Multitask Prompted Training Enables Zero-Shot Task Generalization

Apr 5, 2022

Reading Group

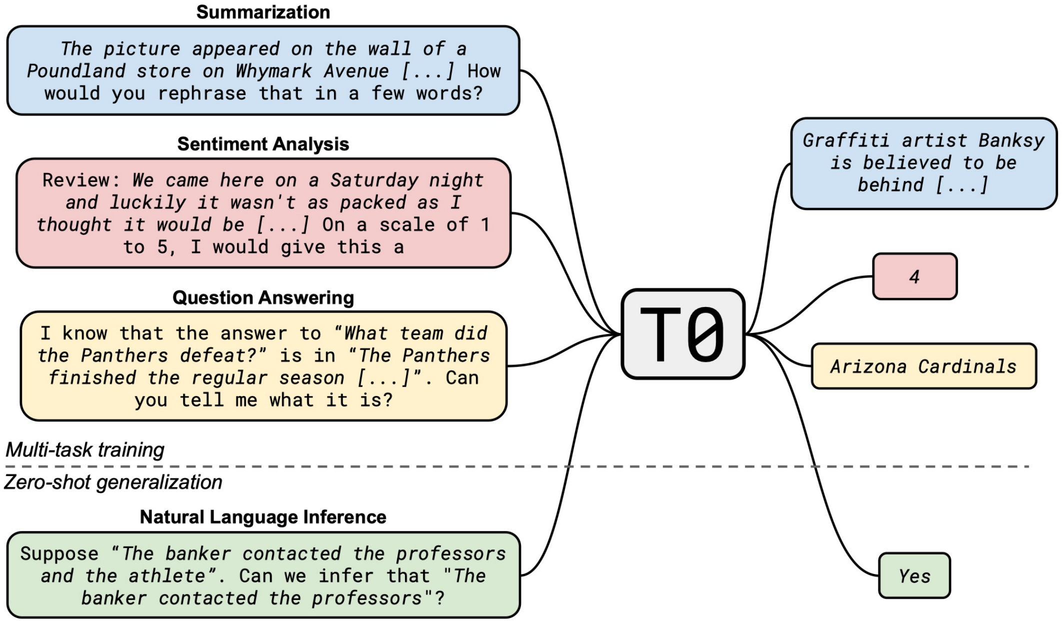
Presented by Mingyu Derek Ma

Introduction

- Recent works show that large LMs exhibit zero-shot generalization ability with only language modeling objectives
- An influential hypothesis is that large language models generalize to new tasks as a result of **an implicit process of multitask learning**
  - Generic text in the pretraining corpus may contain format and structure of QA
  - Given the large training corpus, it's reasonable to expect some tasks would appear explicitly in the pretraining corpora, like lists of trivia QA pairs
- What about convert those implicit signals to explicit ones, by training the model directly in a supervised and massively multi-task fashion

Goal

- Induce a model to better generalize to held-out tasks
- Being more robust to the wording choices of the prompts



Two questions

- Does multitask prompted training improve generalization to held-out tasks?

- Does training on a wider range of prompts improve robustness to prompt wording?

Related Works

- Implicit multitask learning in LM pretraining
- Explicit multitask learning
- Leading hypothesis is that models learn to understand the prompts as task instructions which help them generalize to held-out tasks
  - Depend on semantic meaningfulness of the prompts? Challenged
  - Only claim that prompts serve as a natural format for multitask training which empirically supports generalization to held-out tasks

Task

- Task: to refer to a general NLP ability that is tested by a group of specific datasets
- Create task taxonomy to mitigate fuzzy categorization issue
- 12 tasks and 62 datasets
  - Only English tasks
  - Not require special domain knowledge like biomedicine
  - No tasks about programming languages and structured annotations such as parse trees

Tasks

- Yellow: training mixture
- Green: held out

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

Generalization vs memorization

Contamination analysis of pretraining corpus on test tasks

- Appearance of long common substrings between the **zero-shot test tasks** and **documents in C4**
- NLI premises tend to be sourced from the internet -> high numbers of matches
    - HellaSwag has 9.12% matches
    - ANLI has negligible overlapped hypotheses
    - RTE has high match numbers for both premises and hypotheses

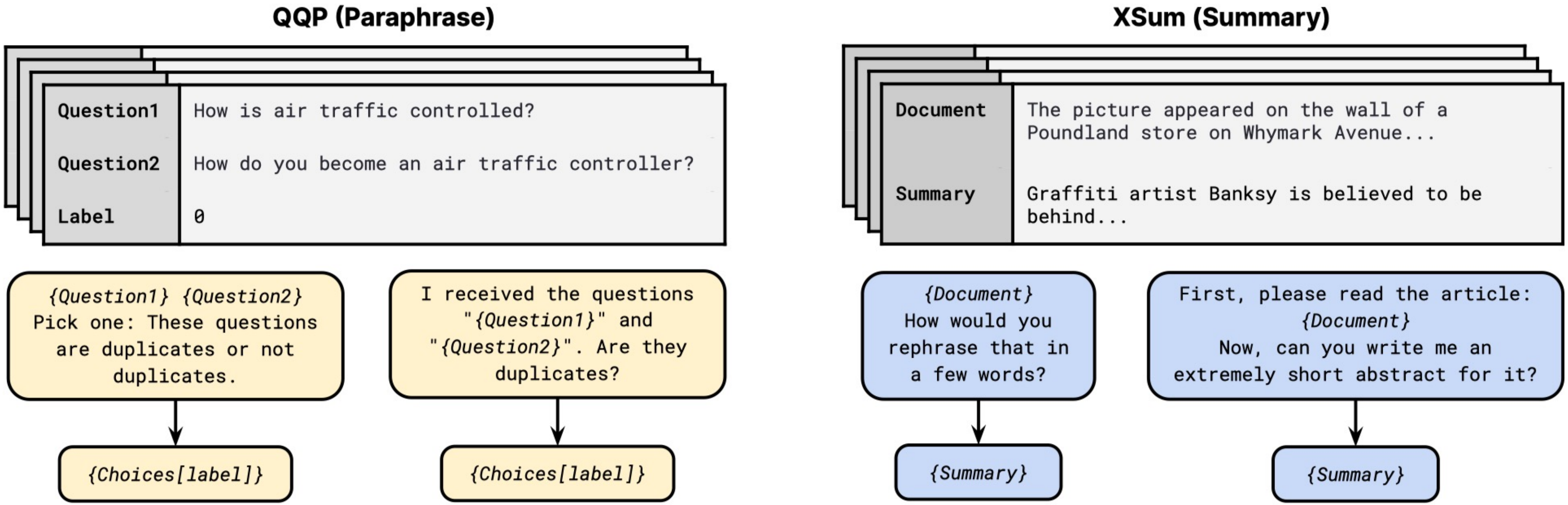
Task	CB	HellaSwag	Lambada	Story Cloze	WiC	Winogrande	WSC
Matches	1/250	912/10000	15/5153	3/1871	20/1400	0/1767	4/146

Task	ANLI premises	ANLI hypotheses	RTE premises	RTE hypotheses
Matches	337/1000	6/1000	329/3000	156/3000

Prompt Templates

- Convert diverse datasets into prompts
- Prompt template
  - Input template
  - Target template
- Built an interface to collect prompts interactively from the research community
- As long as the prompts are grammatical and understandable, creators can be creative
- Public Pool of Prompts (P3)
  - 2073 prompts for 177 datasets
  - 36 contributors
  - Each dataset has multiple prompt template

Prompt Templates



Generative model and training

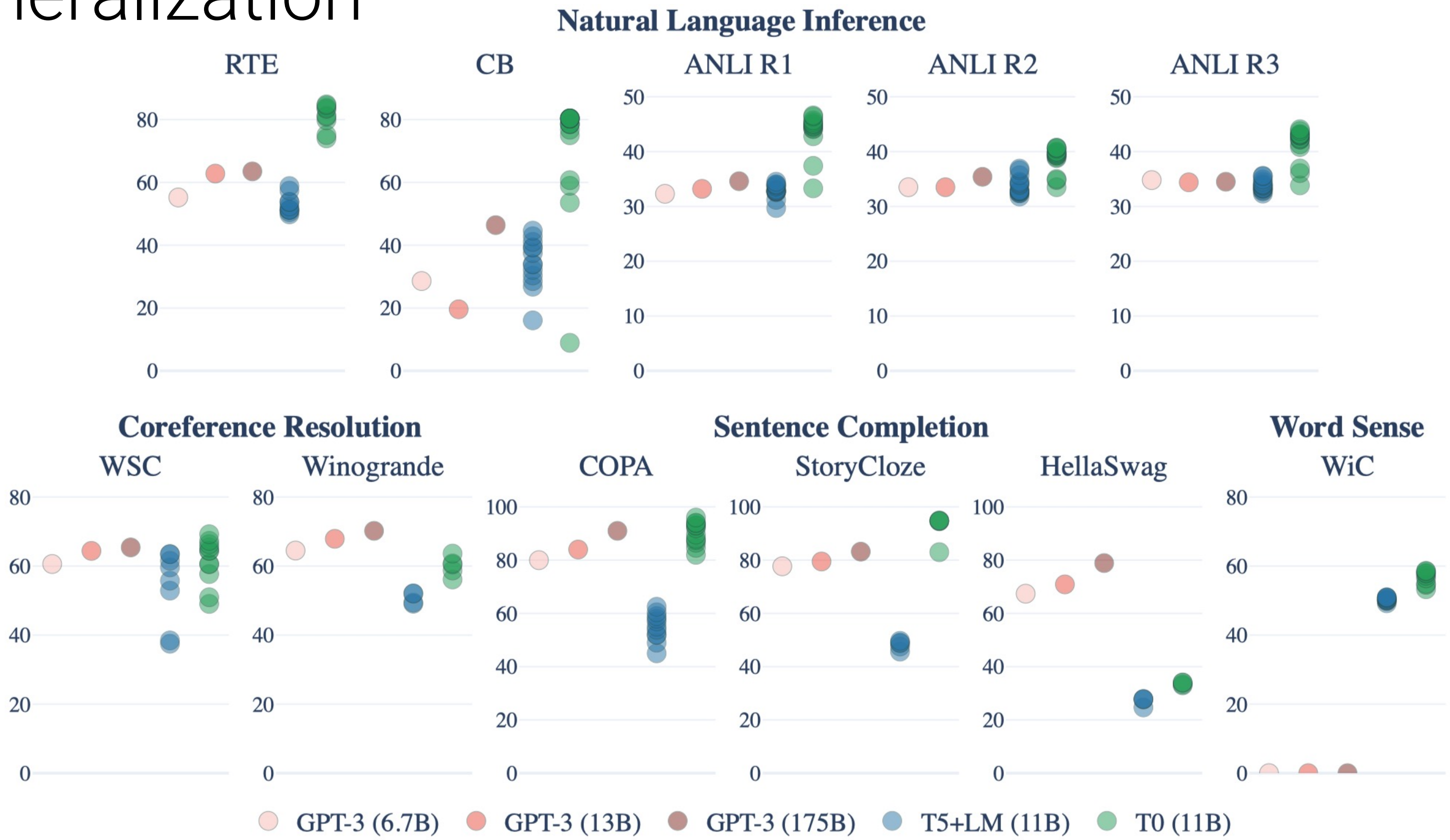
- Encoder-decoder architecture, never trained to generate the input
- Standard maximum likelihood training
- Based on T5
- Three versions
  - T0
  - T0+: T0 but additionally trained on GPT3’s evaluation datasets
  - T0++: further adds SuperGLUE (except RTE and CB) as training dataset
- Two sizes
  - 11B parameters
  - 3B parameters
- Checkpoint selection: choosing the one yield the highest score on the validation splits of the training datasets -> still true zero-shot

Evaluation

- If the task is choosing from several options like multiple choice QA, they apply rank classification
  - Compute log-likelihood of each of the target options under the fine-tuned model
  - Select the option with the highest log-likelihood as the prediction
- Report **median** performance and **interquartile range** across **all prompts** for this dataset

Results: Generalization

- **T0 vs T5+LM**: benefits of multitask prompted training
- **T0 (11B) vs GPT-3 (175B):** matches/exceeds GPT-3 on 9 out of 11 held-out datasets

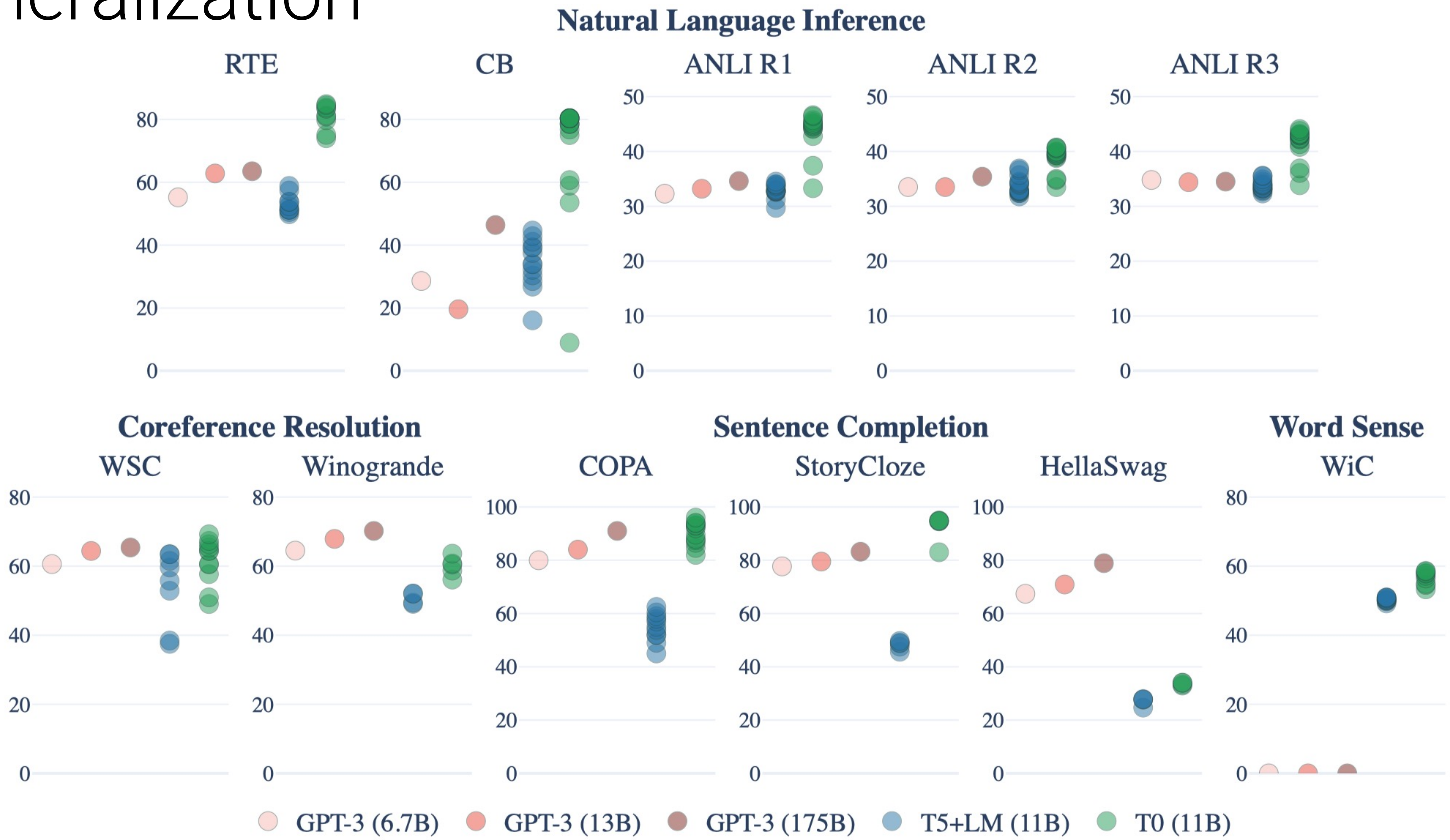


Results: Generalization

T0 (11B) vs GPT-3 (175B): two exceptions, Winogrande and HellaSwag

Wei et al., 2021 also observes similar trend

HellaSwag's median increases from 33.65% to 57.93% if removing instruction, Winogrande performance do not improve though



1.9.2 HELLASWAG

Dataset from Zellers et al. (2019). Used in evaluation.

Data Example

Key	Value
activity_label	Removing ice from car
ctx	Then, the man writes over the snow covering the wi...
ctx_a	Then, the man writes over the snow covering the wi...
ctx_b	then
endings	[', the man adds wax to the windshield and cuts it...
ind	4
label	3
source_id	activitynet~v_-1IBHYS3L-Y
split	train
split_type	indomain

Prompts

Input Template:

```
Complete the description with an appropriate ending:
First, {{ ctx_a.lower() }} Then, {{ ctx_b.lower() }} ...

(a) {{ answer_choices[0] }}

(b) {{ answer_choices[1] }}

(c) {{ answer_choices[2] }}

(d) {{ answer_choices[3] }}
```

Target Template:

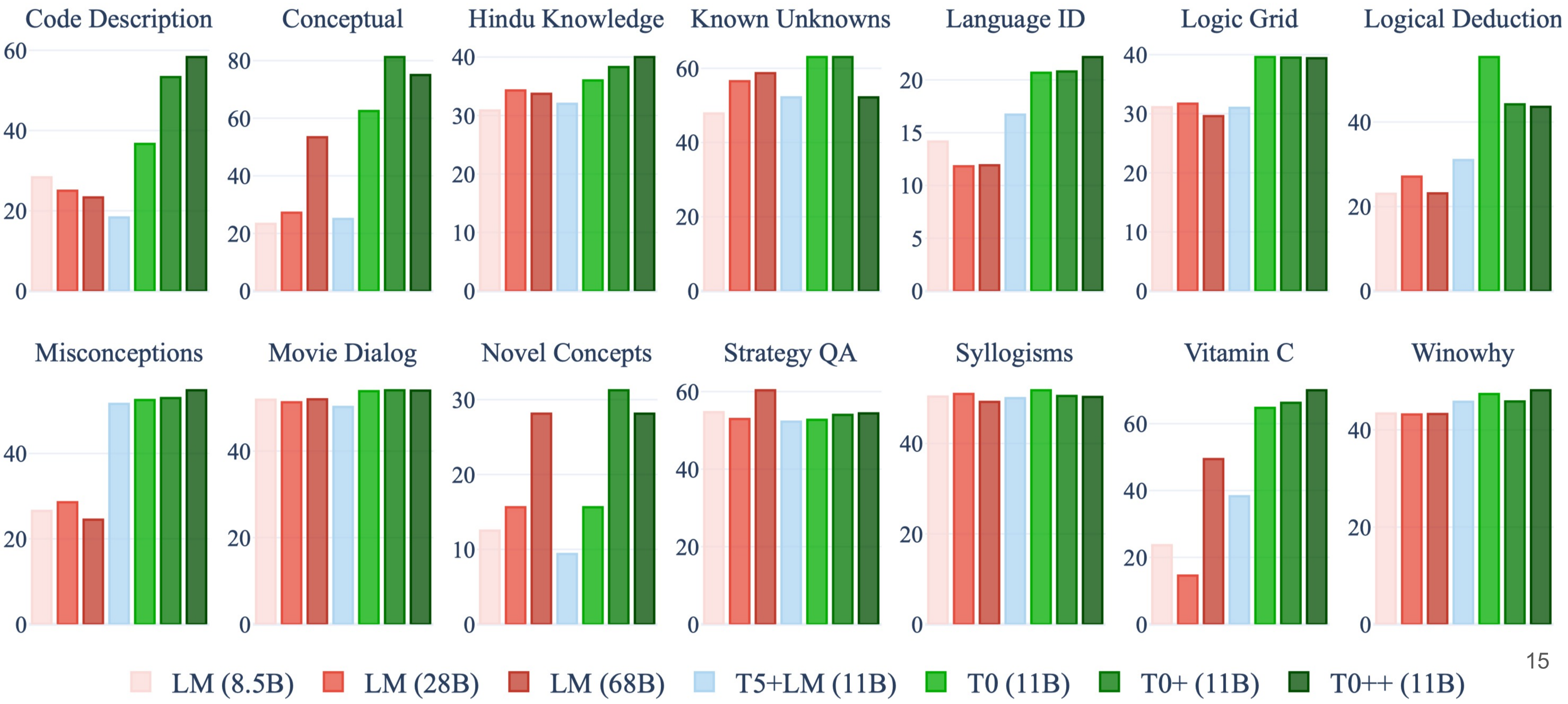
```
{{ answer_choices[label | int()] }}
```

Answer Choices Template:

```
{{endings | join(" ||| ")}}
```

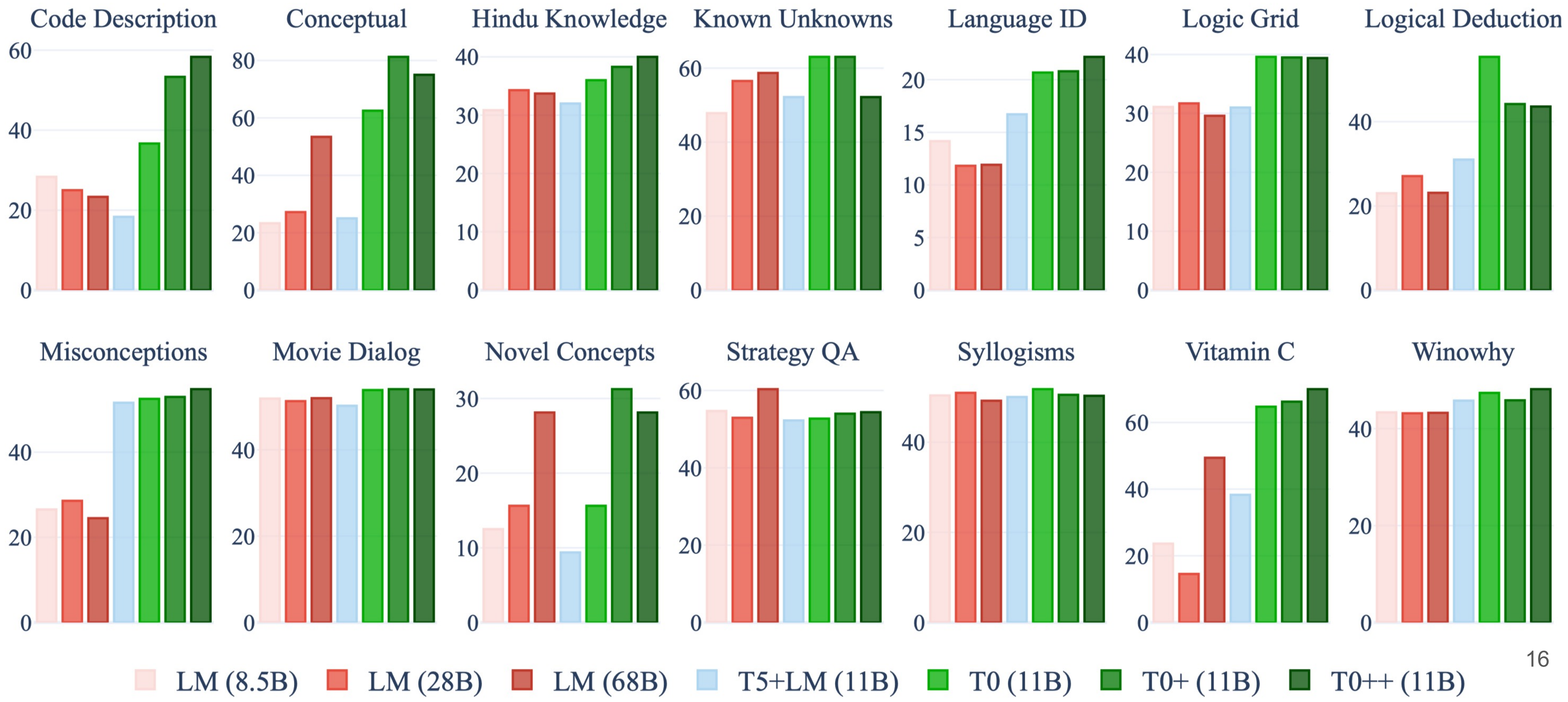
Results: Generalization

- The dataset contains prompt for each sub-dataset
- Baseline LMs are decoder-only Transformer LMs



Results: Generalization

- At least one of the T0 variants outperform all baseline models on all tasks except for StrategyQA
- In most cases, training datasets increases, the better performance (T0++ > T0+ > T0)

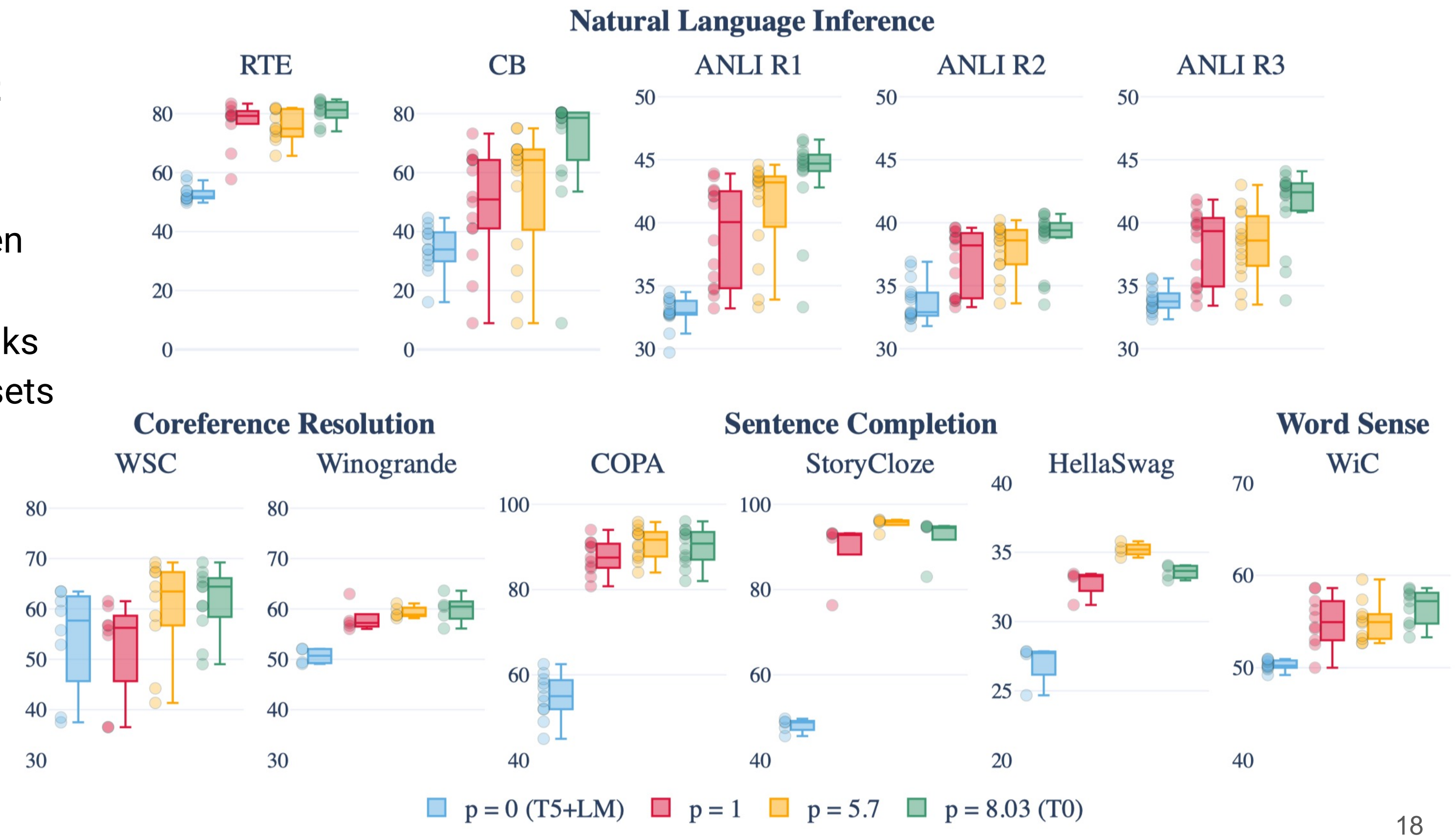


Results: Prompt Robustness

- Wider range of prompts improves robustness to the wording of the prompts?
    - Effect of prompts per datasets
    - Effect of more datasets

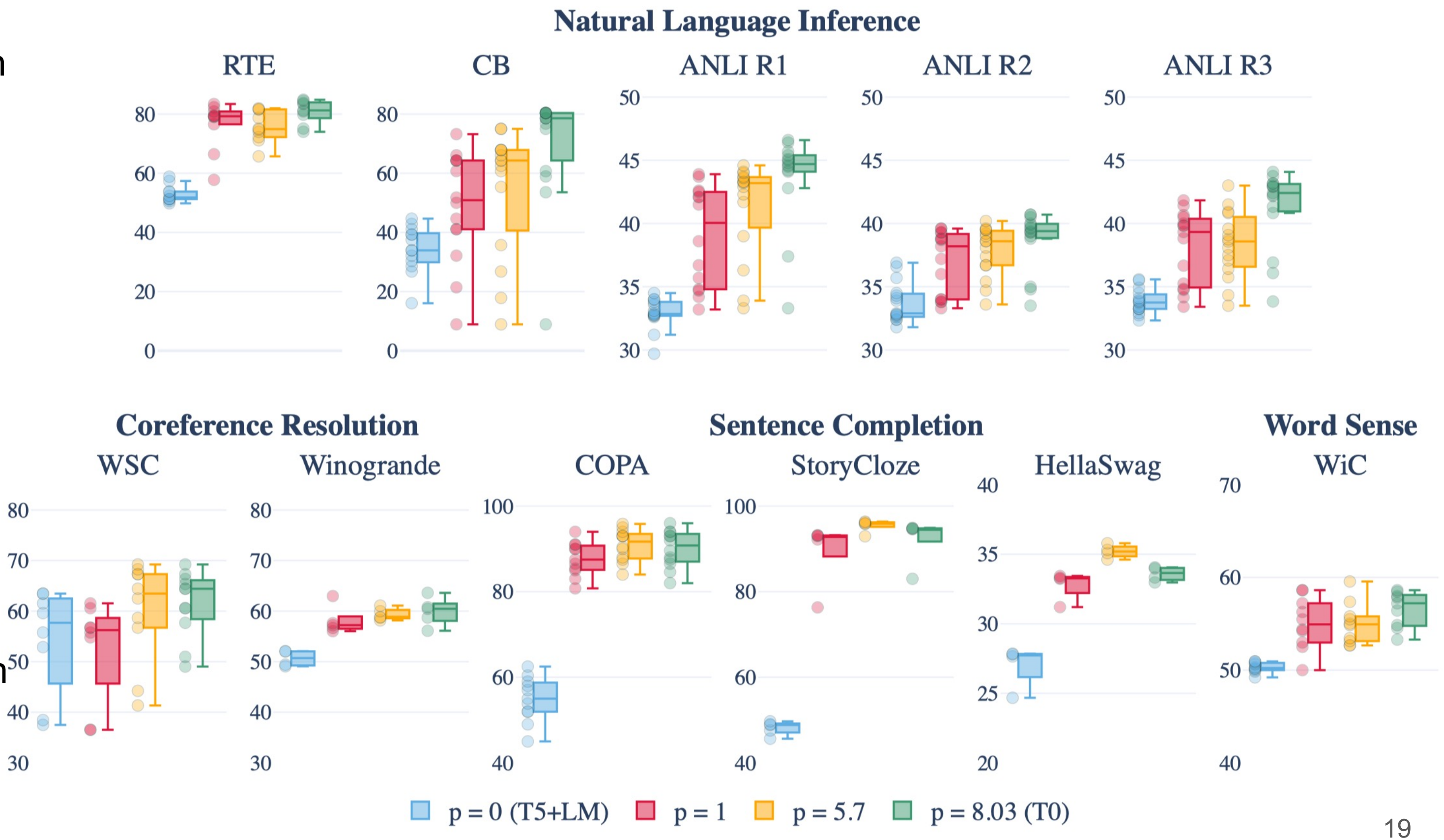
Results: Prompt Robustness

Prompts per dataset
p=0: no prompted training
p=1: randomly chosen prompt
p=5.7: all original-tasks prompts for all datasets
p=8.03: T0 setting



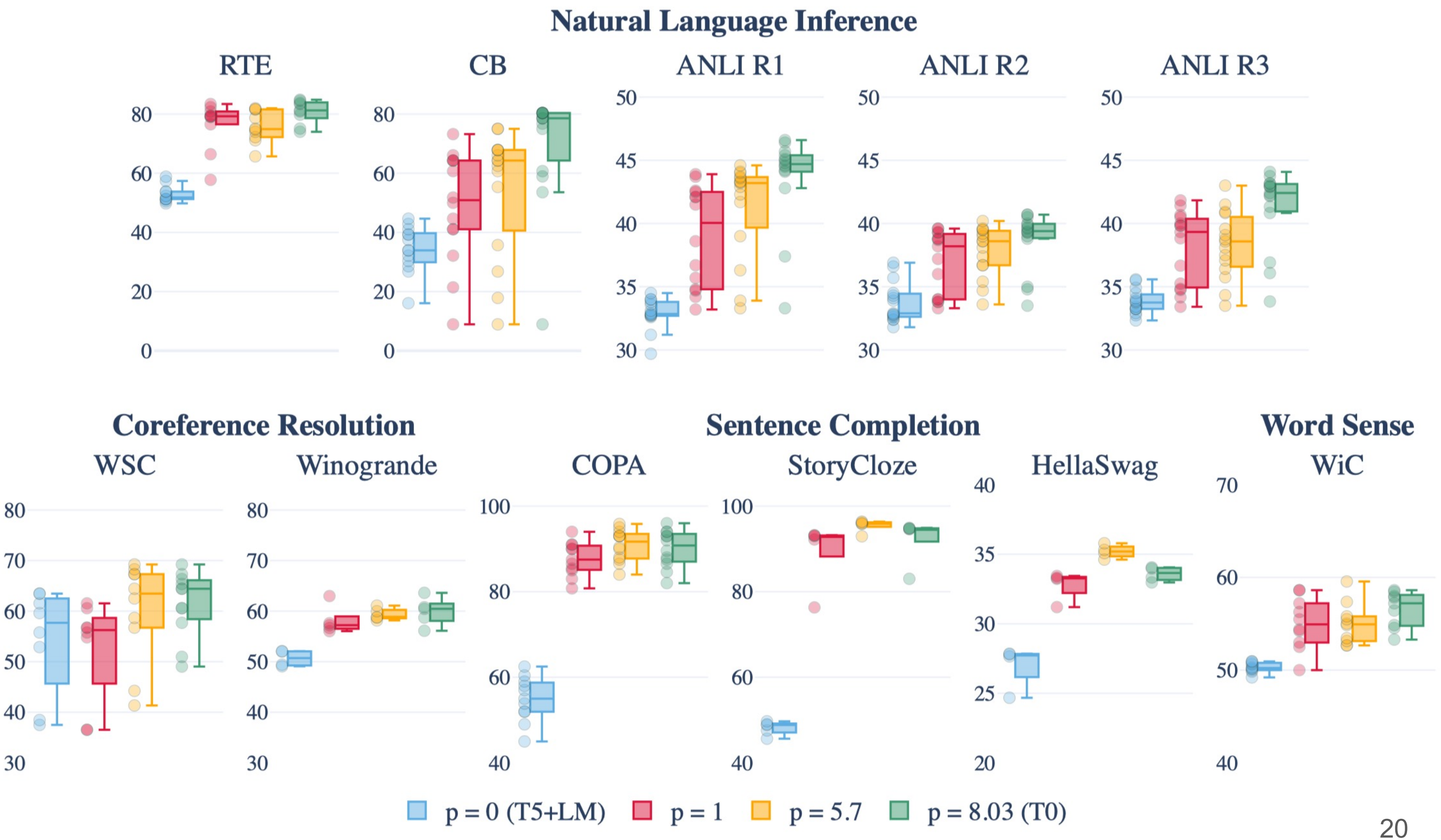
Results: Prompt Robustness

- Large variance when using different evaluation prompts
- Even with 1 prompt, performance on held-out tasks can improve substantially over the non-prompted baseline
- Increase from 1 to 5.7 yields additional improvement in both median and spread for most datasets



Results: Prompt Robustness

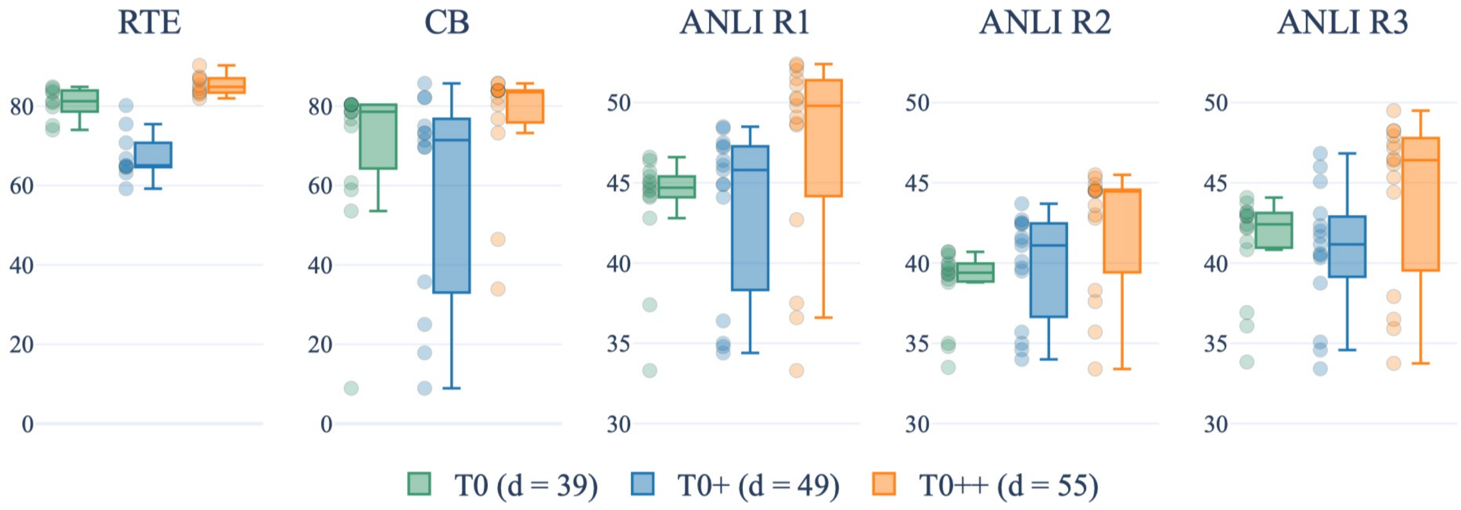
- T0's inclusion all prompts further improves the median (9/11) and spread (8/11) generally
- Training on more prompts per dataset lead to better and more robust generalization
- Training on non-original-task prompts can also be beneficial



Results: Prompt Robustness

- Wider range of prompts improves robustness to the wording of the prompts?
  - Effect of prompts per datasets
  - Effect of more datasets

Results: Prompt Robustness



Prompts from more datasets

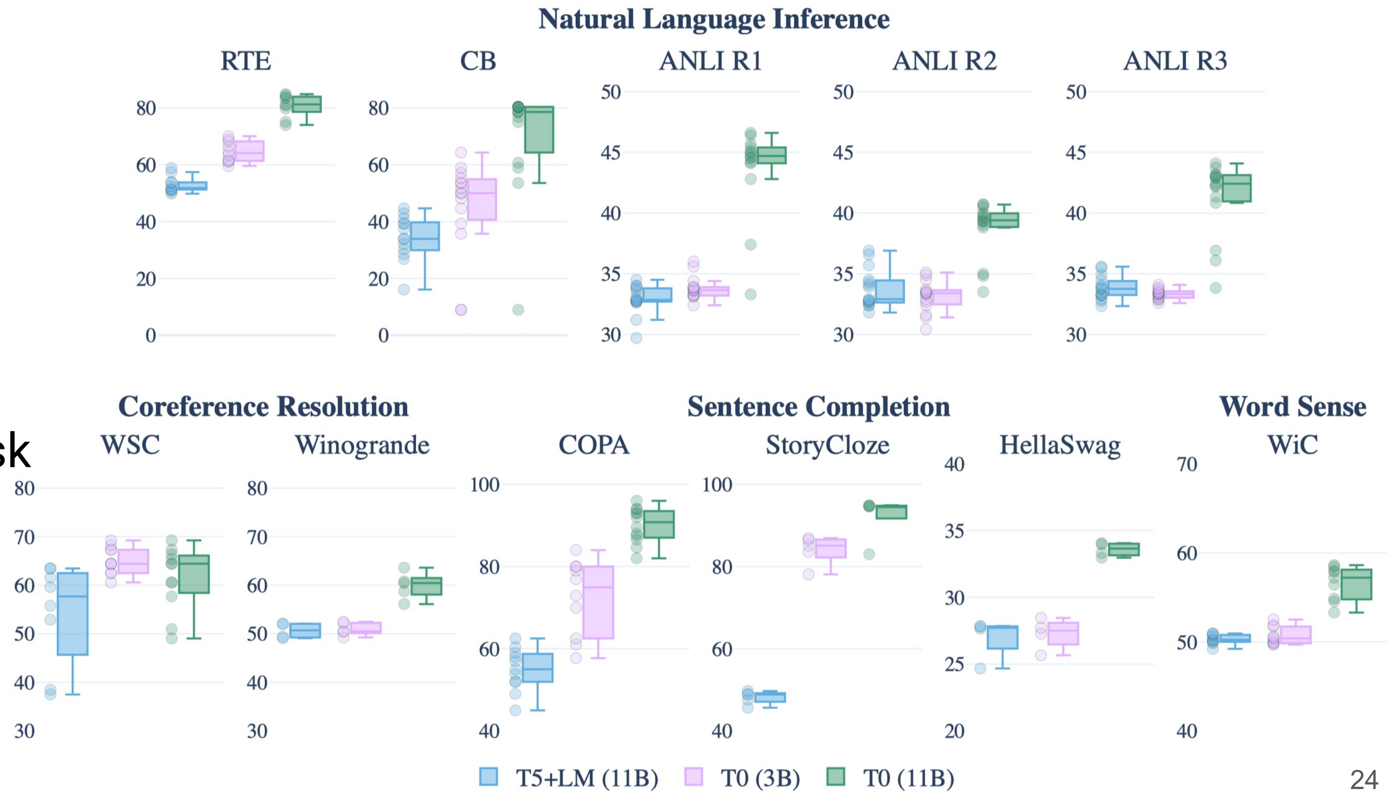
- Fix prompts per dataset, change number of datasets used in training
- Adding more datasets
  - Consistently leads to higher median performance
  - Does not always reduce interquartile range for held-out tasks

Results: Prompt Robustness

- Wider range of prompts improves robustness to the wording of the prompts?
  - Effect of prompts per datasets
    - Training on more prompts per dataset lead to better and more robust generalization
    - Training on non-original-task prompts can also be beneficial
  - Effect of more datasets
    - Increasing number of datasets does not consistently make the model more robust to the wordings of prompts

Results: Model Size

3B parameters version T0 also shows better generalization compared with T5+LM without prompted multitask training



Concurrent work: FLAN (Wei et al., ICLR'2022)

- Similar idea of enabling zero-shot generalization through multitask prompted training
- They train decoder-only LMs, they use single held-out task
- T0 (11B) is 10x smaller than FLAN (137B)
- T0 outperforms FLAN on some datasets, worse than FLAN on some other datasets
- Both T0 and FLAN underperform GPT-3 on Winogrande and HellaSwag on the coreference resolution task

Concurrent work: FLAN (Wei et al., ICLR’2022)

- They perform multi-task prompted training using an 8B model, but observed worse performance than baseline
  - While T0 shows 3B model shows better performance with multi-task prompted training
- FLAN shows more prompts has a negligible impact on performance
- Difference
  - T0 is based on encoder-decoder model
  - T0’s pretraining objective is MLM
  - T0’s prompts are qualitatively more diverse in terms of length and creativity

Conclusion: two questions

- Does multitask prompted training improve generalization to held-out tasks?
  - Multitask training enables zero-shot task generalization
  - T0 matches or exceeds the performance of GPT-3 on 9 out of 11 held-out datasets, with 16x smaller size
- Does training on a wider range of prompts improve robustness to prompt wording?
  - Training on more **prompts** per dataset consistently improves the median and decreases the variability of performance on held-out tasks
  - Training on prompts from a wider range of **datasets** also generally improves the medium but does not consistently decrease the variability