A better way to get language models to do what you ask

Colin Raffel

A better way to get language models to do what you ask

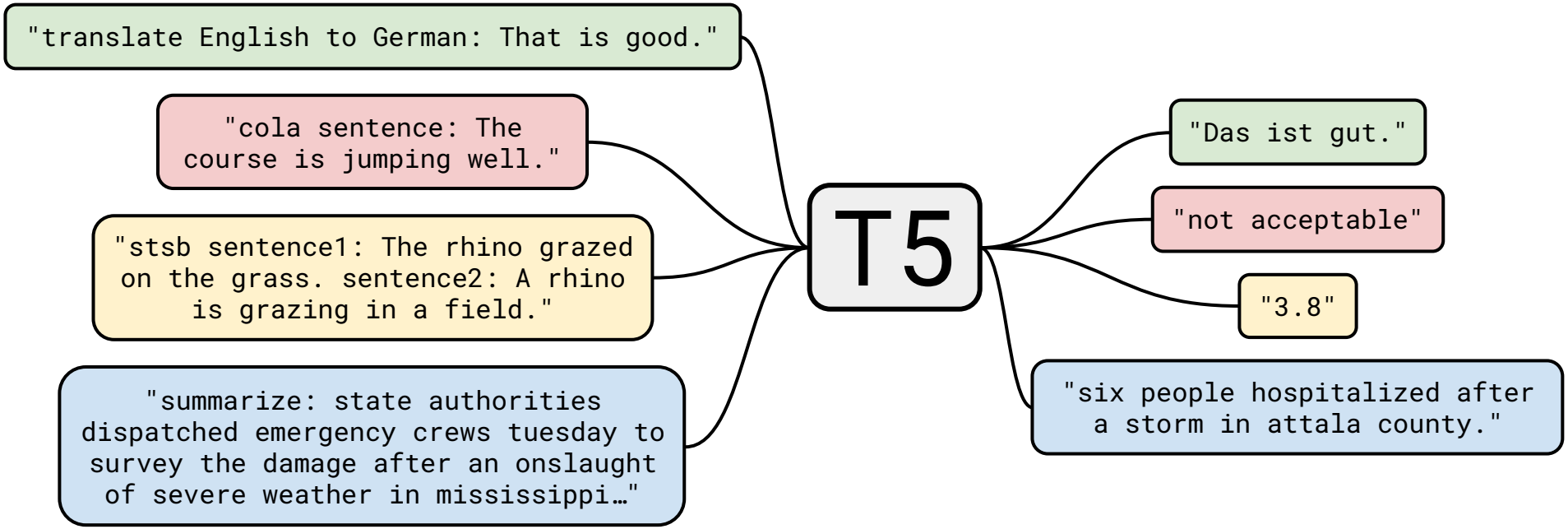


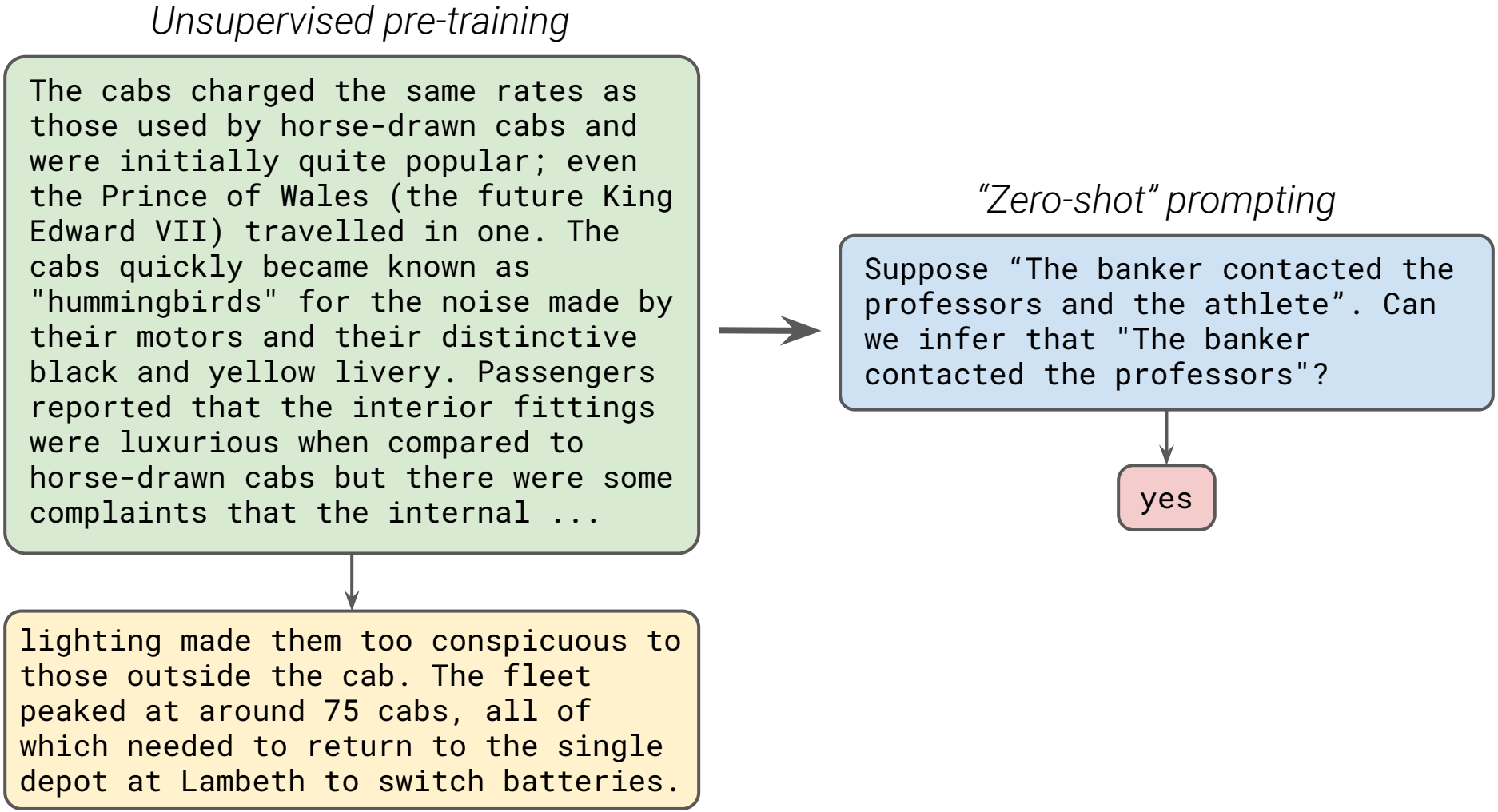

... and many other awesome collaborators

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 ___.

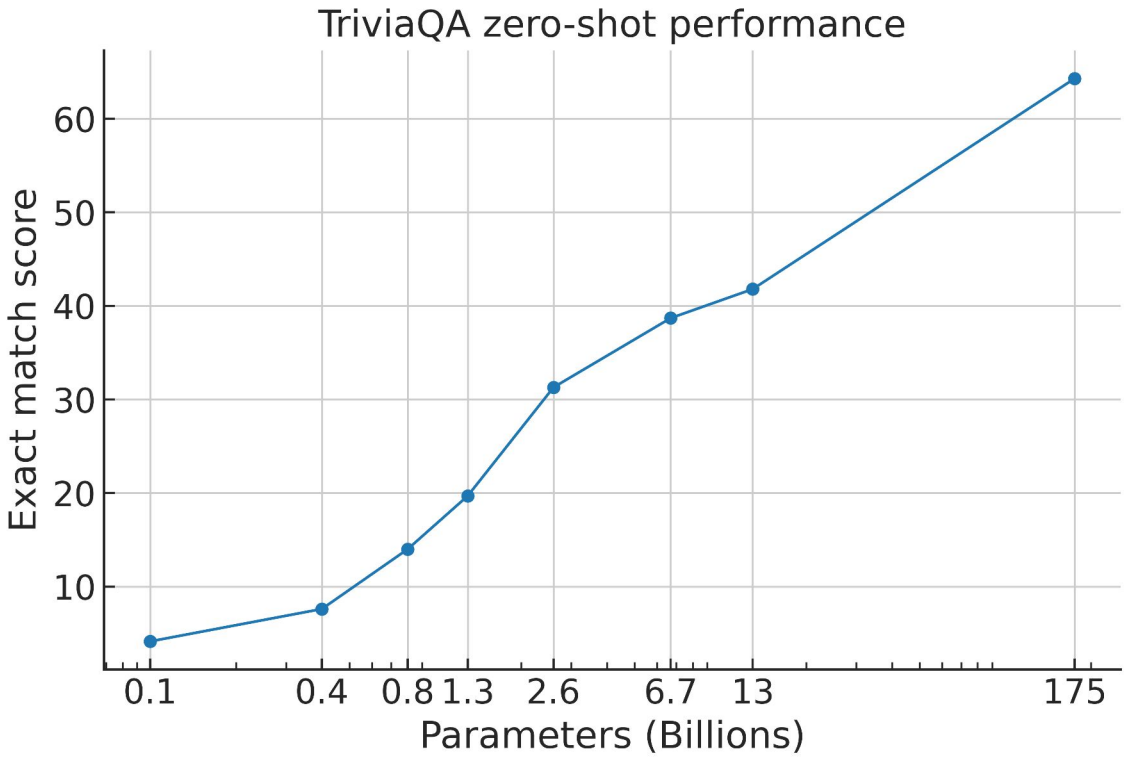
T5

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


"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."


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

Language Models are Few-Shot Learners



1. In what year was the first-ever Wimbledon Championship held? **Answer: 1877.**

2. Hg is the chemical symbol of which element? **Answer: Mercury.**

3. Which email service is owned by Microsoft? **Answer: Hotmail.**

4. Which country produces the most coffee in the world? **Answer: Brazil.**

5. In which city was Jim Morrison buried? **Answer: Paris.**

6. Which song by Luis Fonsi and Daddy Yankee has the most views (of all time) on YouTube? **Answer: "Despacito."**

7. What was the first state? **Answer: Delaware.**

8. What is the capital city of Spain? **Answer: Madrid.**

9. What is the painting "La Gioconda" more usually known as? **Answer: The Mona Lisa.**

from https://www.scarymommy.com/best-trivia-questions-answers/

UNIFIEDQA: Crossing Format Boundaries with a Single QA System

Daniel Khashabi[1] **Sewon Min**[2] **Tushar Khot**[1] **Ashish Sabharwal**[1]
Oyvind Tafjord[1] **Peter Clark**[1] **Hannaneh Hajishirzi**[1,2]

[1]Allen Institute for AI, Seattle, U.S.A.
[2]University of Washington, Seattle, U.S.A.

Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

Measuring Massive Multitask Language Understanding

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23
Answer: B

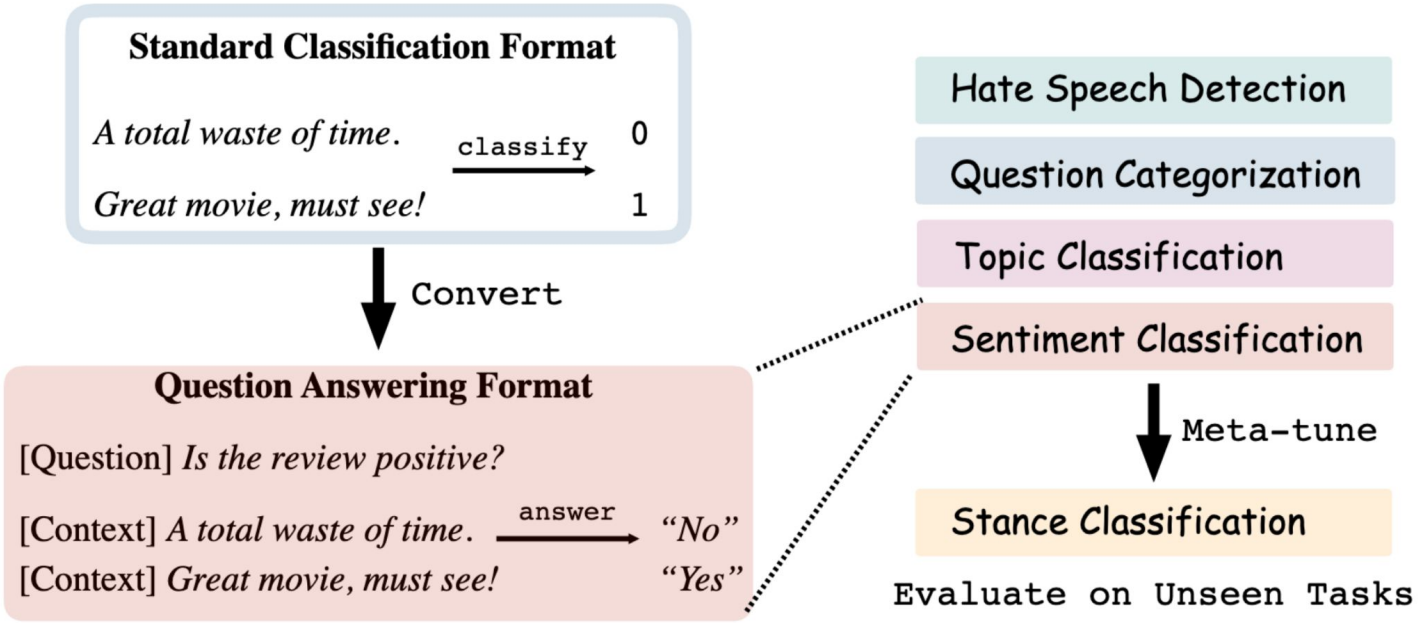
Compute $i + i^2 + i^3 + \cdots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) i (D) $-i$
Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
(A) 28 (B) 21 (C) 40 (D) 30
Answer: C

Model	Humanities	Social Science	STEM	Other	Average
Random Baseline	25.0	25.0	25.0	25.0	25.0
RoBERTa	27.9	28.8	27.0	27.7	27.9
ALBERT	27.2	25.7	27.7	27.9	27.1
GPT-2	32.8	33.3	30.2	33.1	32.4
UnifiedQA	45.6	56.6	40.2	54.6	48.9
GPT-3 Small (few-shot)	24.4	30.9	26.0	24.1	25.9
GPT-3 Medium (few-shot)	26.1	21.6	25.6	25.5	24.9
GPT-3 Large (few-shot)	27.1	25.6	24.3	26.5	26.0
GPT-3 X-Large (few-shot)	40.8	50.4	36.7	48.8	43.9

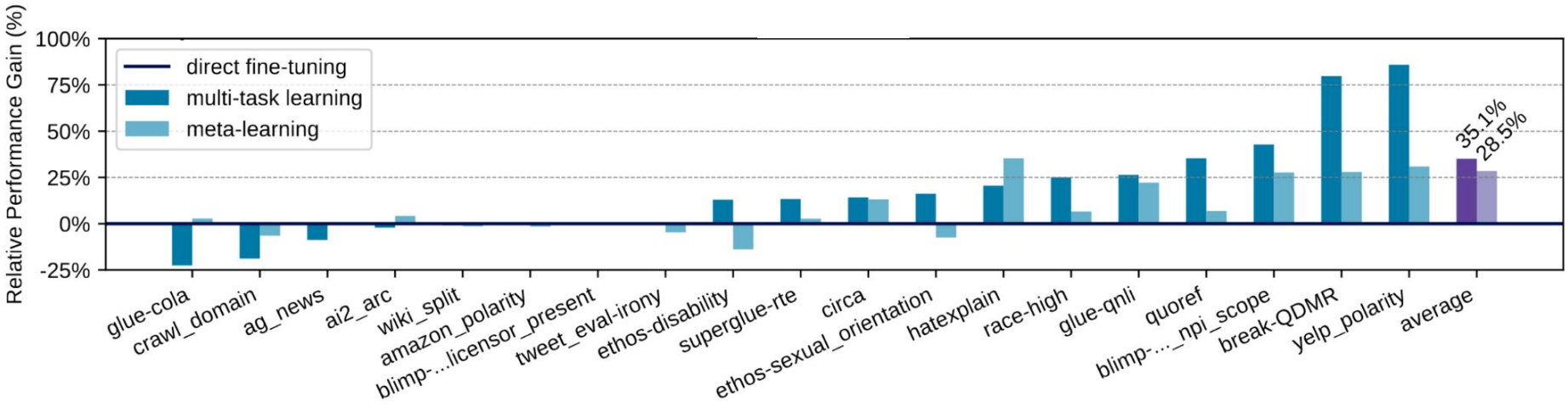
Meta-tuning Language Models to Answer Prompts Better

Ruiqi Zhong **Kristy Lee*** **Zheng Zhang*** **Dan Klein**
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu

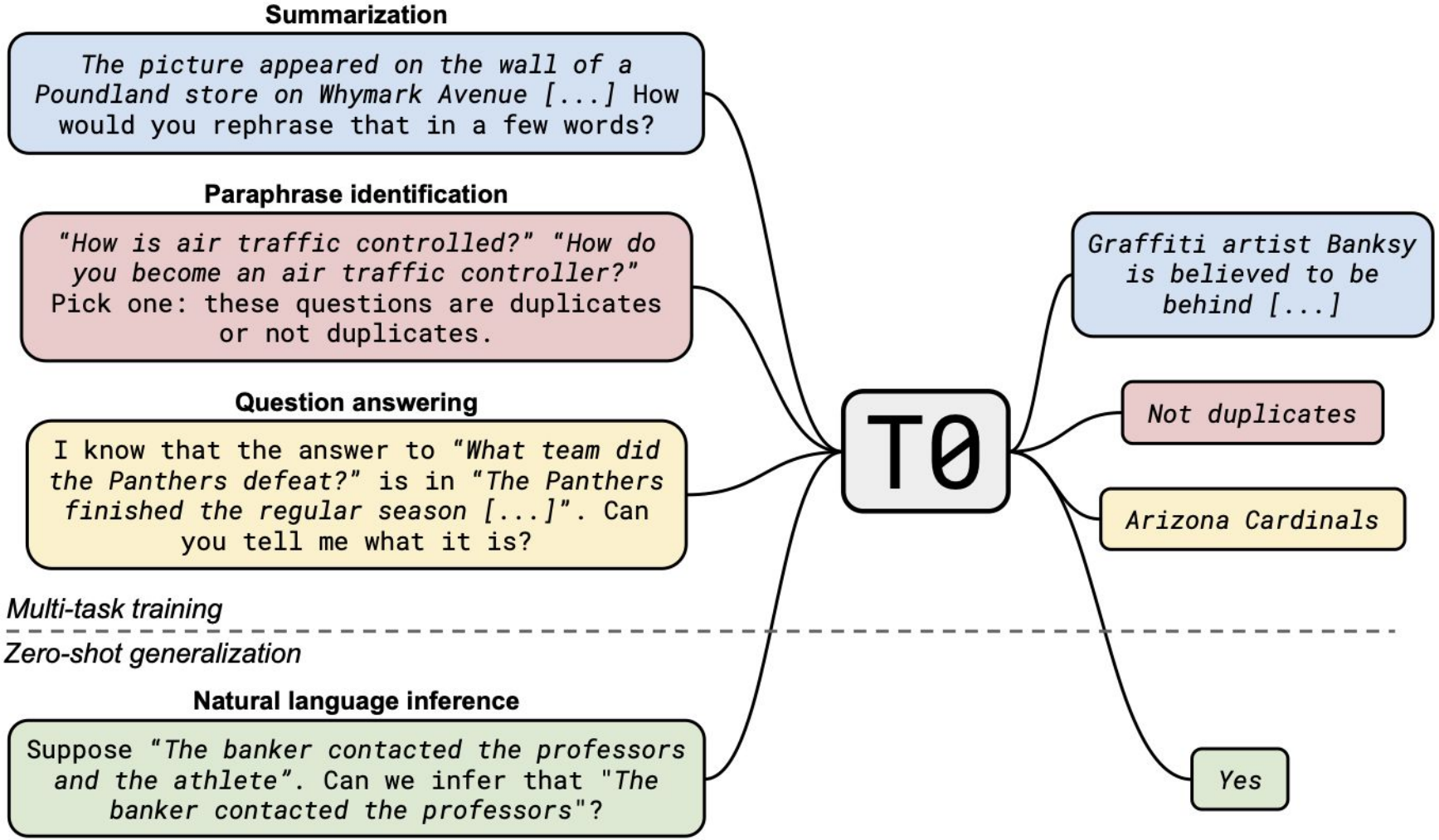


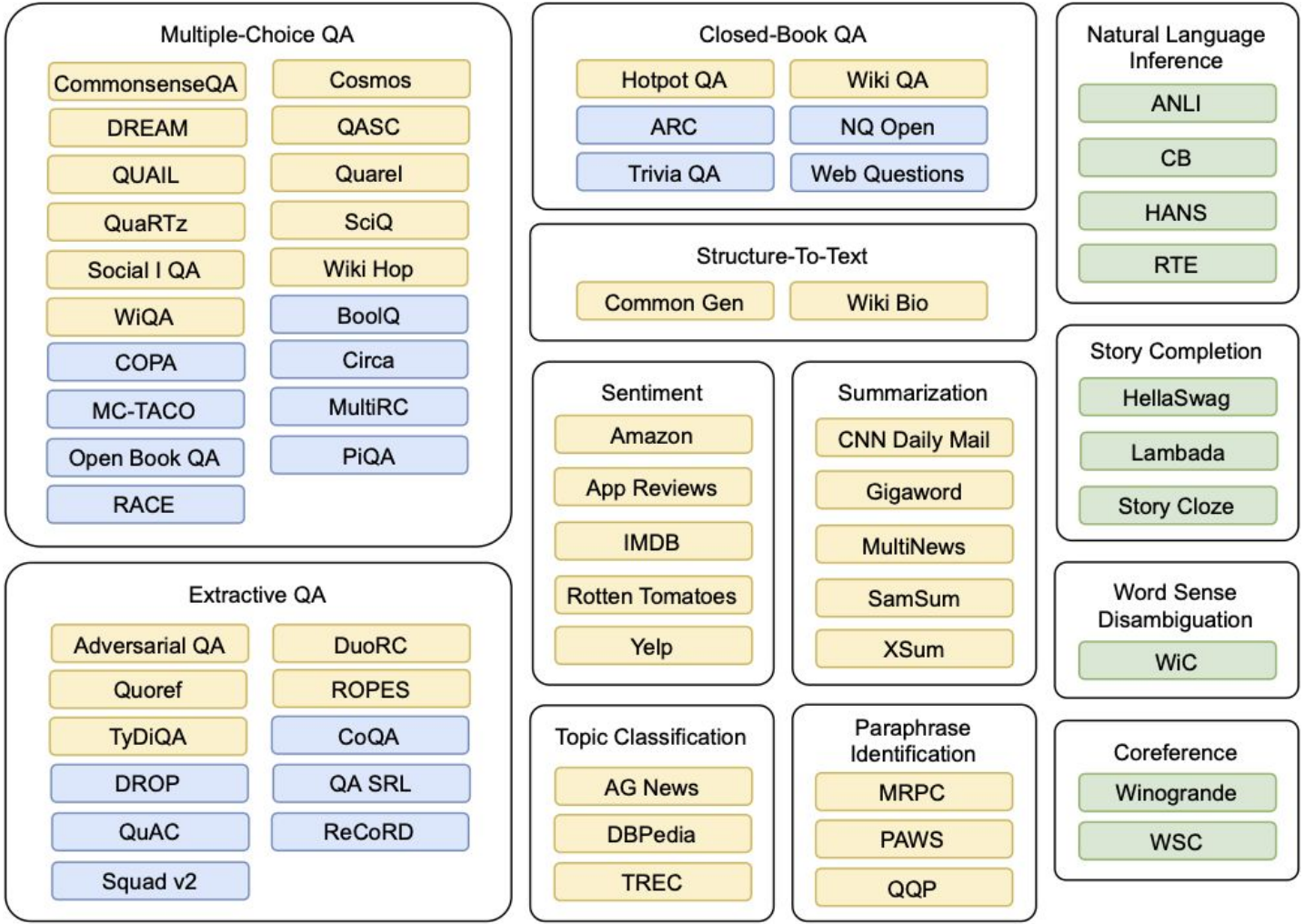
CROSSFIT: A Few-shot Learning Challenge for Cross-task Generalization in NLP

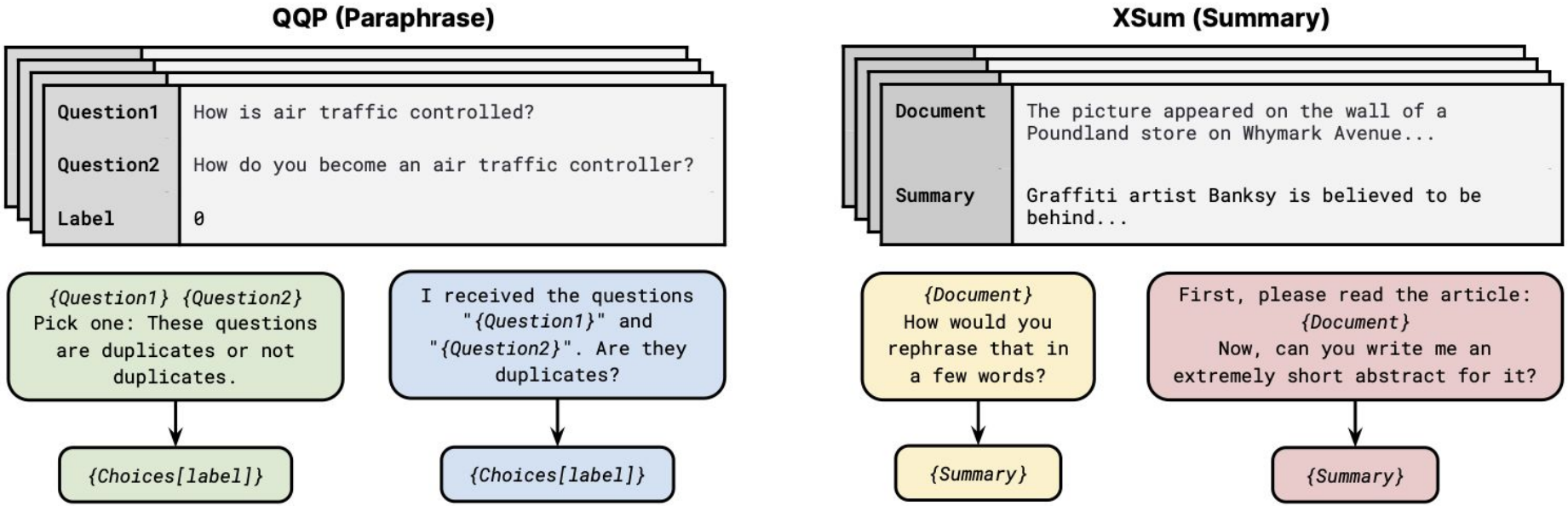
Qinyuan Ye **Bill Yuchen Lin** **Xiang Ren**
University of Southern California
{qinyuany, yuchen.lin, xiangren}@usc.edu



Can we obtain good zero-shot task generalization by training a large language model on a massively multitask mixture of diversely prompted datasets?


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


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


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.
I received the questions "{Question1}" and "{Question2}". Are they duplicates?
{Choices[label]}
{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?
First, please read the article: {Document} Now, can you write me an extremely short abstract for it?
{Summary}
{Summary}

Name

i_am_hesitating

Template Reference

Original Task?

Choices in Prompt?

Metrics

Accuracy

Answer Choices

{{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

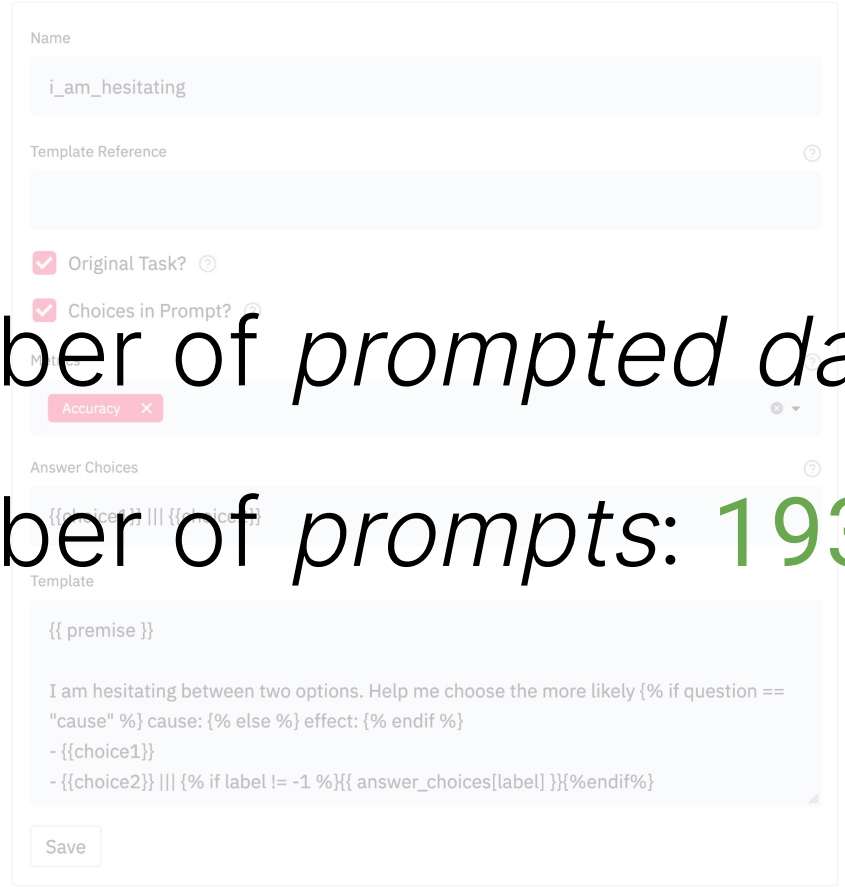
My body cast a shadow over the grass.

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

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

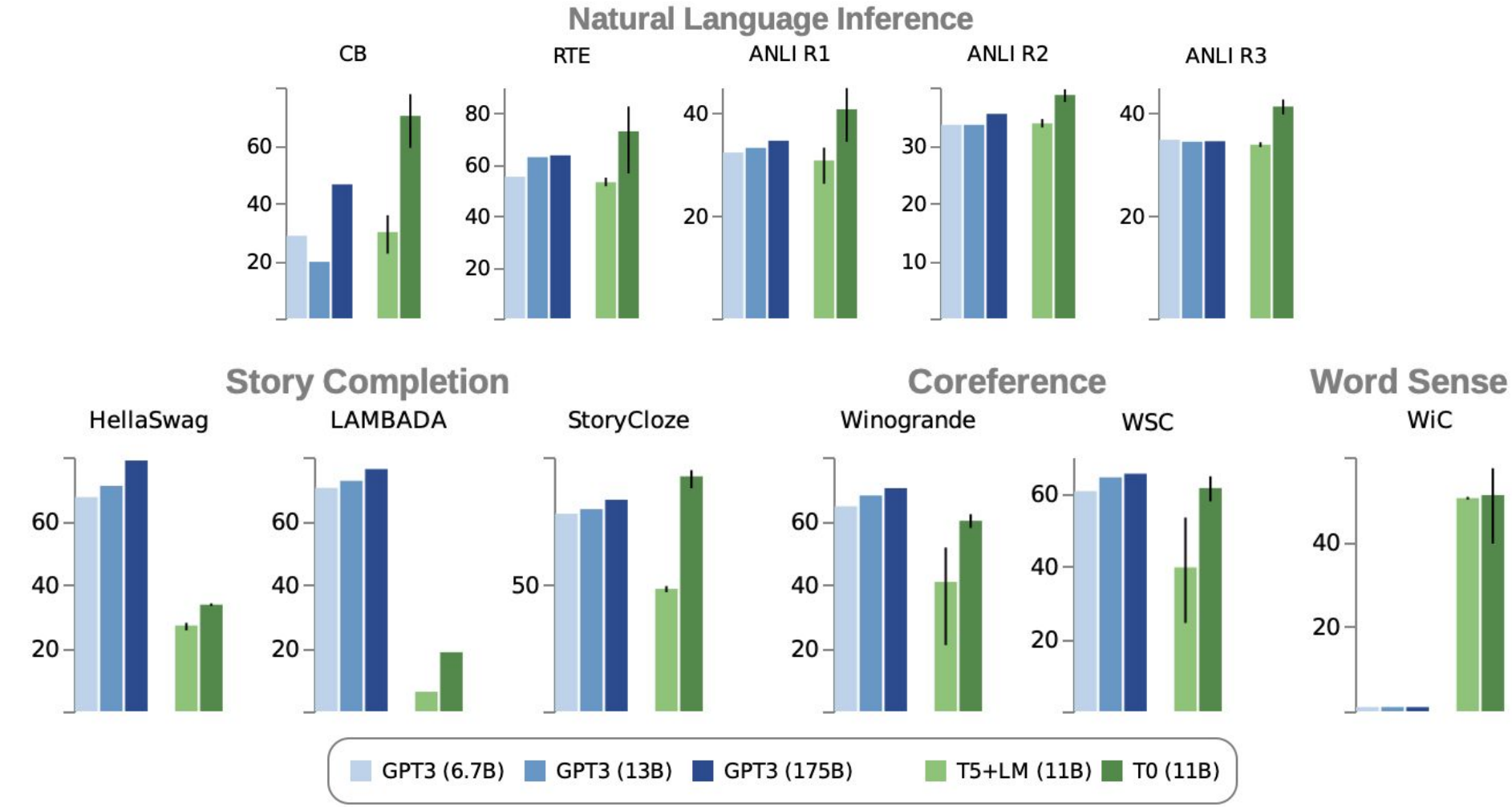
Y

The sun was rising.

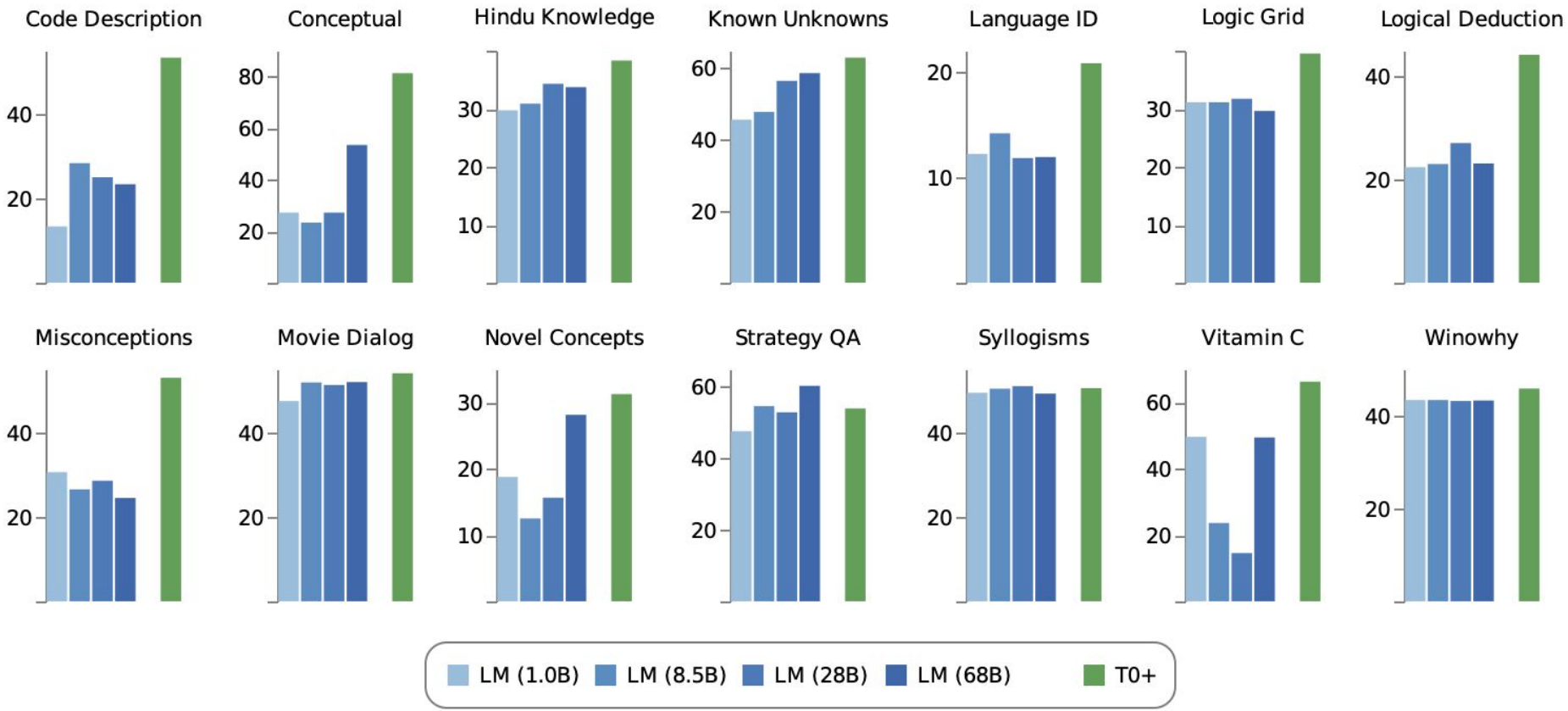


Number of *prompted datasets*: 170

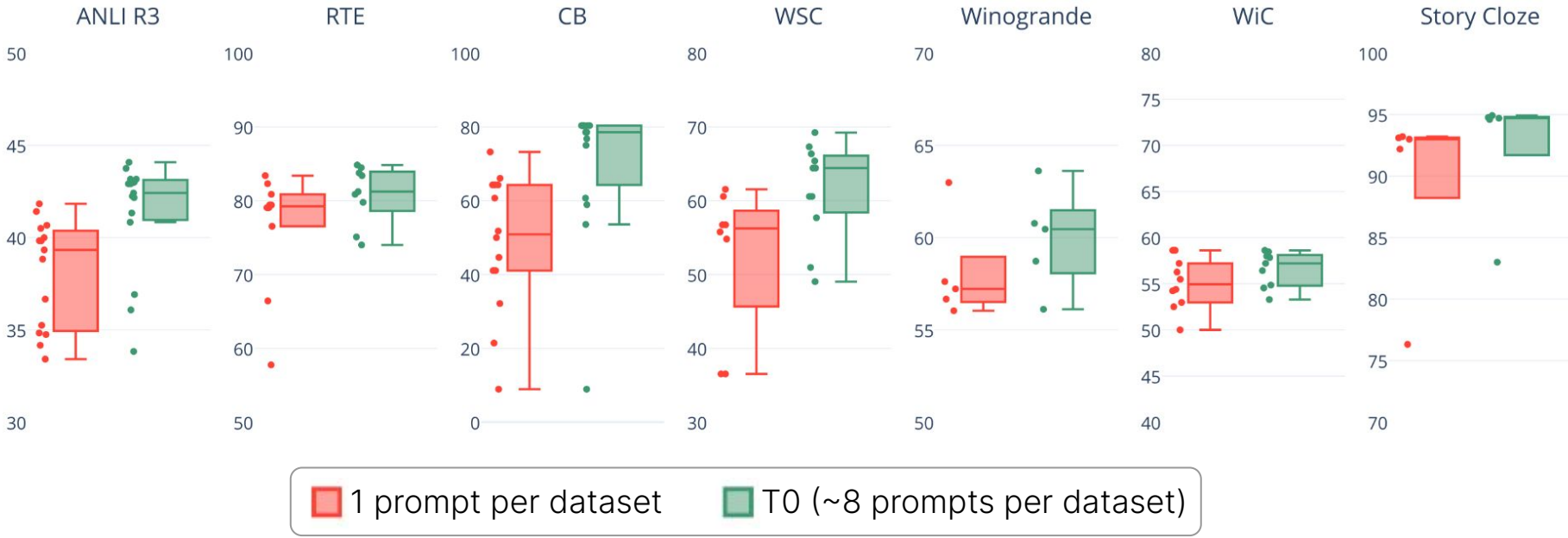
Number of *prompts*: 1939


Natural Language Inference
CB
RTE
ANLI R1
ANLI R2
ANLI R3
Story Completion
HellaSwag
LAMBADA
StoryCloze
Coreference
Winogrande
WSC
Word Sense
WiC
GPT3 (6.7B)
GPT3 (13B)
GPT3 (175B)
T5+LM (11B)
T0 (11B)

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
LM (1.0B)
LM (8.5B)
LM (28B)
LM (68B)
T0+

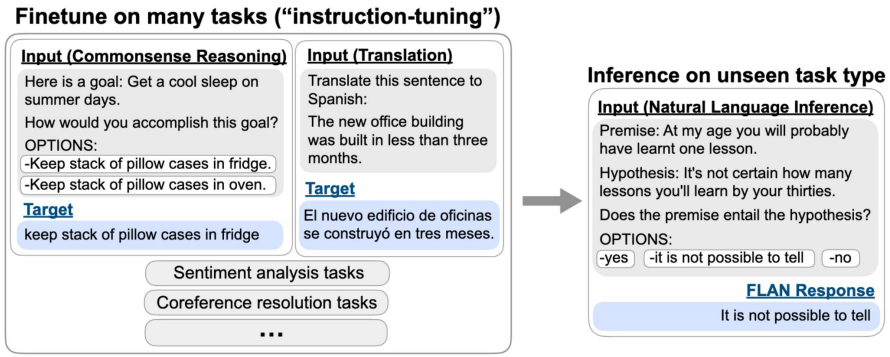
More prompts are better than one

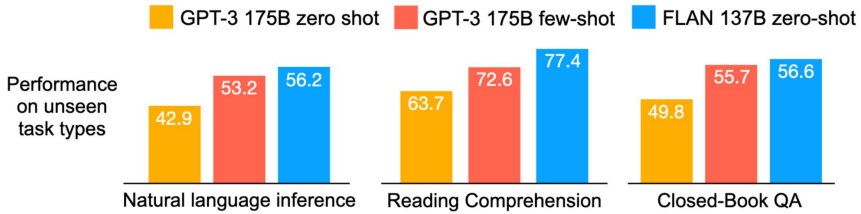


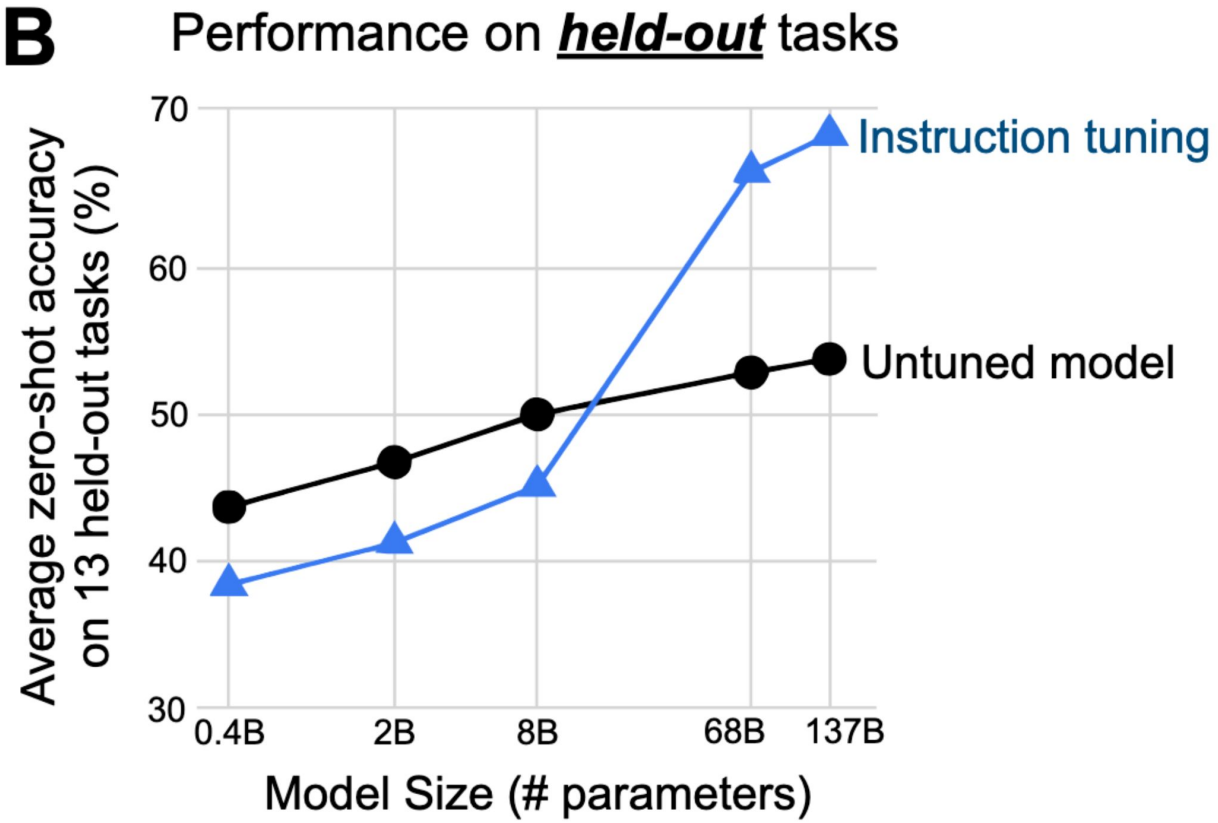
Adding datasets (usually) helps

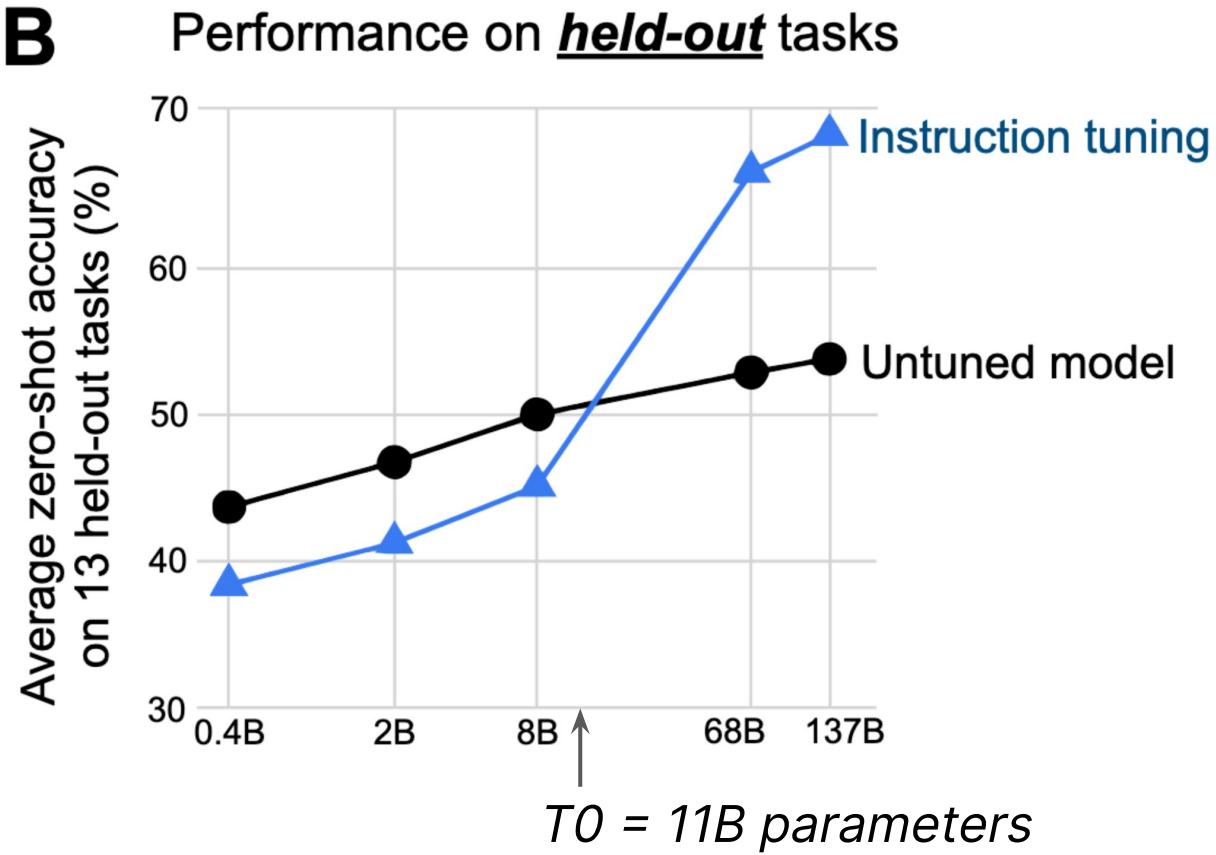
FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

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

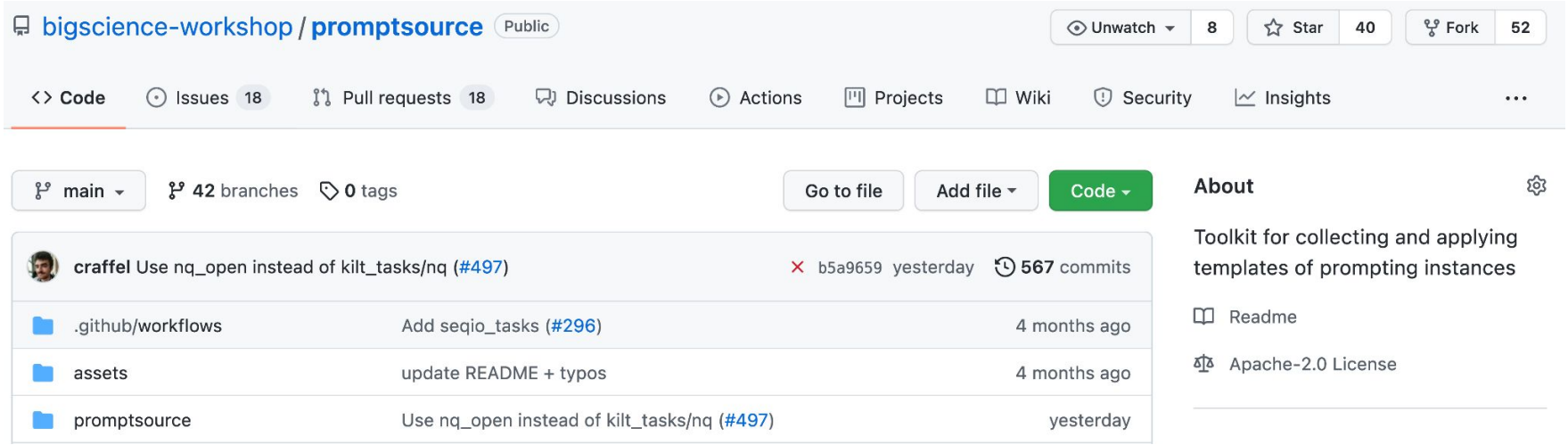




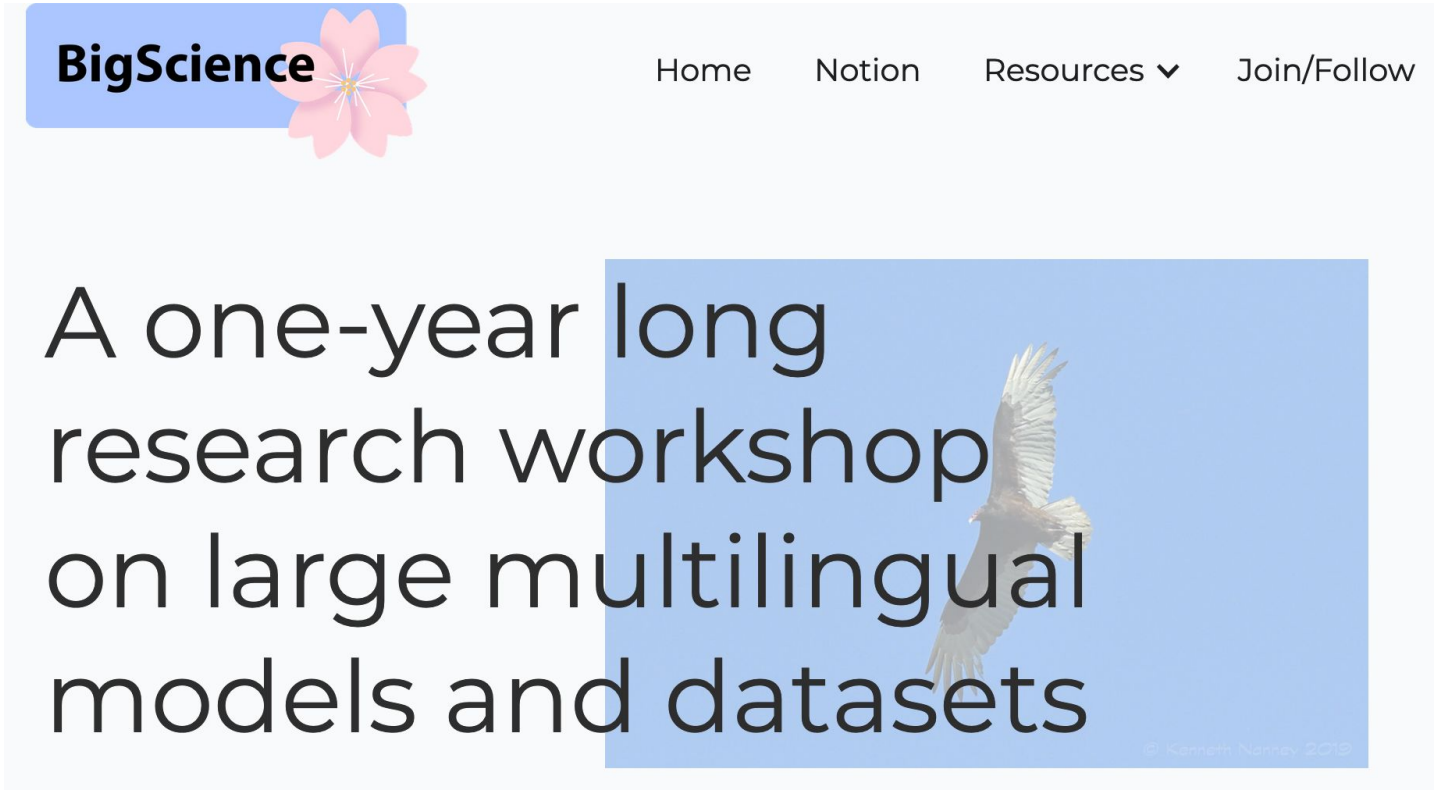

B
Performance on held-out tasks
Average zero-shot accuracy on 13 held-out tasks (%)
70
60
50
40
30
0.4B
2B
8B
68B
137B
Model Size (# parameters)
Instruction tuning
Untuned model


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

	GPT-3	**FLAN**	**T0** (ours)
Size	175B	137B	11B
Multitask supervision	Implicit	Explicit	Explicit
Zero-shot performance	Decent	Better	Better
Architecture	Decoder	Decoder	Encoder/decoder
Pre-training	LM	LM	MLM→LM
Multiple prompts?	No	Yes	Yes
Prompt diversity	N/A	Some	Lots
Public	No	No	Yes



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



https://bigscience.huggingface.co/

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Bach* Brown University

Lintang Sutawika BigScience
Zaid Alyafeai KFUPM
Antoine Chaffin IRISA, IMATAG
Arnaud Stiegler Hyperscience
Arun Raja A*STAR

Manan Dey SAP
M Saiful Bari NTU
Canwen Xu UCSD/HF
Urmish Thakker SambaNova Systems
Shanya Sharma Walmart Labs

Eliza Szczechla BigScience
Taewoon Kim VU Amsterdam
Gunjan Chhablani BigScience
Nihal V. Nayak Brown University

Debajyoti Datta University of Virginia
Jonathan Chang ASUS
Mike Tian-Jian Jiang ZEALS
Han Wang NYU
Matteo Manica IBM Research

Sheng Shen UC Berkeley
Zheng-Xin Yong Brown University
Harshit Pandey BigScience
Michael McKenna Parity
Rachel Bawden Inria, France

Thomas Wang Inria, France
Trishala Neeraj BigScience
Jos Rozen BigScience
Abheesht Sharma BITS Pilani, India
Andrea Santilli Sapienza

Thibault Fevry BigScience
Jason Alan Fries Stanford University
Ryan Teehan Charles River Analytics
Teven Le Scao Hugging Face

Stella Biderman EleutherAI
Leo Gao EleutherAI
Tali Bers Brown University
Thomas Wolf Hugging Face
Alexander M. Rush Hugging Face

Thanks!