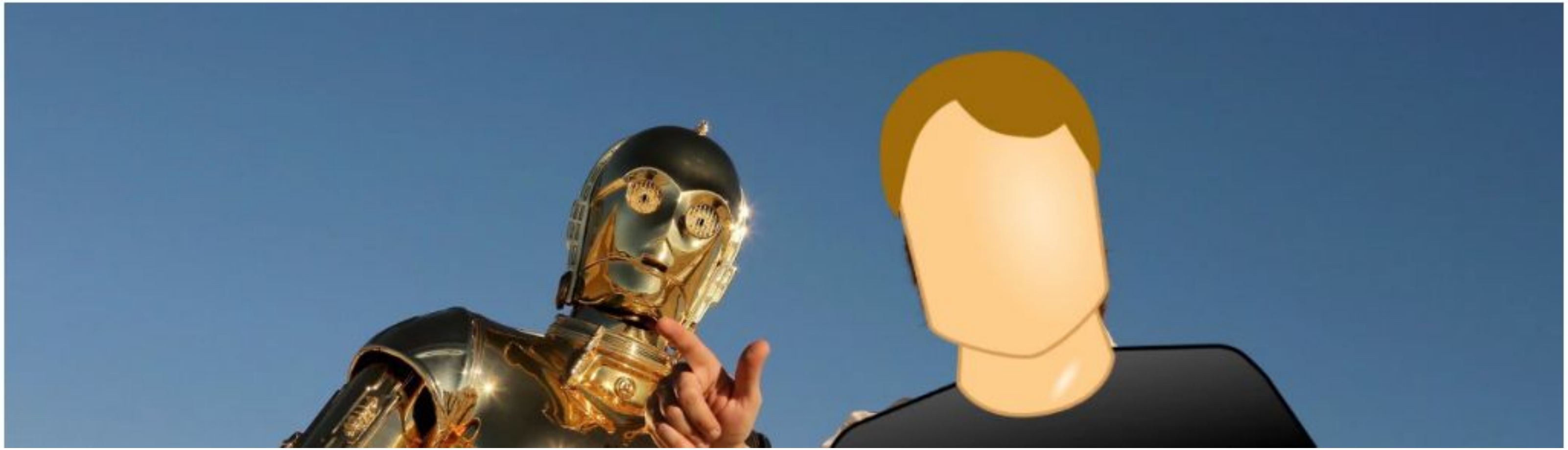
Open Reporting



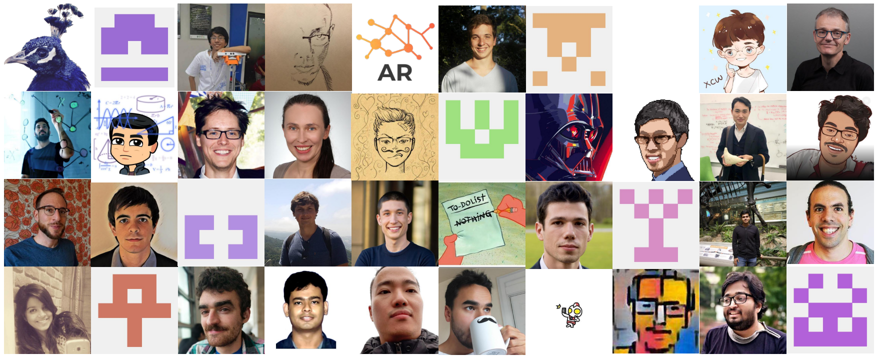


Release: BLOOM

Ya puedes usar BLOOM, una IA de código abierto más potente que GPT-3 que es capaz de generar texto en 59 lenguajes



BigScience



https://github.com/bigscience-workshop/t-zero
https://huggingface.co/bigscience/T0{p,pp,_3B}