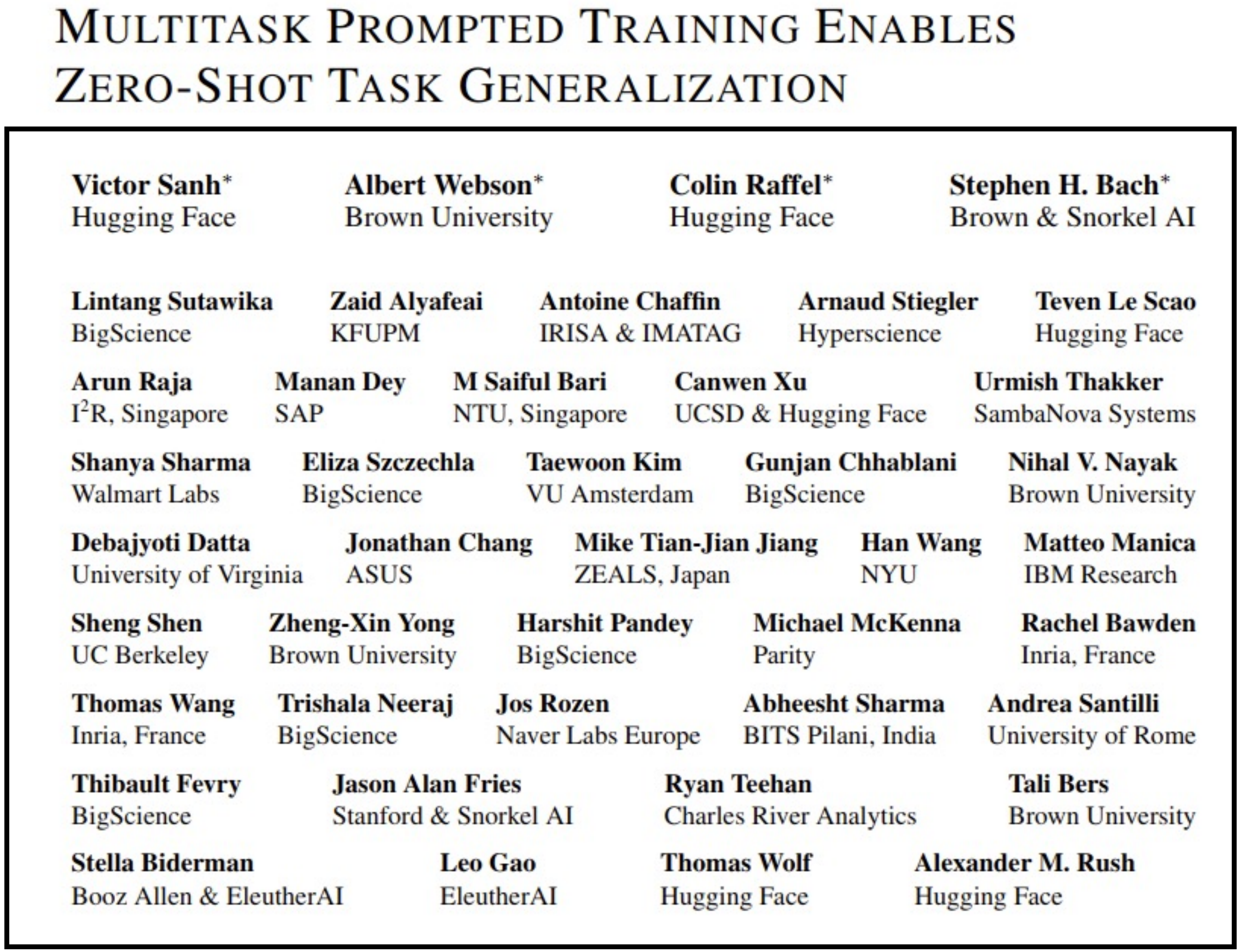


MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh*
Hugging Face

Albert Webson*
Brown University

Colin Raffel*
Hugging Face

Stephen H. Bach*
Brown & Snorkel AI

Lintang Sutawika
BigScience

Zaid Alyafeai
KFUPM

Antoine Chaffin
IRISA & IMATAG

Arnaud Stiegler
Hyperscience

Teven Le Scao
Hugging Face

Arun Raja
I^{2}R, Singapore

Manan Dey
SAP

M Saiful Bari
NTU, Singapore

Canwen Xu
UCSD & Hugging Face

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, Japan

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
Naver Labs Europe

Abheesht Sharma
BITS Pilani, India

Andrea Santilli
University of Rome

Thibault Fevry
BigScience

Jason Alan Fries
Stanford & Snorkel AI

Ryan Teehan
Charles River Analytics

Tali Bers
Brown University

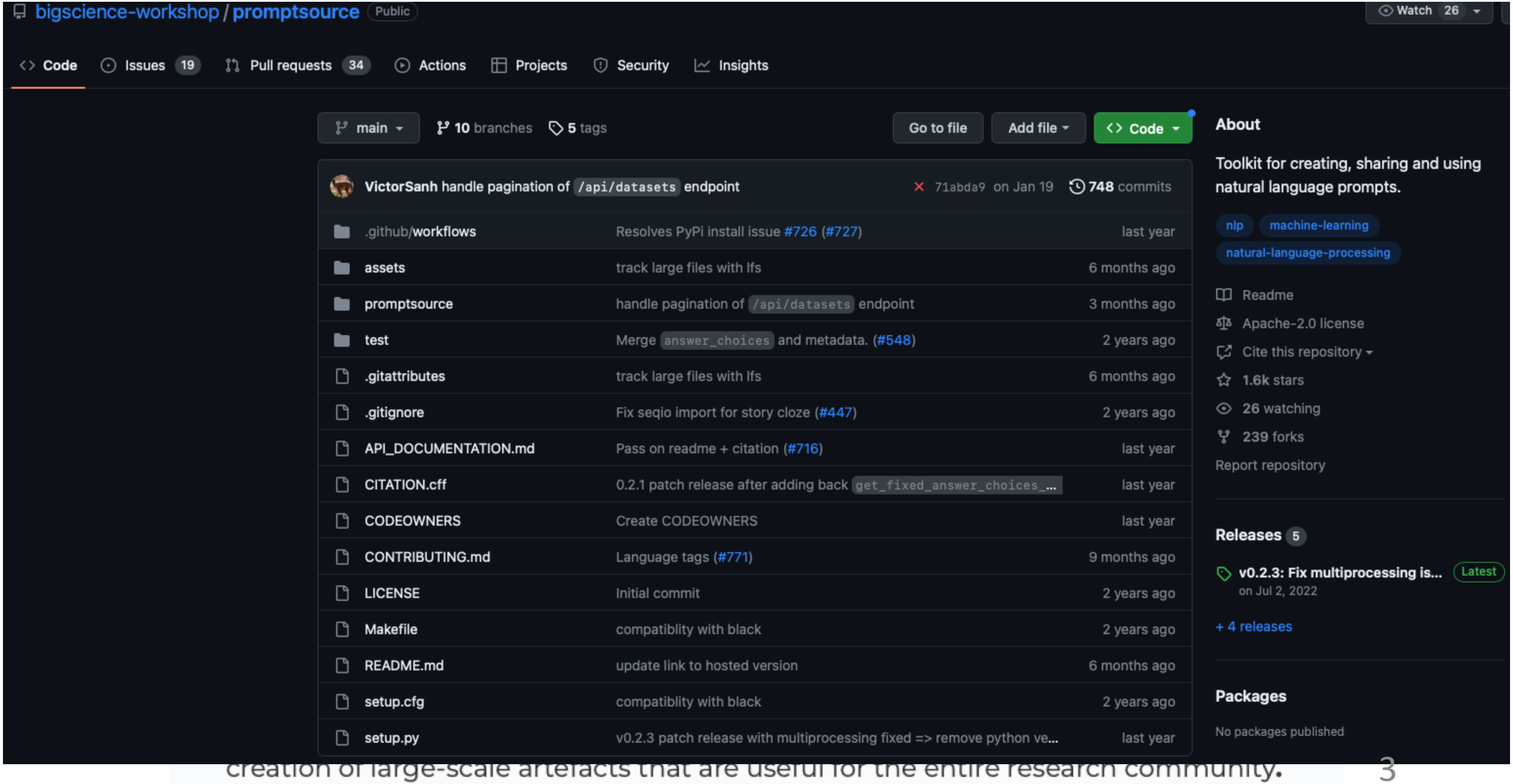
Stella Biderman
Booz Allen & EleutherAI

Leo Gao
EleutherAI

Thomas Wolf
Hugging Face

Alexander M. Rush
Hugging Face

- BigScience
- HuggingFace
- PromptSource



VictorSanh handle pagination of `/api/datasets` endpoint ✕ 71abda9 on Jan 19 748 commits

.github/workflows	Resolves PyPi install issue #726 (#727)	last year
assets	track large files with lfs	6 months ago
promptsource	handle pagination of `/api/datasets` endpoint	3 months ago
test	Merge `answer_choices` and metadata. (#548)	2 years ago
.gitattributes	track large files with lfs	6 months ago
.gitignore	Fix seqio import for story cloze (#447)	2 years ago
API_DOCUMENTATION.md	Pass on readme + citation (#716)	last year
CITATION.cff	0.2.1 patch release after adding back `get_fixed_answer_choices_...`	last year
CODEOWNERS	Create CODEOWNERS	last year
CONTRIBUTING.md	Language tags (#771)	9 months ago
LICENSE	Initial commit	2 years ago
Makefile	compatiblity with black	2 years ago
README.md	update link to hosted version	6 months ago
setup.cfg	compatiblity with black	2 years ago
setup.py	v0.2.3 patch release with multiprocessing fixed => remove python ve...	last year

About

Toolkit for creating, sharing and using natural language prompts.

Releases 5

v0.2.3: Fix multiprocessing is... Latest
on Jul 2, 2022

+ 4 releases

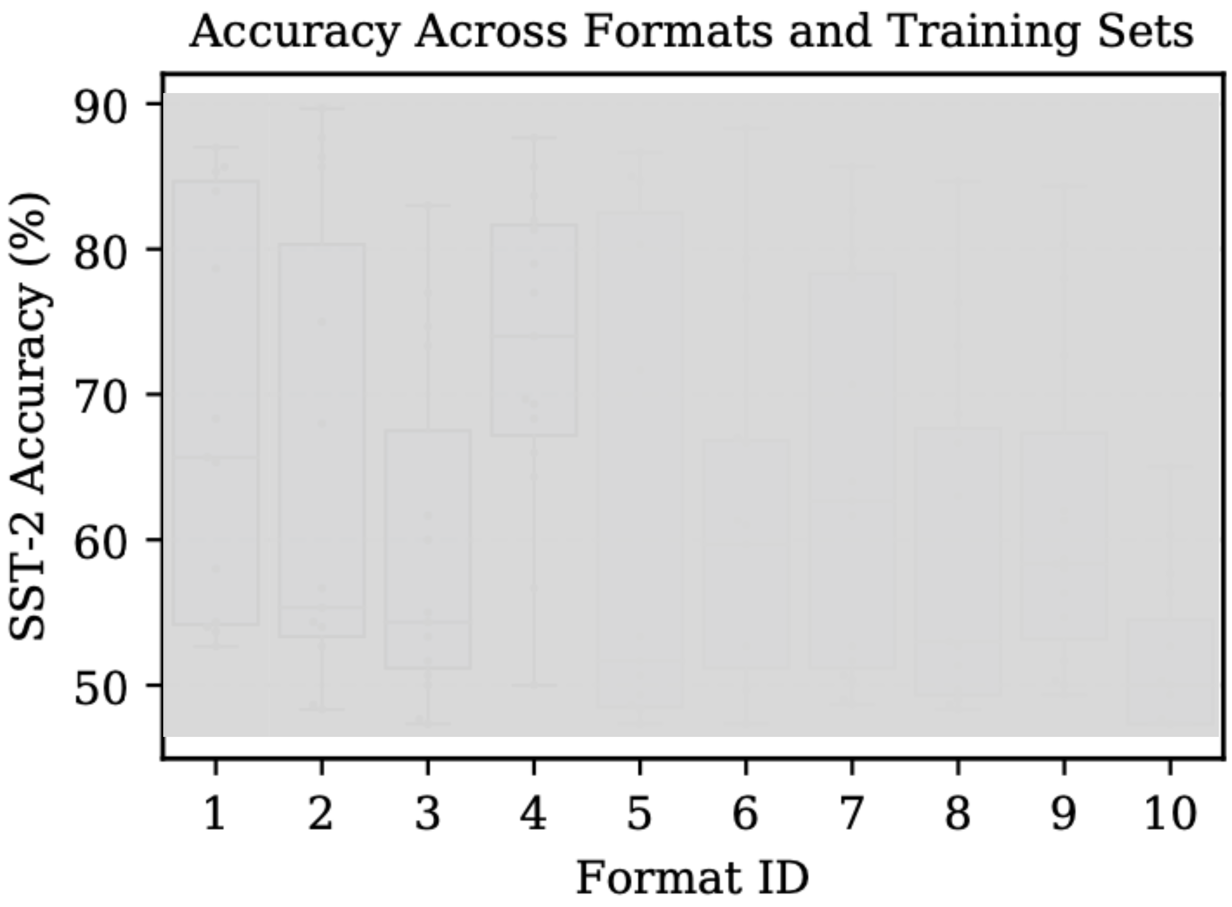
Packages

No packages published

creation of large-scale artefacts that are useful for the entire research community.

Challenge 1: **Huge variance** against **semantically equivalent prompts**

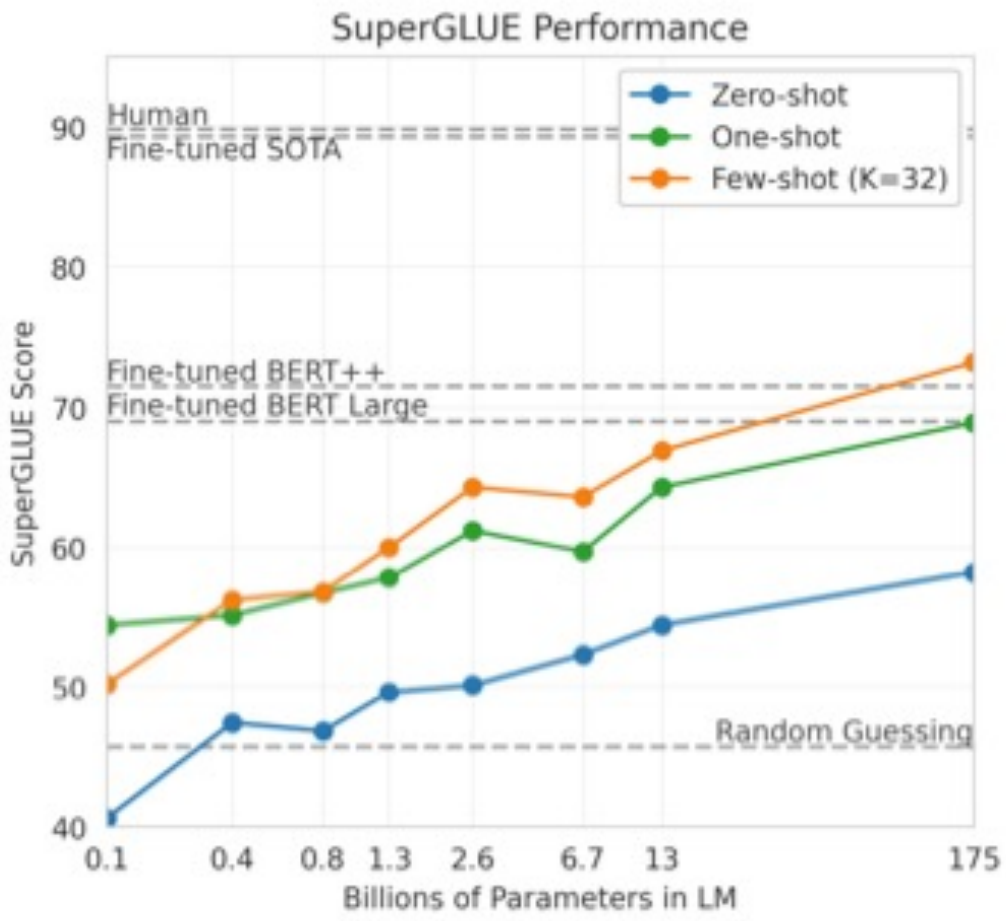
GPT3:



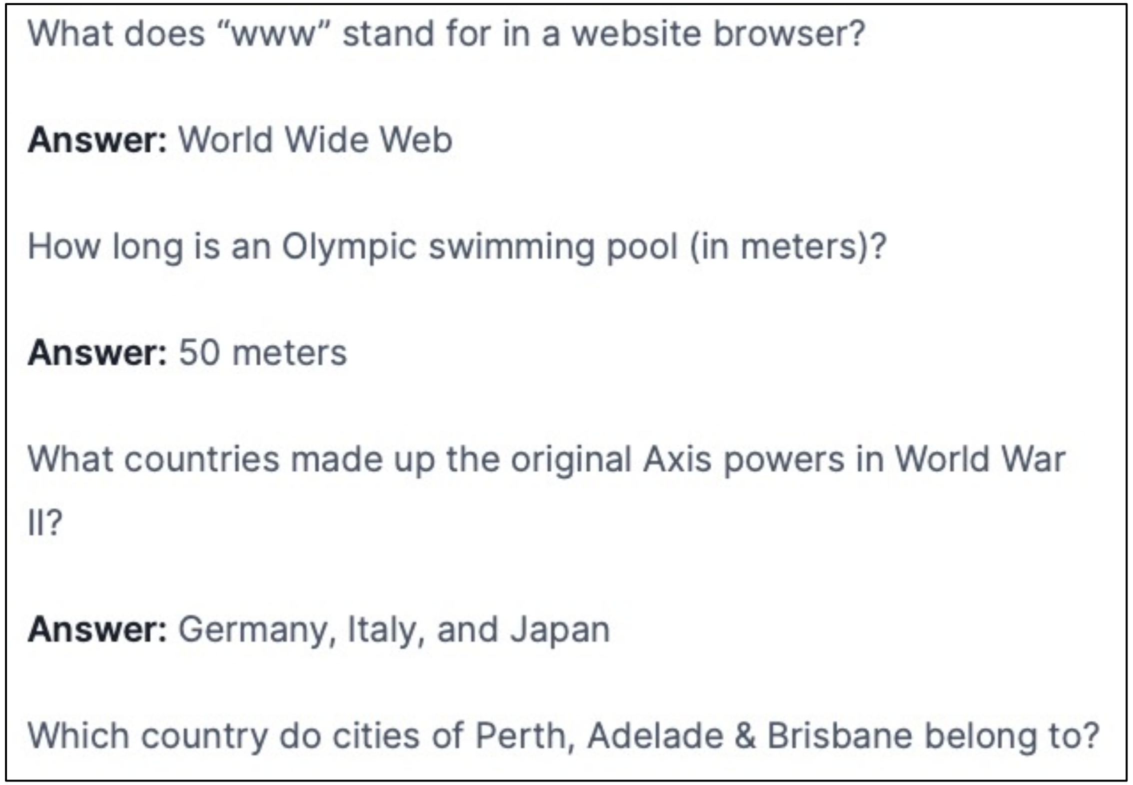
1. Review: [example]   Answer: [positive/negative]
2. Review: [example]   Question: Did the author think that the movie was good or bad? Answer: [good/bad]
3. My review for last night's film: [example] The critics agreed that the movie was [good/bad]
4. Critical reception [edit] In a contemporary review, Roger Ebert wrote [example] Entertainment weekly agreed and the overall critical reception of the film was [good/bad]

[1] Zhao, Zihao, et al. "Calibrate before use: Improving few-shot performance of language models." *International Conference on Machine Learning*. PMLR, 2021.

Challenge 2: **Zero-shot** only works with a **giant model (>100B)**



Hypothesis: Large models undergo **implicit** multi-task training in their **pre-training** corpora.



What does “www” stand for in a website browser?

Answer: World Wide Web

How long is an Olympic swimming pool (in meters)?

Answer: 50 meters

What countries made up the original Axis powers in World War II?

Answer: Germany, Italy, and Japan

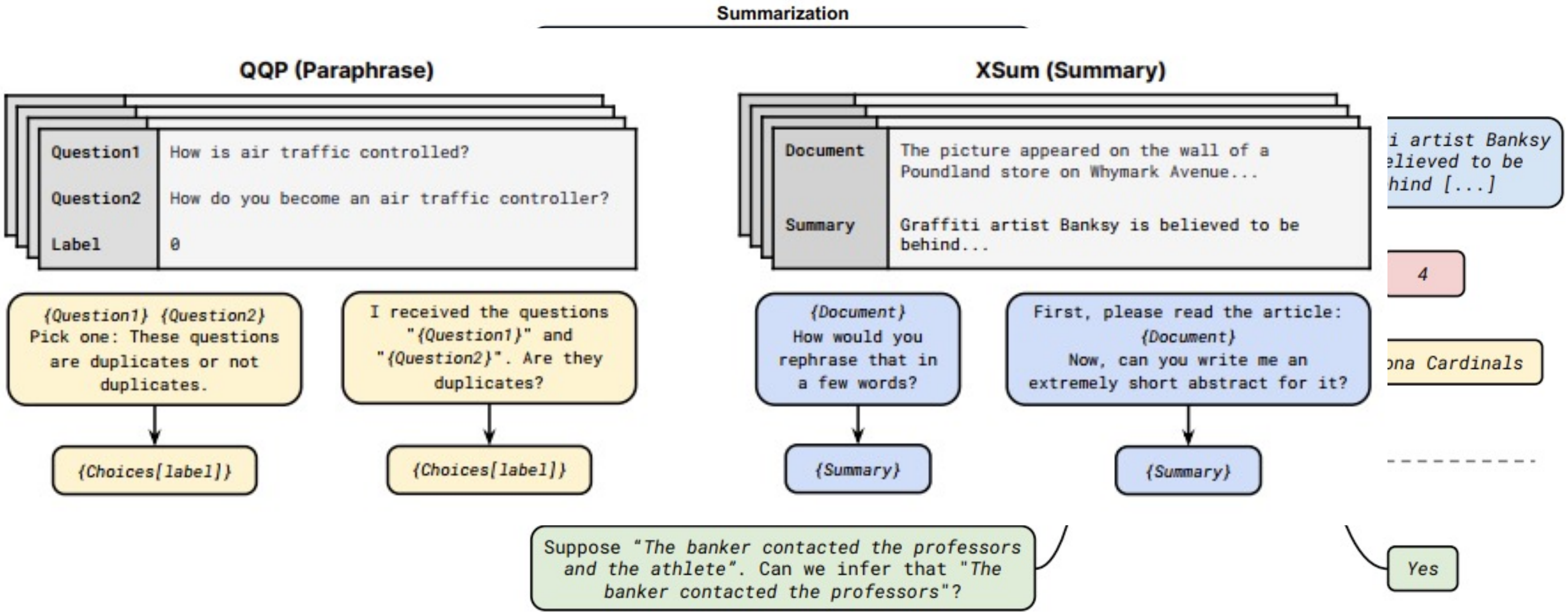
Which country do cities of Perth, Adelade & Brisbane belong to?

[1] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Proposal: Make **implicit** multi-task learning **explicit**

- Bette tasks
- More choic

5-10 prompts for each dataset



Summarization

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?

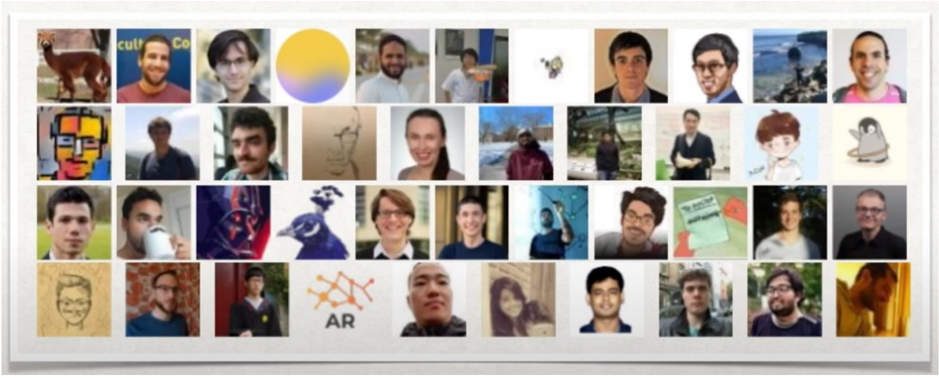
{Summary}

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

Yes

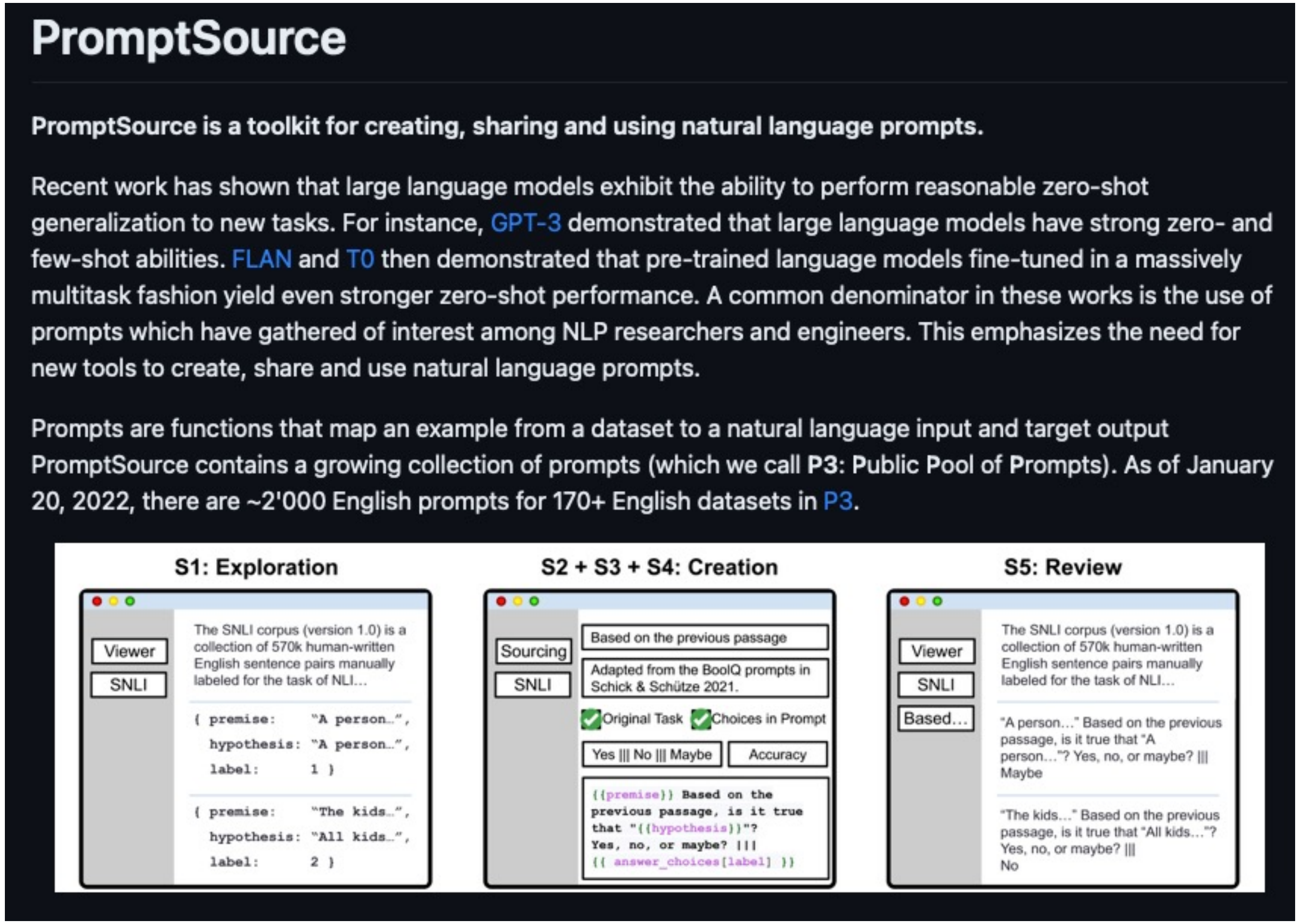
Proposal: Make **implicit** multi-task learning **explicit**

Open call for contributing prompts



- ➢ 62 Datasets
- ➢ 520 prompts

Interface to write, review, and debug prompts



PromptSource

PromptSource is a toolkit for creating, sharing and using natural language prompts.

Recent work has shown that large language models exhibit the ability to perform reasonable zero-shot generalization to new tasks. For instance, GPT-3 demonstrated that large language models have strong zero- and few-shot abilities. FLAN and T0 then demonstrated that pre-trained language models fine-tuned in a massively multitask fashion yield even stronger zero-shot performance. A common denominator in these works is the use of prompts which have gathered of interest among NLP researchers and engineers. This emphasizes the need for new tools to create, share and use natural language prompts.

Prompts are functions that map an example from a dataset to a natural language input and target output PromptSource contains a growing collection of prompts (which we call **P3**: Public Pool of Prompts). As of January 20, 2022, there are ~2'000 English prompts for 170+ English datasets in P3.

S1: Exploration

Viewer

SNLI

The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for the task of NLI...

```
{ premise:    "A person…",
  hypothesis: "A person…",
  label:       1 }

{ premise:    "The kids…",
  hypothesis: "All kids…",
  label:       2 }
```

S2 + S3 + S4: Creation

Sourcing

SNLI

Based on the previous passage

Adapted from the BoolQ prompts in Schick & Schütze 2021.

☑ Original Task ☑ Choices in Prompt

Yes ||| No ||| Maybe

Accuracy

```
{{premise}} Based on the
previous passage, is it true
that "{{hypothesis}}"?
Yes, no, or maybe? |||
{{ answer_choices[label] }}
```

S5: Review

Viewer

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...

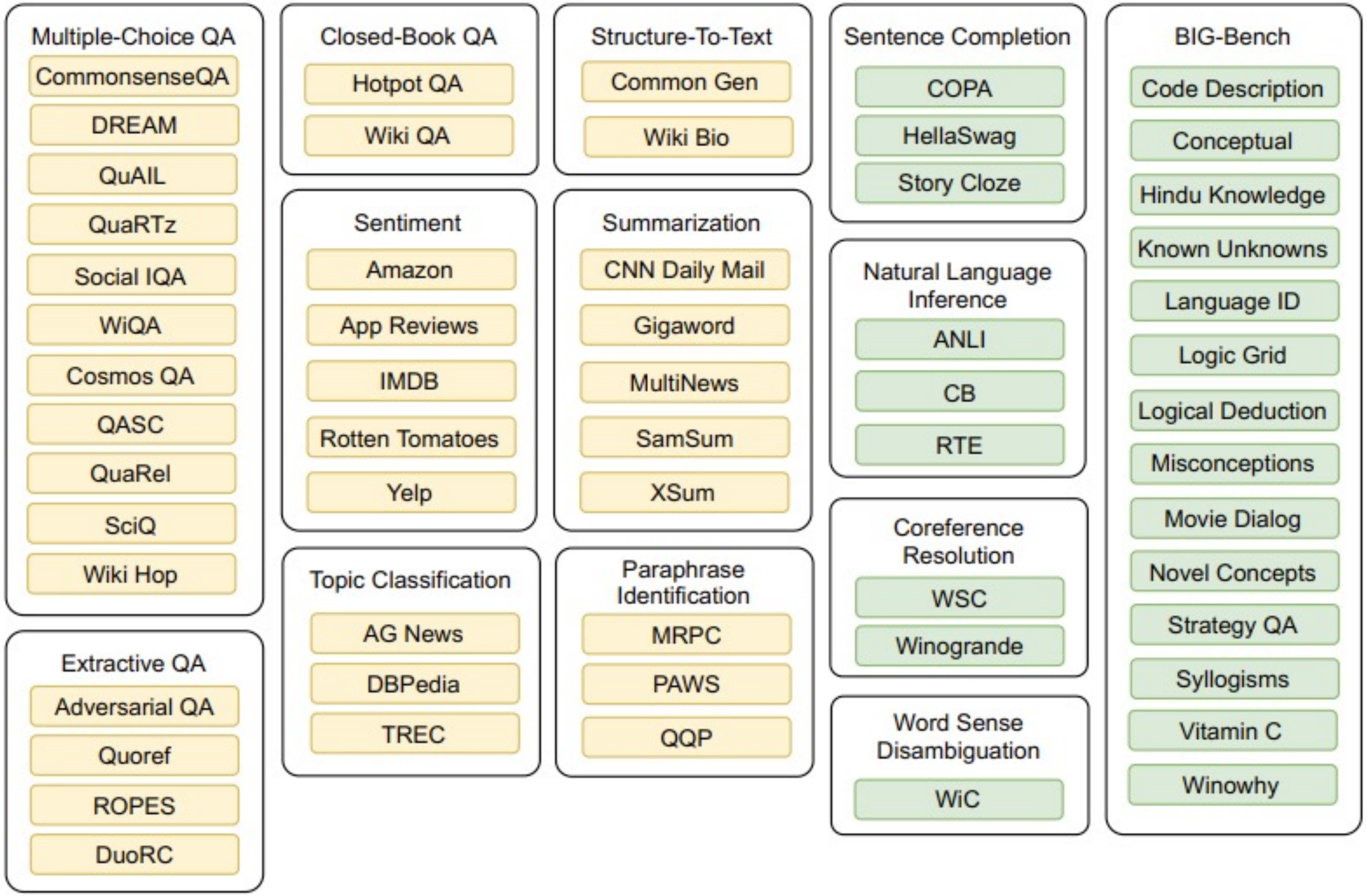
"A person..." Based on the previous passage, is it true that "A person..."? Yes, no, or maybe? ||| Maybe

"The kids..." Based on the previous passage, is it true that "All kids..."? Yes, no, or maybe? ||| No

T0 Does not see any of the held-out datasets/tasks during **training**

Training datasets/tasks

Held-out datasets/tasks



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

Training and evaluation:

Base model

T5:

a Transformer-based encoder-decoder language model pretrained with a masked language modeling-style objective.

T0:

11B/3B parameters, Trained on datasets mentioned earlier.

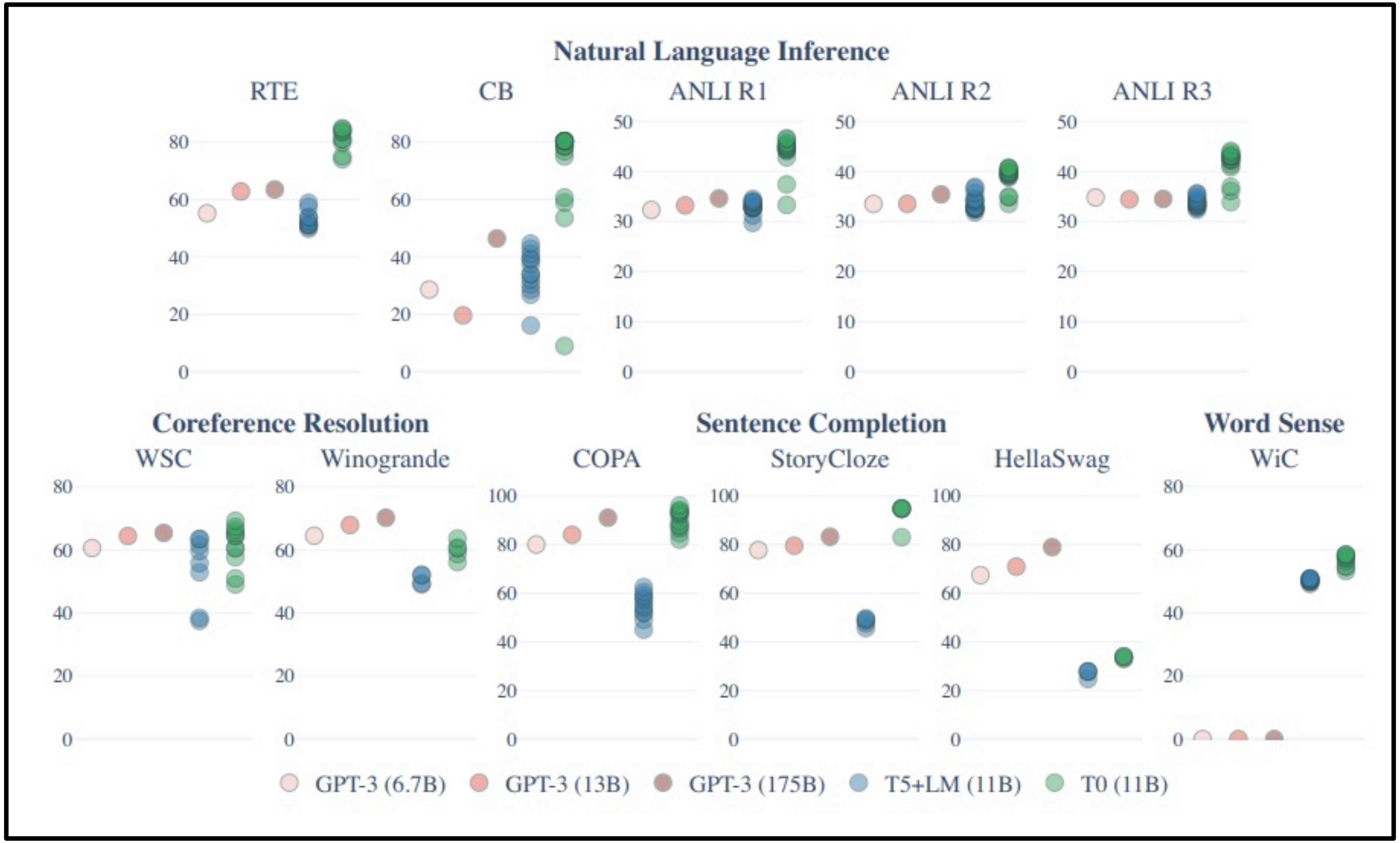
T0+:

Same hyperparameters and size but with added training sets from GPT3's evaluation datasets.

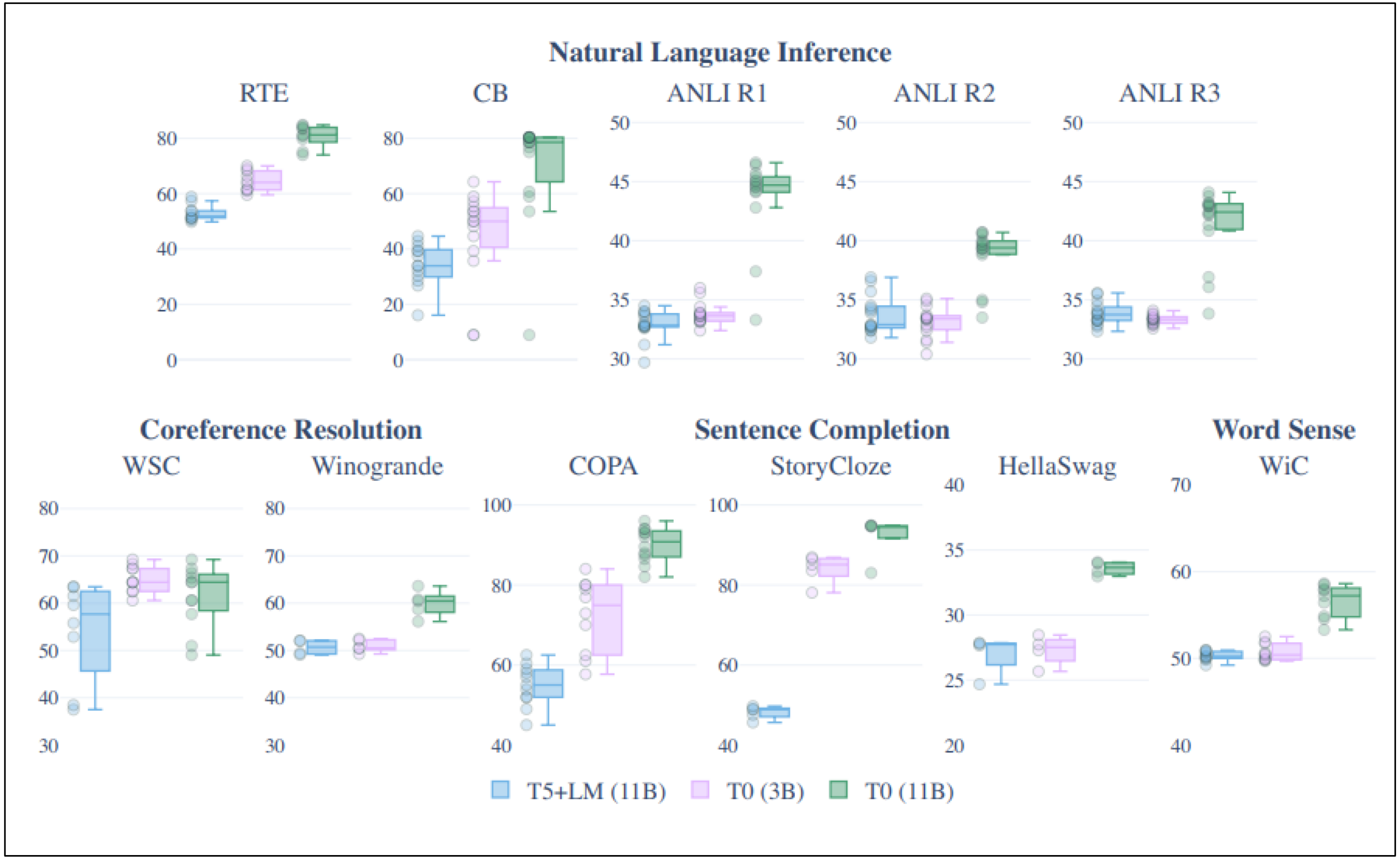
T0++:

Same hyperparameters and size but with added training sets from GPT3's evaluation datasets and SuperGlue.

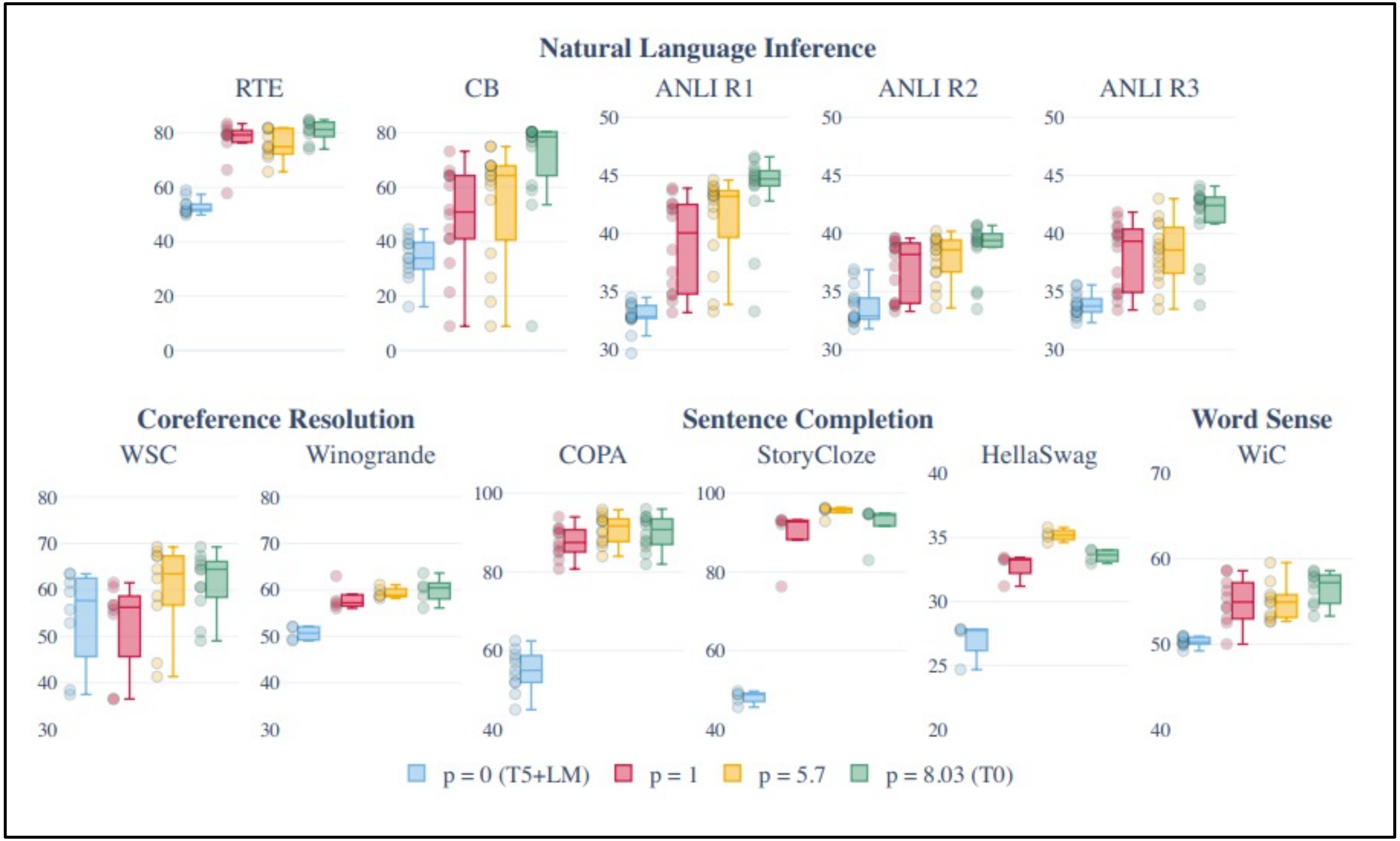
Zero-Shot **Performance** on **Held-out Tasks**



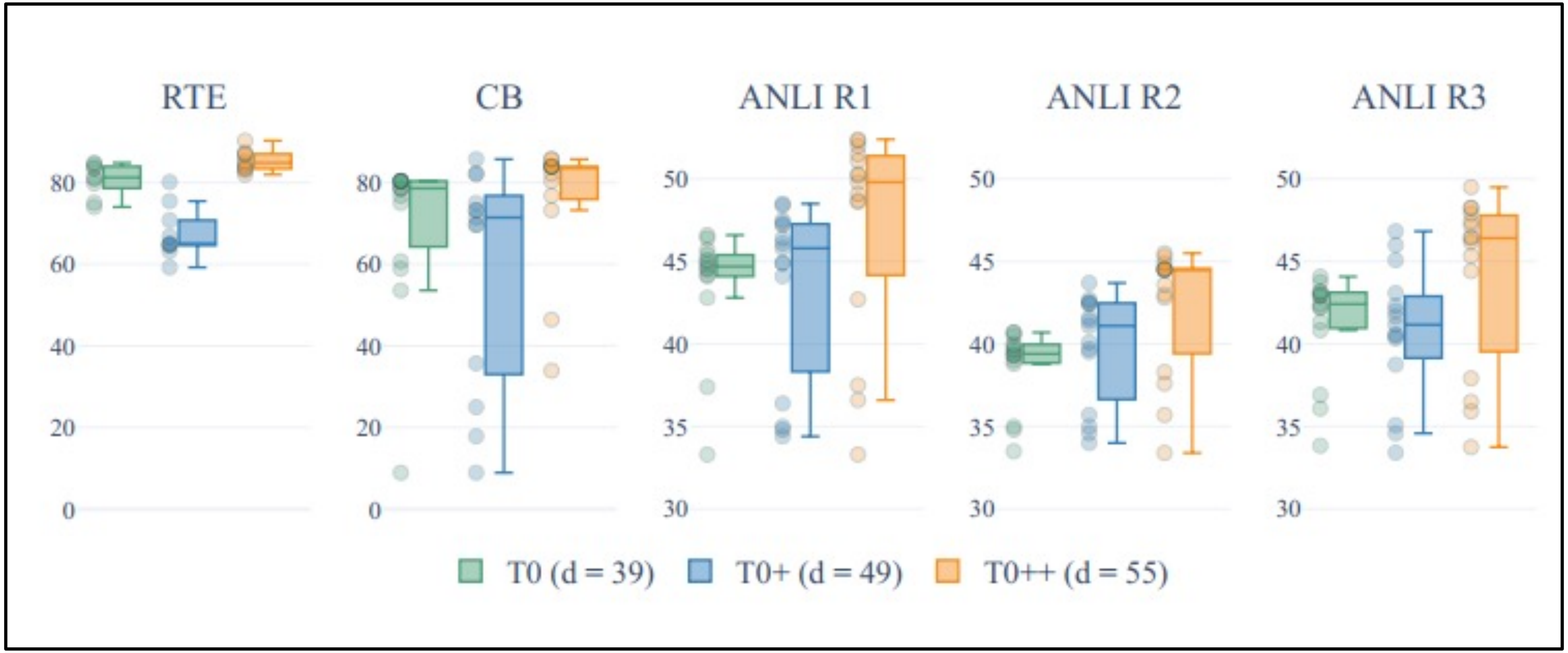
Effect of **size** of the pretrained model



Training on **more diverse prompts** improves **robustness** on held-out tasks



Effect of number of training datasets:



Strengths:

- Effectiveness of multitask prompted training in achieving strong zero-shot generalization abilities.
- Evidence that T0, a smaller model, can outperform larger models like GPT-3 in several held-out tasks, showcasing its efficiency.
- Extensive ablation studies, highlighting the importance of diverse prompts and the impact of increasing the number of datasets in each task.
- Releasing all trained models, the collection of prompts, and the prompt annotation tool to the research community, fostering future work in the area of zero-shot generalization.