2023-1 Natural Language Processing

👌T0
: Multitask Prompted Training Enables Zero-Shot Task Generalization

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach et al.

BigScience, 2021

ICLR 2022 Spotlight

Yejin Yoon

NLP Lab.

Dept. of Artificial Intelligence Application, Hanyang University

stillwithyou@hanyang.ac.kr

What are Covered in this Presentation

- **Details of T0's method**
  - **T0 :** Sanh, Victor, et al. "Multitask prompted training enables zero-shot task generalization." arXiv preprint arXiv:2110.08207 (Oct. 2021).
- **Intuitive concepts of predecessors**
  - **T5 :** Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.
- **Intuitive concepts of successors (peers)**
  - **FLAN :** Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (Sep. 2021).
  - **Instruct-GPT** : Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.

What are NOT Covered in this Presentation

- **Details of datasets & tasks**

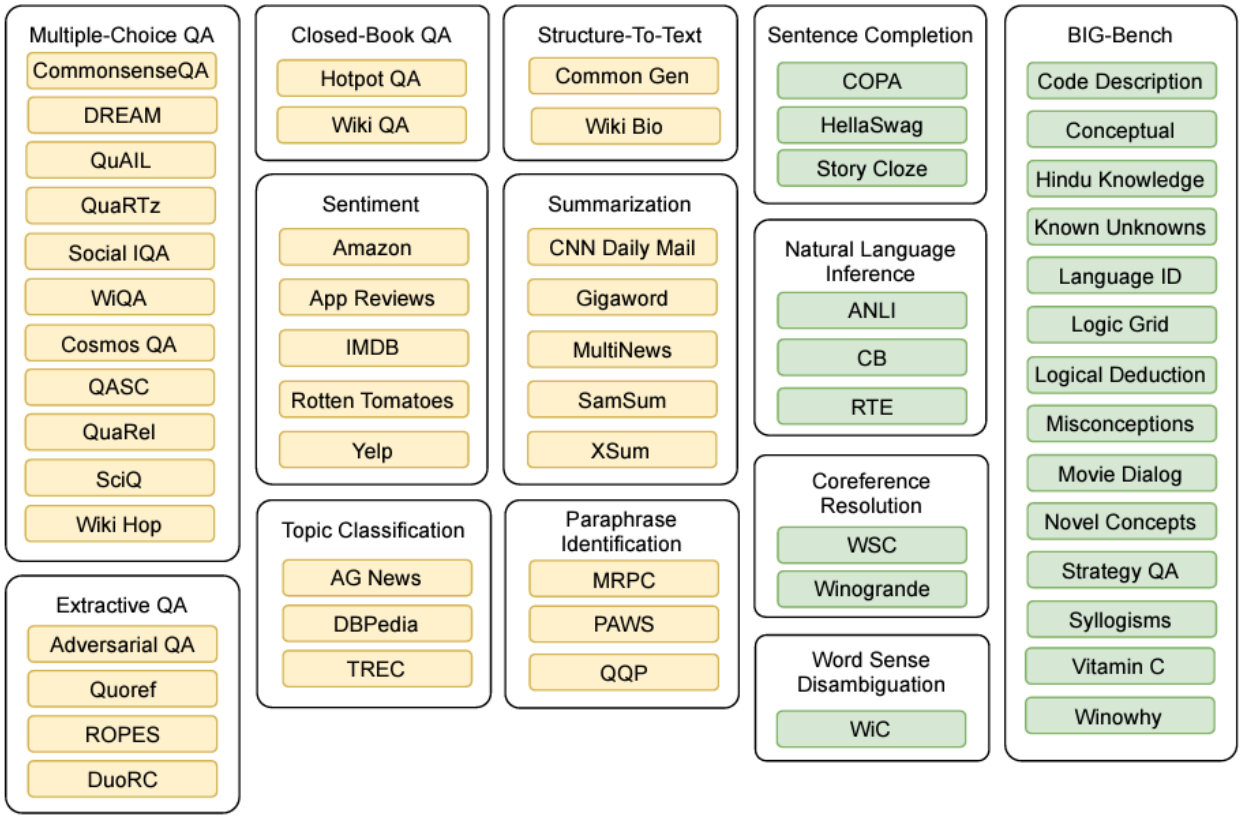


Figure 2: T0 datasets and task taxonomy. (T0+ and T0++ are trained on additional datasets. See Table 5 for the full list.) Color represents the level of supervision. Yellow datasets are in the training mixture. Green datasets are held out and represent tasks that were not seen during training. Hotpot QA is recast as closed-book QA due to long input length.

1. Pre-Requisites

- BigScience Workshop
- T5
- Multi-task learning
- Instruction Tuning

Pre-Requisites : BigScience Workshop

BigScience is an open science project composed of hundreds of researchers around the world.

Models

- T0
- 13B English decoder model
- BLOOM: BigScience 176B Model

Datasets

- BigScience Data Catalogue
- P3 Prompting dataset
- Masader

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

- **BigScience is not a consortium nor an officially incorporated entity.**
  - An open collaboration boot-strapped by HuggingFace, GENCI and IDRIS…
  - It gathers academic, industrial and independent researchers from many affiliations and whose research interests span many fields of research across AI, NLP, social sciences, legal, ethics and public policy.

Pre-Requisites : T5

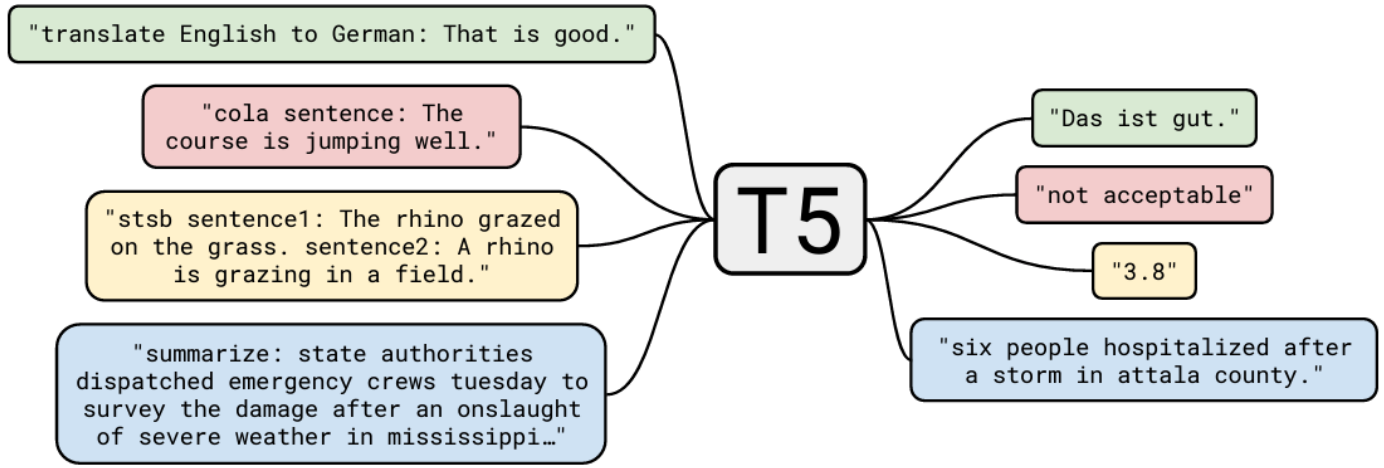


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." JMLR 21.1 (2020): 5485-5551.

- **"T**ext-**t**o-**T**ext **T**ransfer **T**ransformer"
  - All NLP tasks can be treated as text-to-text tasks.

Pre-Requisites : T5

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." JMLR 21.1 (2020): 5485-5551.

- **Model Structures**

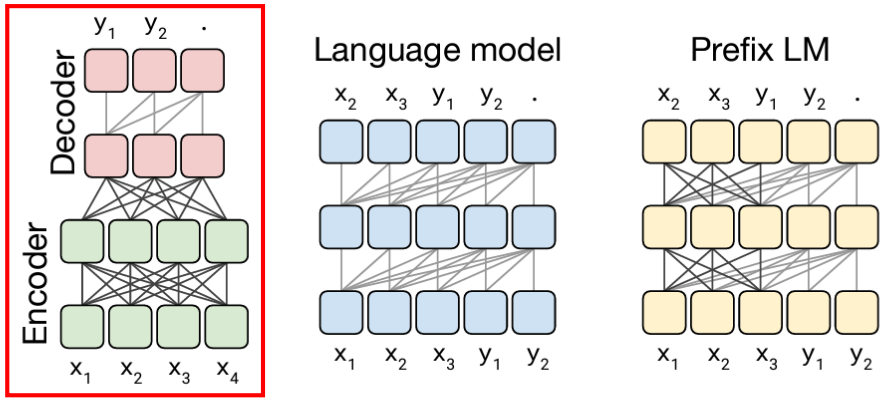


Figure 4: Schematics of the Transformer architecture variants we consider. as x and y respectively. Left: A standard encoder-decoder architecture uses fully-visible masking in the encoder and the encoder-decoder attention, with causal masking in the decoder. Middle: A language model consists of a single Transformer layer stack and is fed the concatenation of the input and target, using a causal mask throughout. Right: Adding a prefix to a language model corresponds to allowing fully-visible masking over the input.

- **Objectives**

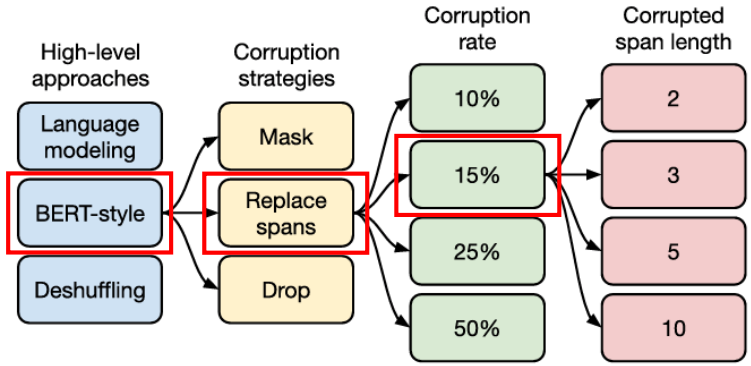
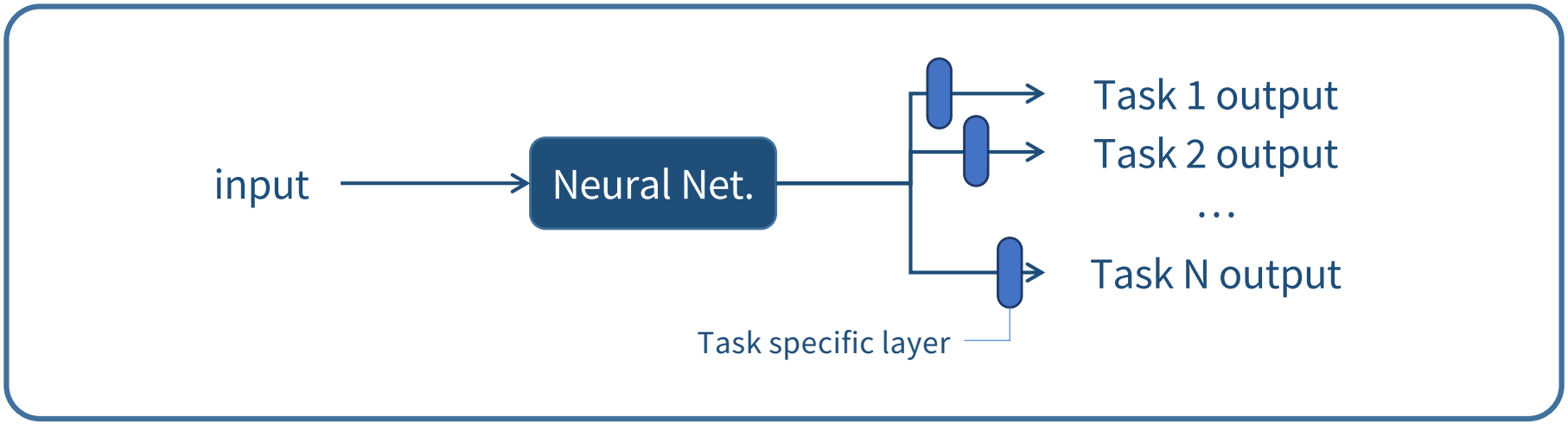


Figure 5: A flow chart of our exploration of unsupervised objectives. We first consider a few disparate approaches in Section 3.3.1 and find that a BERT-style denoising objective performs best. Then, we consider various methods for simplifying the BERT objective so that it produces shorter target sequences in Section 3.3.2. Given that replacing dropped-out spans with sentinel tokens performs well and results in short target sequences, in Section 3.3.3 we experiment with different corruption rates. Finally, we evaluate an objective that intentionally corrupts contiguous spans of tokens in Section 3.3.4.

Pre-Requisites : Multi-task learning

- **Multi-task learning**

  Ruder, Sebastian. "An overview of multi-task learning in deep neural networks." arXiv preprint arXiv:1706.05098 (2017).



> Advantage

- Knowledge Transfer
- Decreasing overfitting
- Computational efficiency

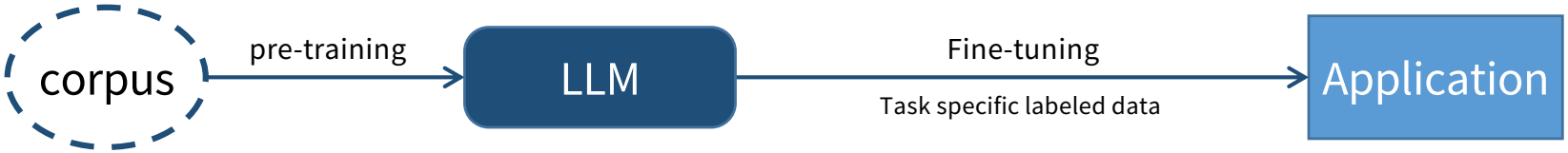
> Disadvantage

- Negative Transfer
- Hard to task balancing

Pre-Requisites : Instruction Tuning

- **Instruction tuning**
  - Fine-tuning various NLP tasks by transforming them into natural language instructions.

* **Vanilla Transfer learning** : BERT, T5, …



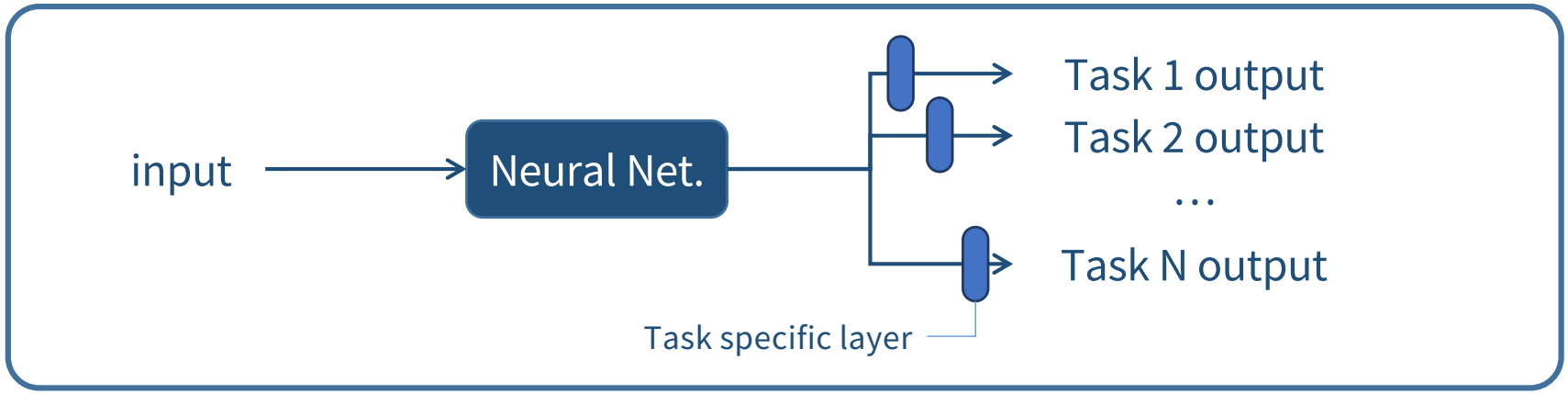
* **In-context learning** : GPT-3, …

* **Instruction tuning (=multi-task prompted training)** : T0, FLAN, …

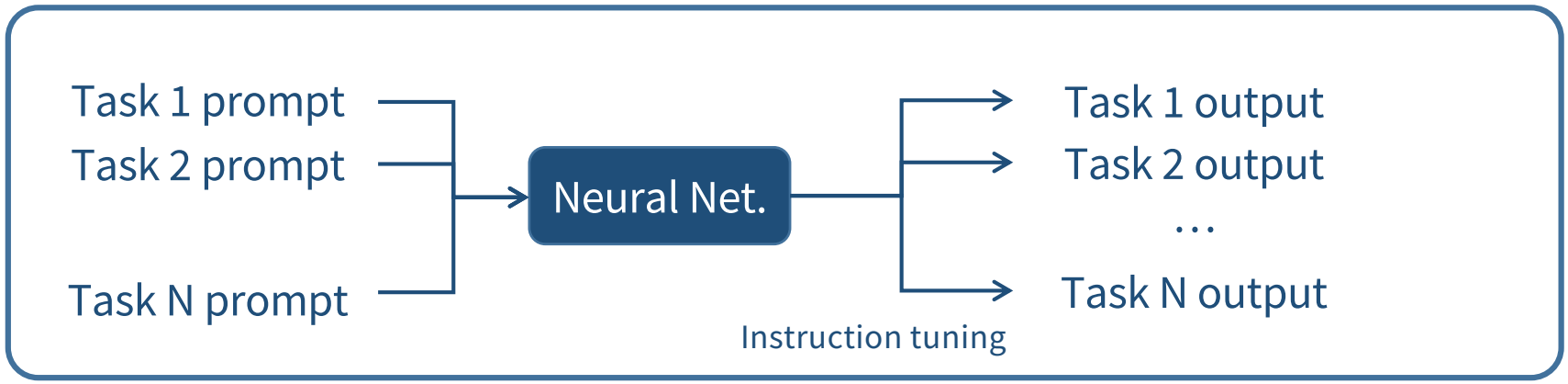
Pre-Requisites

- **Multi-task learning w/ instruction tuning**

<Multi-task learning w/ vanilla transfer-learning>



<Multi-task learning w/ instruction tuning>



T0
: Multitask Prompted Training Enables Zero-Shot Task Generalization

Victor Sanh, Albert Webson,
Colin Raffel, Stephen H. Bach et al.

ICLR 2022

Yejin Yoon

Contents

2. Introduction

- TL; DR
- Background
- Problem States

TL; DR

- **Multi-task prompted training can enable strong zero-shot generalization abilities in LM**
- **Demonstrating ablation study for robustness of prompt wording**
- **Releasing all prompt template and model**
  - T0 model
  - P3 prompt

Background

- **Implicit multi-task learning**
  - LLM have recently been shown to perform reasonable zero-shot generalization on a diverse set of NLP tasks
    - Being trained on only language modeling objectives, LM can perform relatively well at new tasks that they have not been explicitly trained to perform
  - *Hypothesis*: LLM generalize to new tasks as a result of an implicit process of multi-task learning
    - By learning to predict new word, LLM is forced to learn from mixture of implicit tasks included in their pretraining corpus
  - This ability requires **a sufficiently large model** and is **sensitive to the wording of its prompts**

- **Explicit multi-task learning**
  - Explicitly training language model in a supervised and massively multi-task style
  - *Hypothesis*: Multitask supervision in pretraining plays a large role in zero-shot generalization

Problem States

- How to make **a model to better generalize to unseen tasks ?**
  (1) w/o requiring massive scale
  (2) as well as being more robust to the wording choices of the prompts

- **RQ1: Does multi-task prompted training improve generalization to unseen task?**

- **RQ2: Does training on a wider range of prompts improve robustness to prompt wording?**

3. Main Points: Method

- T0
- P3

Methods: T0

- **T0: encoder-decoder model that takes textual inputs and produces target responses**

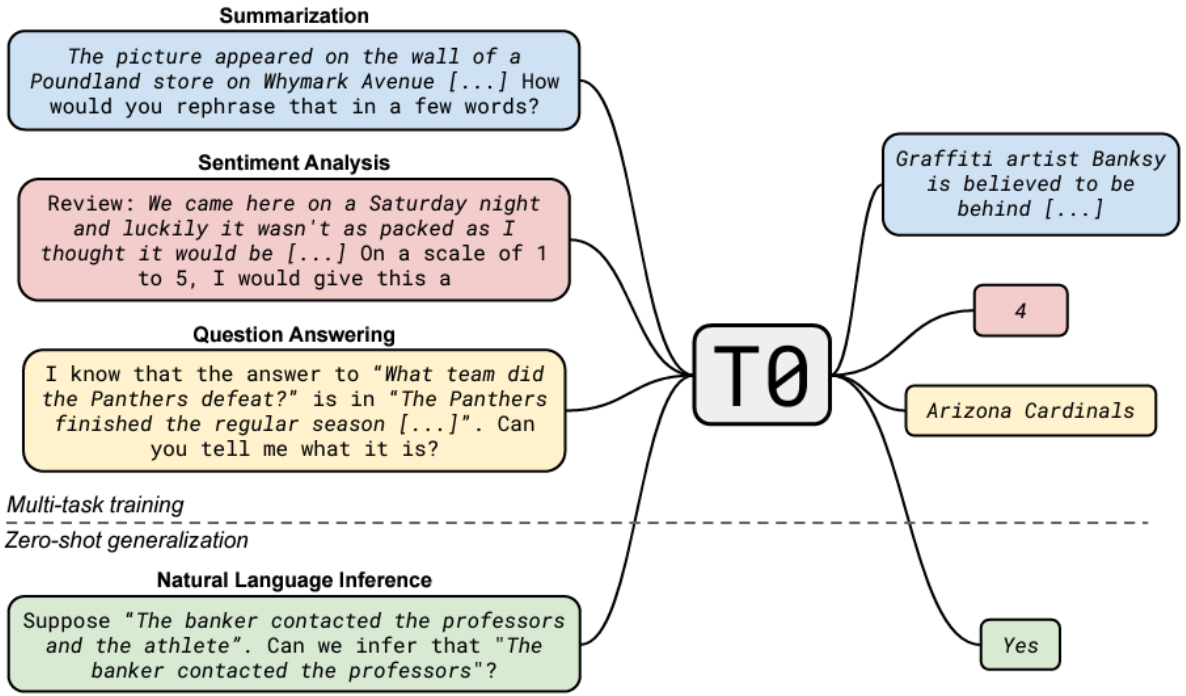


Figure 1: Our model and prompt format. T0 is an encoder-decoder model that consumes textual inputs and produces target responses. It is trained on a multitask mixture of NLP datasets partitioned into different tasks. Each dataset is associated with multiple prompt templates that are used to format example instances to input and target pairs. Italics indicate the inserted fields from the raw example data. After training on a diverse mixture of tasks (top), our model is evaluated on zero-shot generalization to tasks that are not seen during training (bottom).

Methods: T0

- **T0: an encoder-decoder model architecture with input text fed to the encoder and target text produced by the decoder**

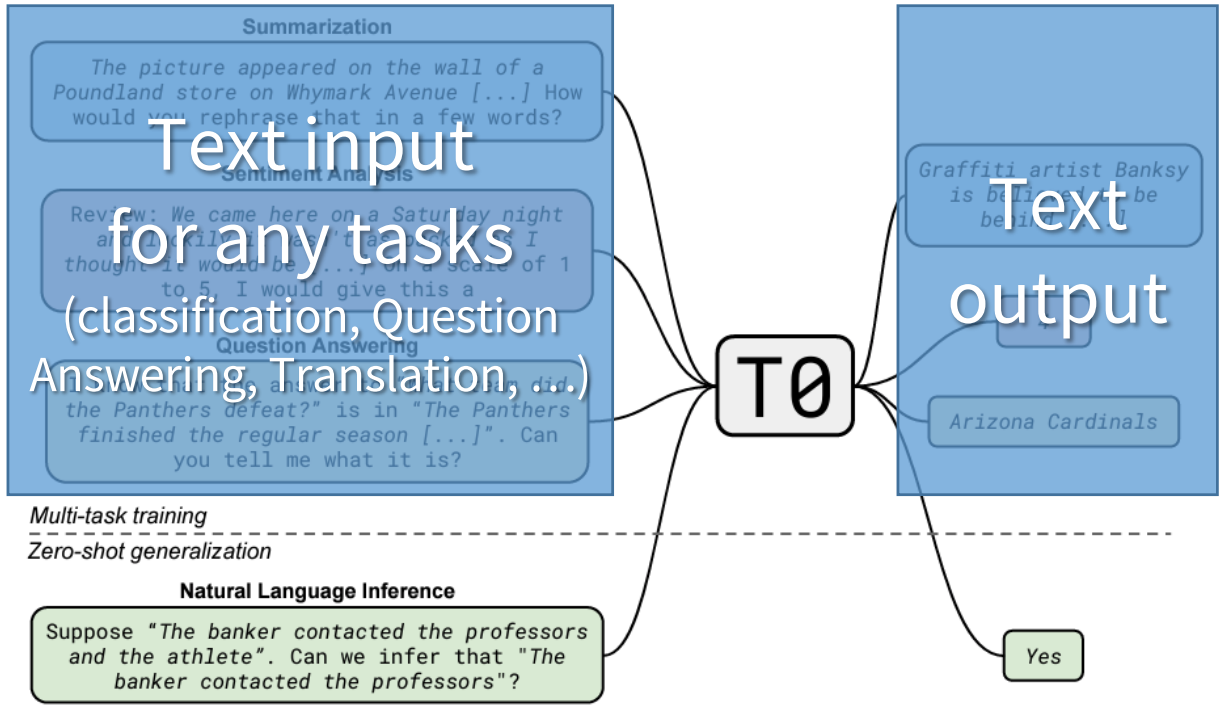
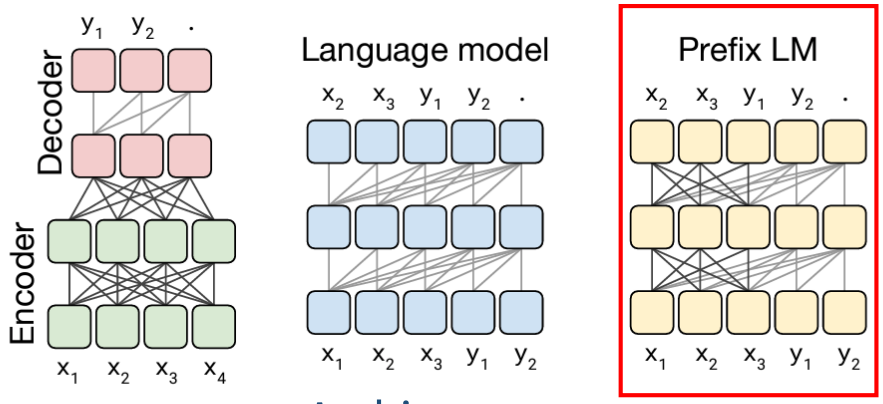


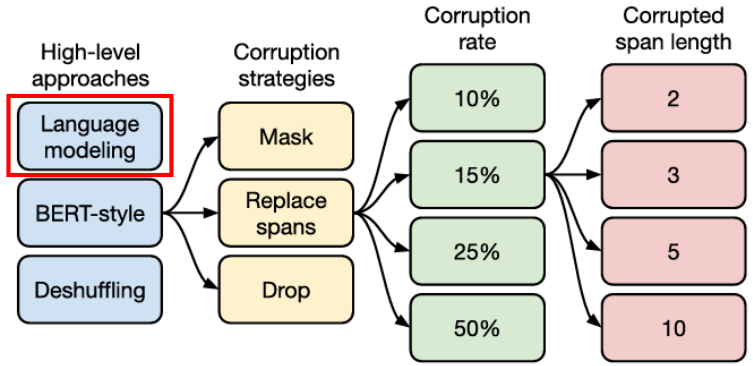
Figure 1: Our model and prompt format. T0 is an encoder-decoder model that consumes textual inputs and produces target responses. It is trained on a multitask mixture of NLP datasets partitioned into different tasks. Each dataset is associated with multiple prompt templates that are used to format example instances to input and target pairs. Italics indicate the inserted fields from the raw example data. After training on a diverse mixture of tasks (top), our model is evaluated on zero-shot generalization to tasks that are not seen during training (bottom).

Methods: T0

- **T0: an encoder-decoder model architecture with input text fed to the encoder and target text produced by the decoder**
  - Architecture: based on LM-adapted T5 model (referred to as **T5+LM**)



< Architecture >



<Objective >

*** Prefix LM example**

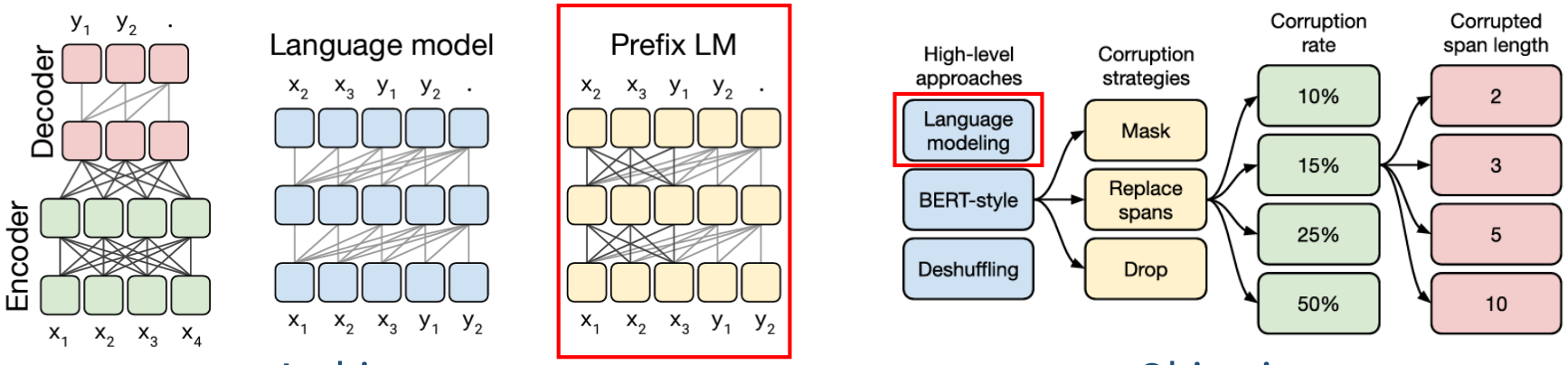
Translate English to Korean : How much is it? Target: 얼마예요?

Prefix
bi-directional attention

Target
uni-directional attention

Methods: T0

- **T0: an encoder-decoder model architecture with input text fed to the encoder and target text produced by the decoder**
  - Architecture: based on LM-adapted T5 model (referred to as **T5+LM**)



< Architecture > <Objective >

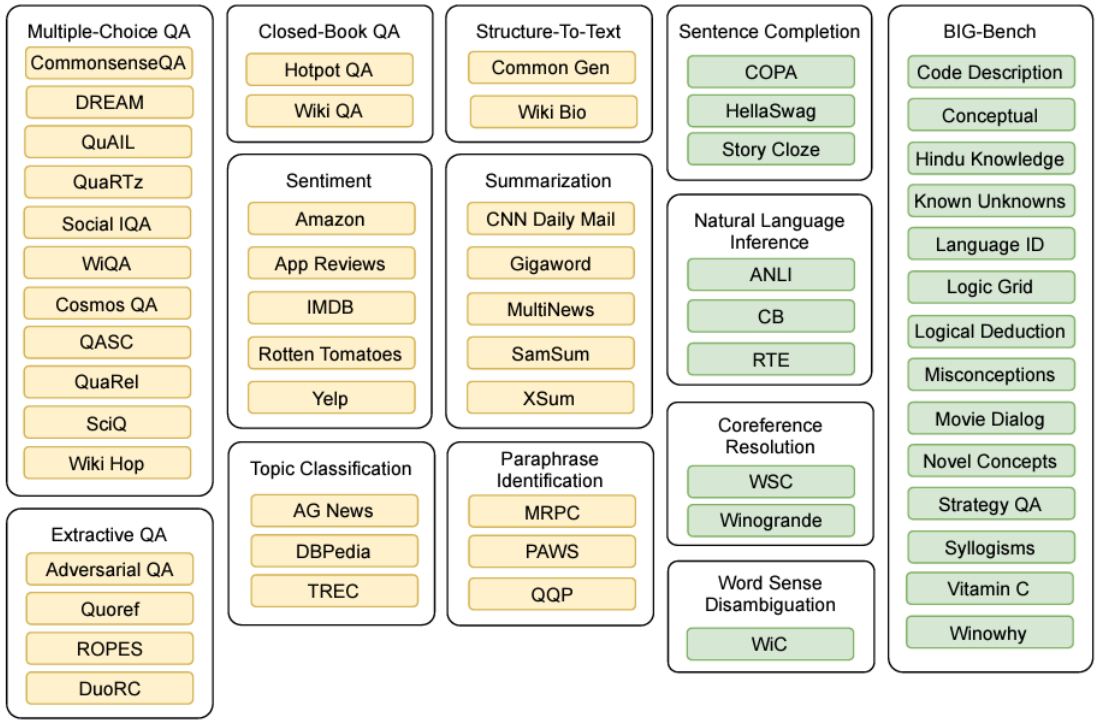
  - Objective: standard maximum likelihood loss
    * Never trained to generate the input
  - Fine-tune a PLM on multi-task training mixture of natural language prompted dataset

Methods: T0

- **Training Details**
  - Model variants
    - T5+LM: baseline, T5 on 100B additional tokens from C4 on a standard language modeling objective (11B parameters)
    - T0: trained on the multi-task mixture
    - T0+: T0 + GPT-3's evaluation datasets
    - T0++: T0 + GPT-3's evaluation datasets + SuperGLUE (except held-out tasks)
  - Checkpoint selection: highest score on the validation splits
    - For satisfying true zero-shot setting, do not use any examples form any of the held-out tasks to select the best checkpoints
  - Optimizer: Adafactor
  - Batch: 1024
  - Learning rate: 12-3
  - Dropout rate: 0.1

Methods: T0

- **Dataset**
  - 12 tasks and 62 datasets with publicly contributed prompts in mixtures
    - Training mixture
      Multiple Choice QA,
      Close-Book QA,
      Structure-To-Text,
      Sentiment Analysis,
      Summarization,
      Extractive QA,
      Topic Classification,
      Paraphrase Identification
    - Held-out* mixture
      Sentence Completion,
      BIG-Bench, NLI,
      Coreference Resolution,
      Word Sense Disambiguation

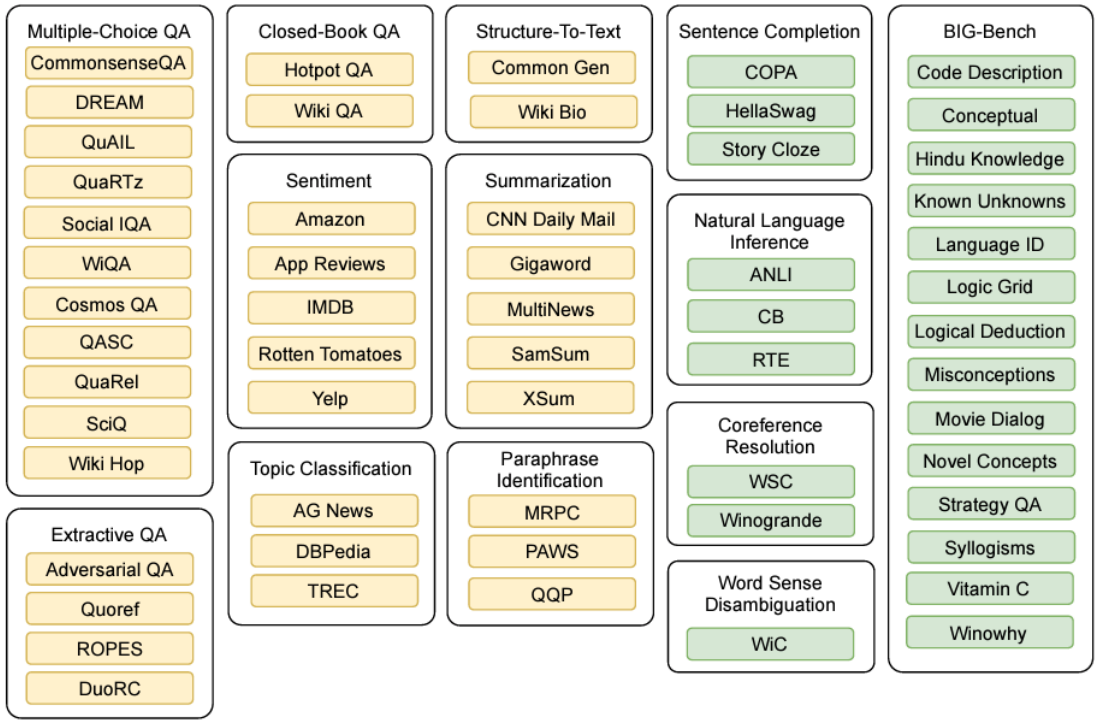


* Held-out task: zero-shot evaluation tasks (=unseen tasks)

Figure 2: T0 datasets and task taxonomy. (T0+ and T0++ are trained on additional datasets. See Table 5 for the full list.) Color represents the level of supervision. Yellow datasets are in the training mixture. Green datasets are held out and represent tasks that were not seen during training. Hotpot QA is recast as closed-book QA due to long input length.

Methods: T0

- **Dataset**
  - 12 tasks and 62 datasets with publicly contributed prompts in mixtures
    - Training mixture
      Multiple Choice QA, Close-Book QA, Structure-To-Text, Sentiment Analysis, Summarization, Extractive QA, Topic Classification, Paraphrase Identification
    - Held-out* mixture
      Sentence Completion, BIG-Bench, NLI, Coreference Resolution, Word Sense Disambigua

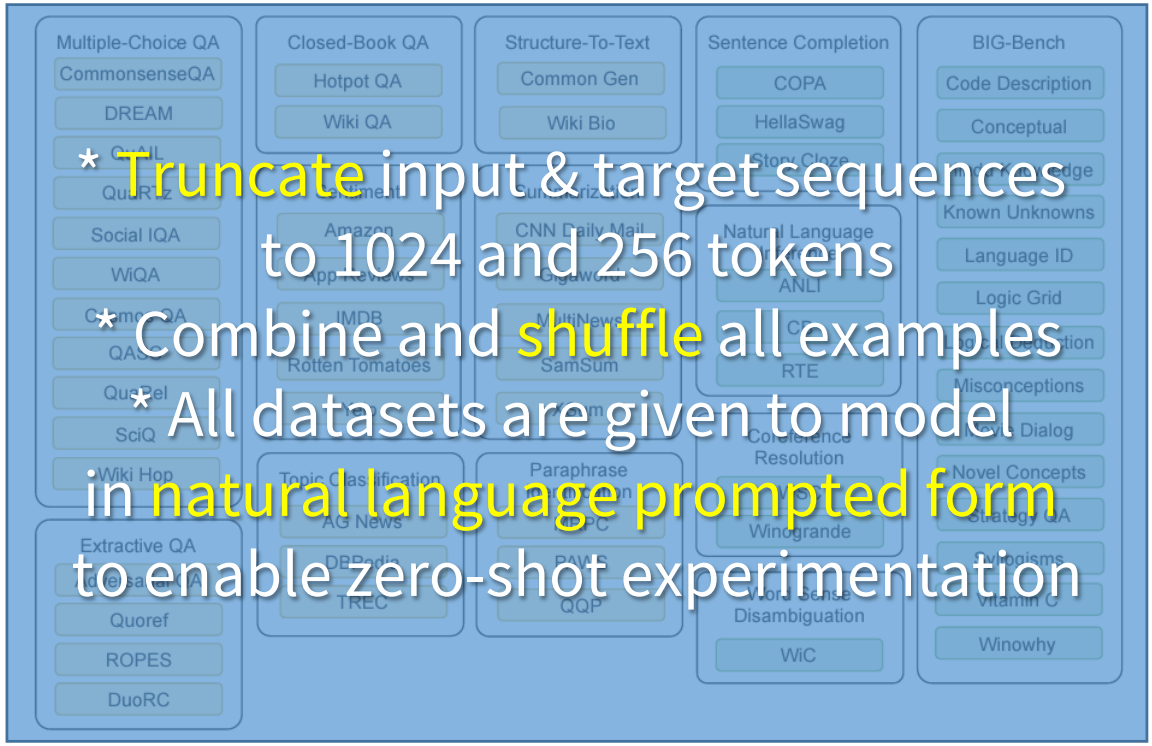


*** NLI example**

Premise: 흡연자는 발코니에서 흡연이 가능합니다.
Hypothesis: 어떤 방에서도 흡연은 금지됩니다.
{ Label: entailment, neutral, contradiction }

Methods: T0

- **Dataset**
  - 12 tasks and 62 datasets with publicly contributed prompts in mixtures
    - Training mixture
      Multiple Choice QA,
      Close-Book QA,
      Structure-To-Text,
      Sentiment Analysis,
      Summarization,
      Extractive QA,
      Topic Classification,
      Paraphrase
      Identification
    - Held-out* mixture
      Sentence Completion,
      BIG-Bench, NLI,
      Coreference Resolution,
      Word Sense Disambigua



*** NLI example**

Premise: 흡연자는 발코니에서 흡연이 가능합니다.
Hypothesis: 어떤 방에서도 흡연은 금지됩니다.
{ Label: entailment, neutral, contradiction }

Methods: Public Pool of Prompts

- **Unified Prompt Format (w/ 36 contributors)**

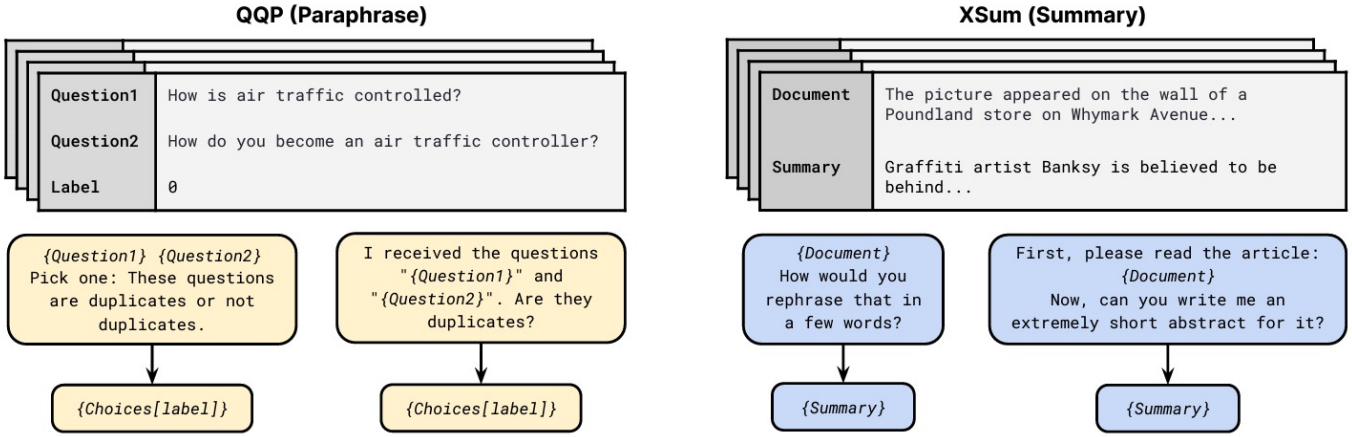


Figure 3: Prompt templates from the P3 prompt collection. Each dataset has multiple prompt templates consisting of an input and a target template. These use the fields of the raw data examples as well as template metadata, e.g., the left paraphrasing identification prompts use *Choices*, a template-level list variable `['Not duplicates', 'Duplicates']`. These templates are materialized to produce the prompted instance shown in Figure 1. The complete set of prompt templates used in T0 is given in Appendix G.

- Prompt: consist of an input template and a target template along with a collection of associated meta-data* (*meta-data: labels, …)
- Develop an application that make it easy to convert diverse datasets into prompts to facilitate writing a large collection of prompts

Methods: Public Pool of Prompts

- **Unified Prompt Format (w/ 36 contributors)**
  - Prompt: consist of an input template and a target template along with a collection of associated meta-data* (*meta-data: labels, …)
  - Develop an **application** that make it easy to convert diverse datasets into prompts to facilitate writing a large collection of prompts
  - Prompts were both formal and creative and various ordering of the data
    - As the question “what makes a prompt effective” is not solved, the project team encouraged contributors to be open in their **style and create a diverse set of prompts**
  - Prompts that required explicit counting or numerical indexing* were **removed** in favor of natural language variants * e.g. span extraction
  - Only **English** dataset is included, excluding ones that included potentially harmful content or non-natural language such as programing languages
    - 2,073 prompts for 177 datasets (11.7 prompts per dataset on average)
  - Original-task prompt vs. **Non-original-task** prompt
    - Some of the prompts correspond directly to a version of the original proposed task, others also are allowed to permuted the original task

Methods: Public Pool of Prompts

- **Structure to Text** (Lin et al. 2020): COMMON_GEN
  - Data Example

Key	Value
concept_set_idx	0
concepts	['ski', 'mountain', 'skier']
target	Skier skis down the mountain

  - Original-task prompt

```
# Ignoring the order of the concepts: {{ concepts | join(", ") }};
  Generate a sentence with all the concepts :
```

  - Non-original-task prompt

```
# What are the topics in the sentence: {{target}}
```

```
# We have the sentence: {{target}};
  Extract all the key concepts:
```

- Original-task prompt vs. **Non-original-task** prompt
  - Some of the prompts correspond directly to a version of the original proposed task, others also are allowed to permuted the original task

4. Effects: Experiments

- Evaluation
- RQ1
- RQ2

Evaluation

- **Evaluate zero-shot generalization on 11 datasets in 4 held-out NLP tasks**
  - Natural language inference, Coreference resolution, Word sense disambiguation, Sentence completion

  + 14 novel tasks from BIG-bench

- **All reported datasets use accuracy as their metric**
  - Some tasks that involve choosing the correct completion from several options (e.g. multi choice QA) use rank classification
    - Do Not apply length normalization to the log-likelihoods of the target options

- **Do <u>NOT</u> perform prompt selection by comparing the performance of different prompts on the validation split for "true" zero-shot setting**
  - Report **median** performance across all prompts for dataset along with their **interquartile range** to measure the model's robustness to the wording of the prompts

RQ1 Does multi-task prompted training improve generalization to unseen task?

- **Held-out tasks**

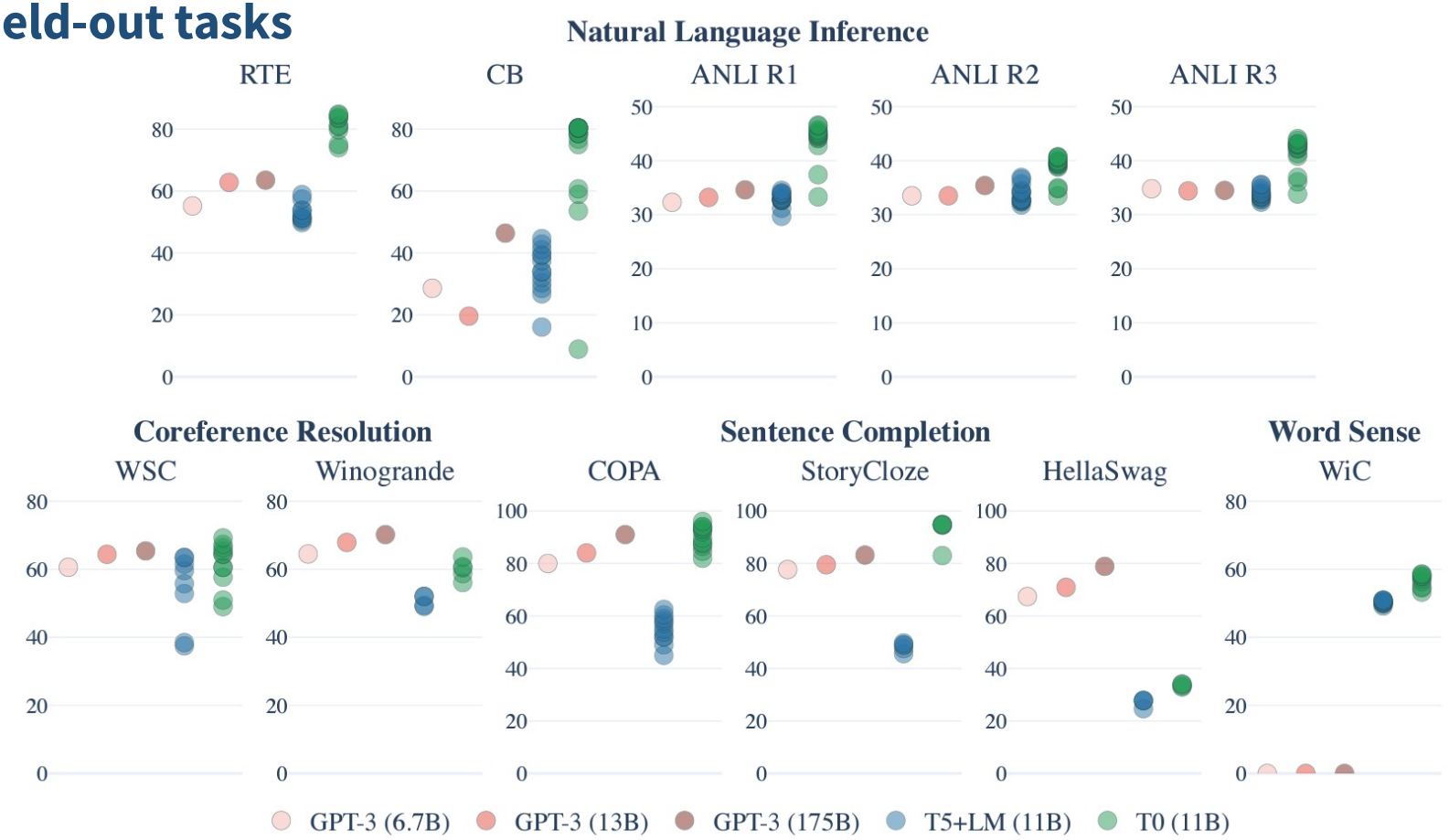
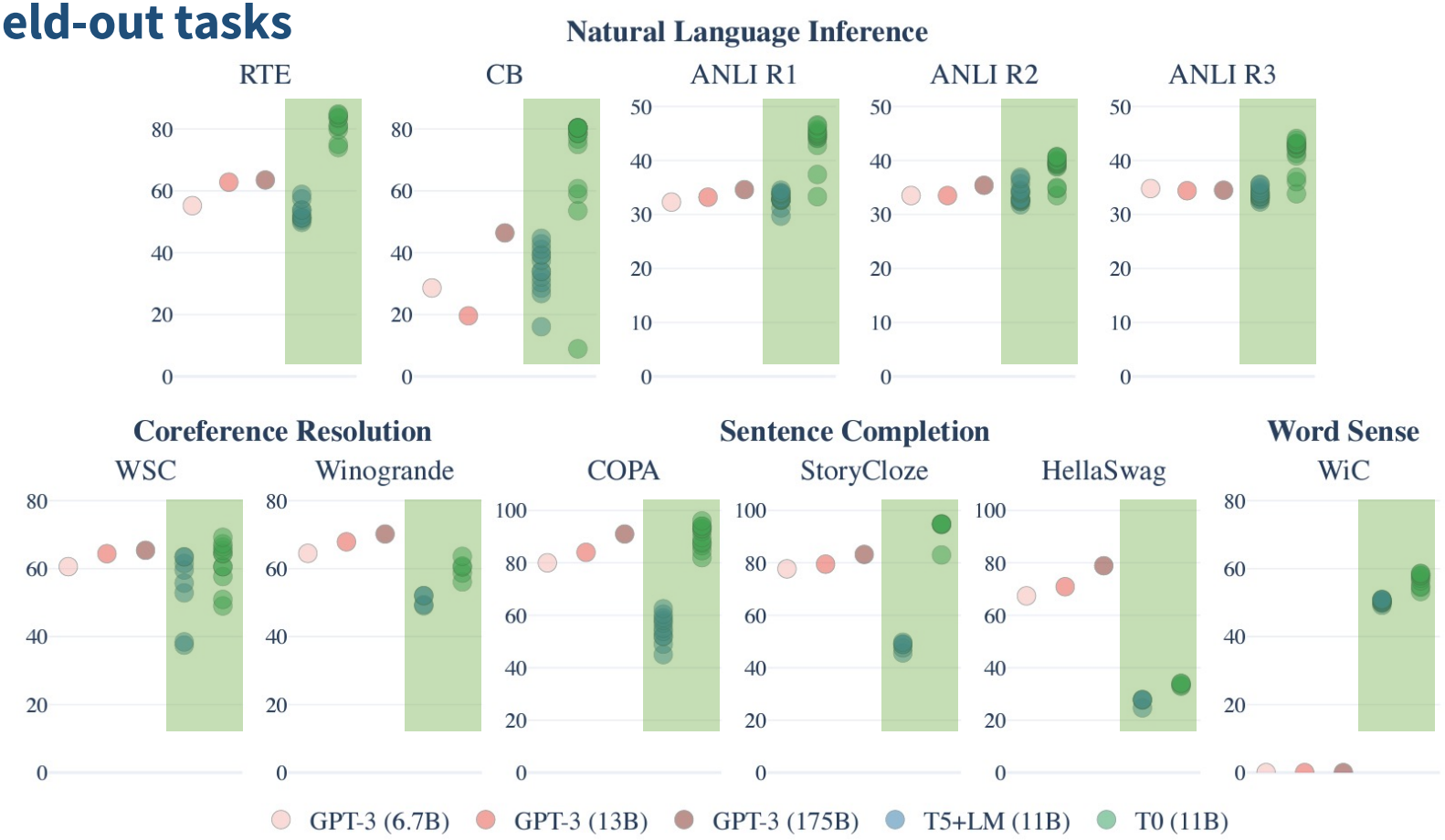


Figure 4: Results for T0 task generalization experiments compared to GPT-3 (Brown et al., 2020). Each dot is the performance of one evaluation prompt. The baseline T5+LM model is the same as T0 except without multitask prompted training. GPT-3 only reports a single prompt for each dataset.

RQ1 Does multi-task prompted training improve generalization to unseen task?

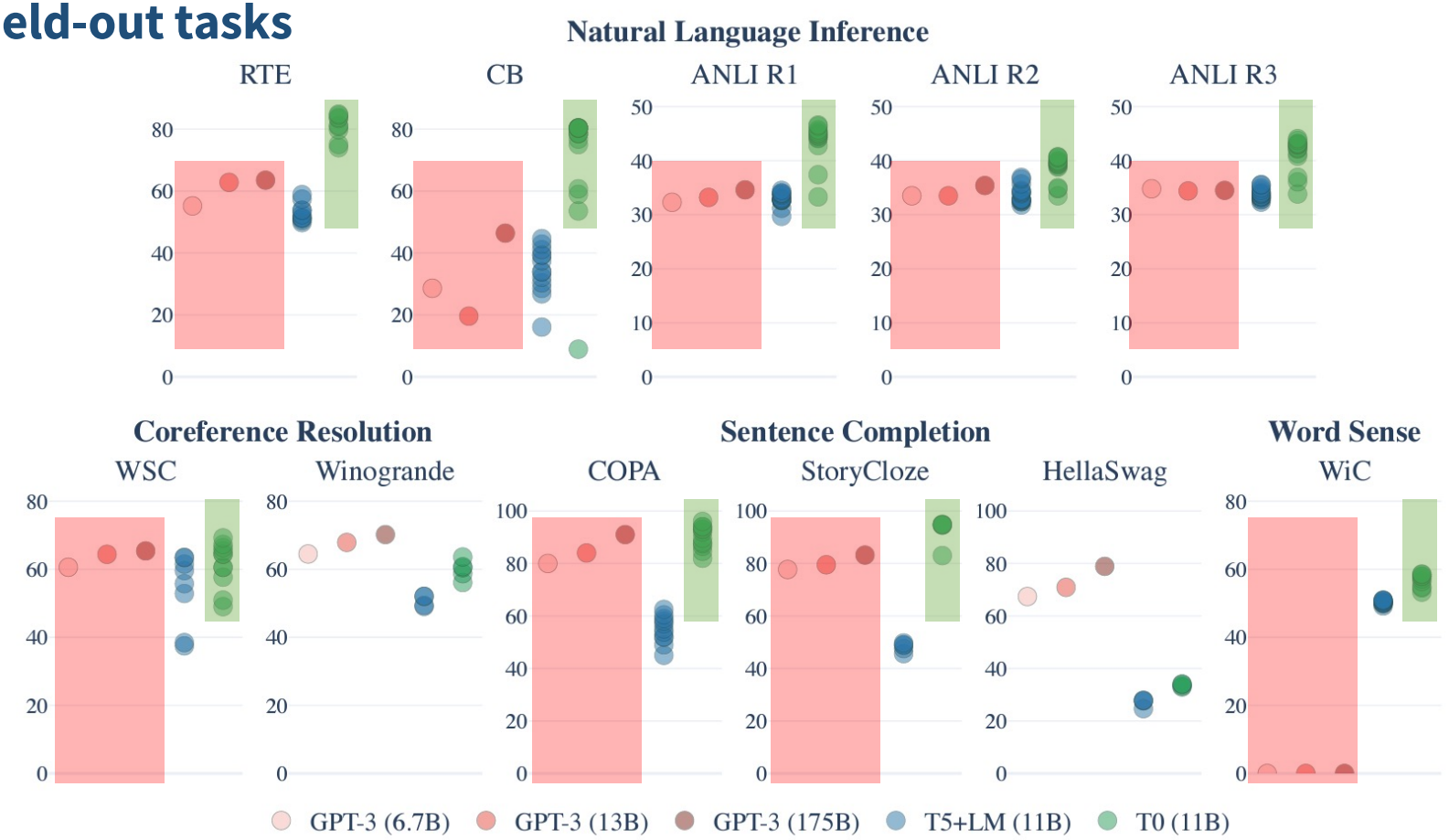
- **Held-out tasks**



T0 > T5+LM(baseline): multi-task prompted training improve generalization to unseen task

RQ1 Does multi-task prompted training improve generalization to unseen task?

- **Held-out tasks**

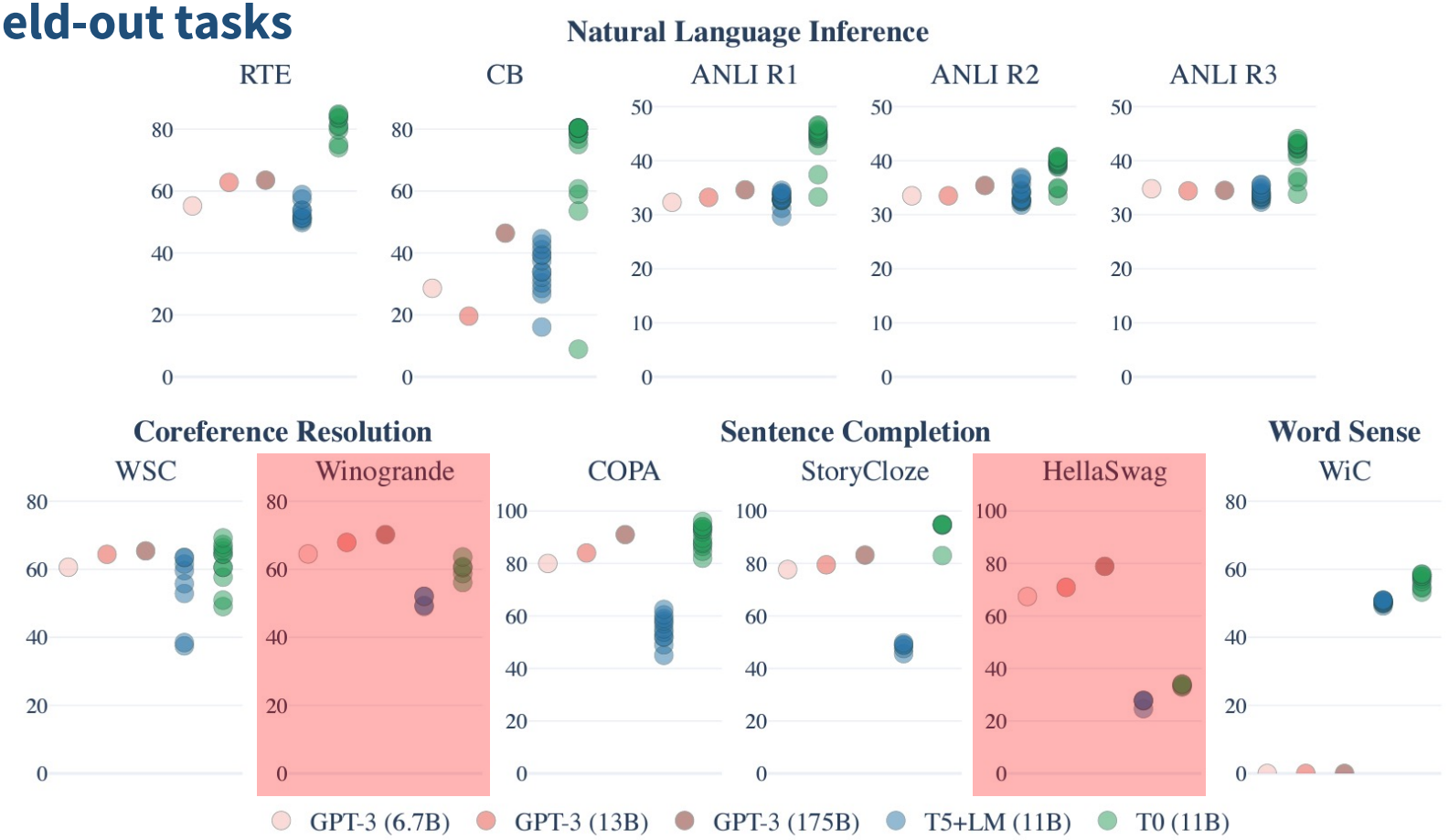


T0 > T5+LM(baseline): multi-task prompted training improve generalization to unseen task

11B T0 > 175B GPT-3: more efficient and effective

RQ1 Does multi-task prompted training improve generalization to unseen task?

- **Held-out tasks**



T0 > T5+LM(baseline): multi-task prompted training improve generalization to unseen task

11B T0 > 175B GPT-3: more efficient and effective

RQ1 Does multi-task prompted training improve generalization to unseen task?

- **BIG-bench**

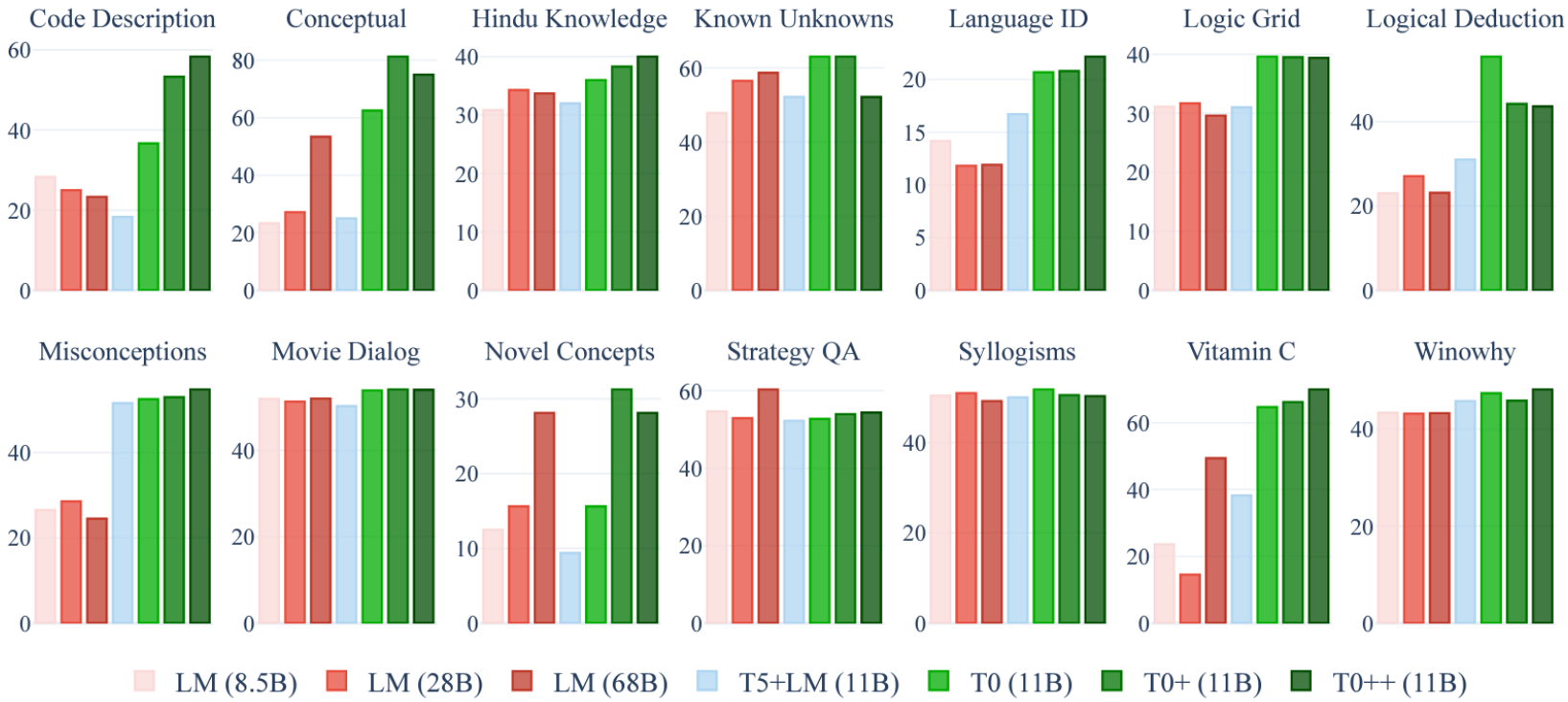


Figure 5: Results for a subset of BIG-bench which has available baselines. The baseline models are Transformer-based language models provided by BIG-bench maintainers, who also provide one prompt per dataset. T0, T0+ and T0++ are identical except for increasing the number of training datasets (§5). BIG-bench Tasks are all zero-shot for all the reported models.

RQ1 Does multi-task prompted training improve generalization to unseen task?

- **BIG-bench**

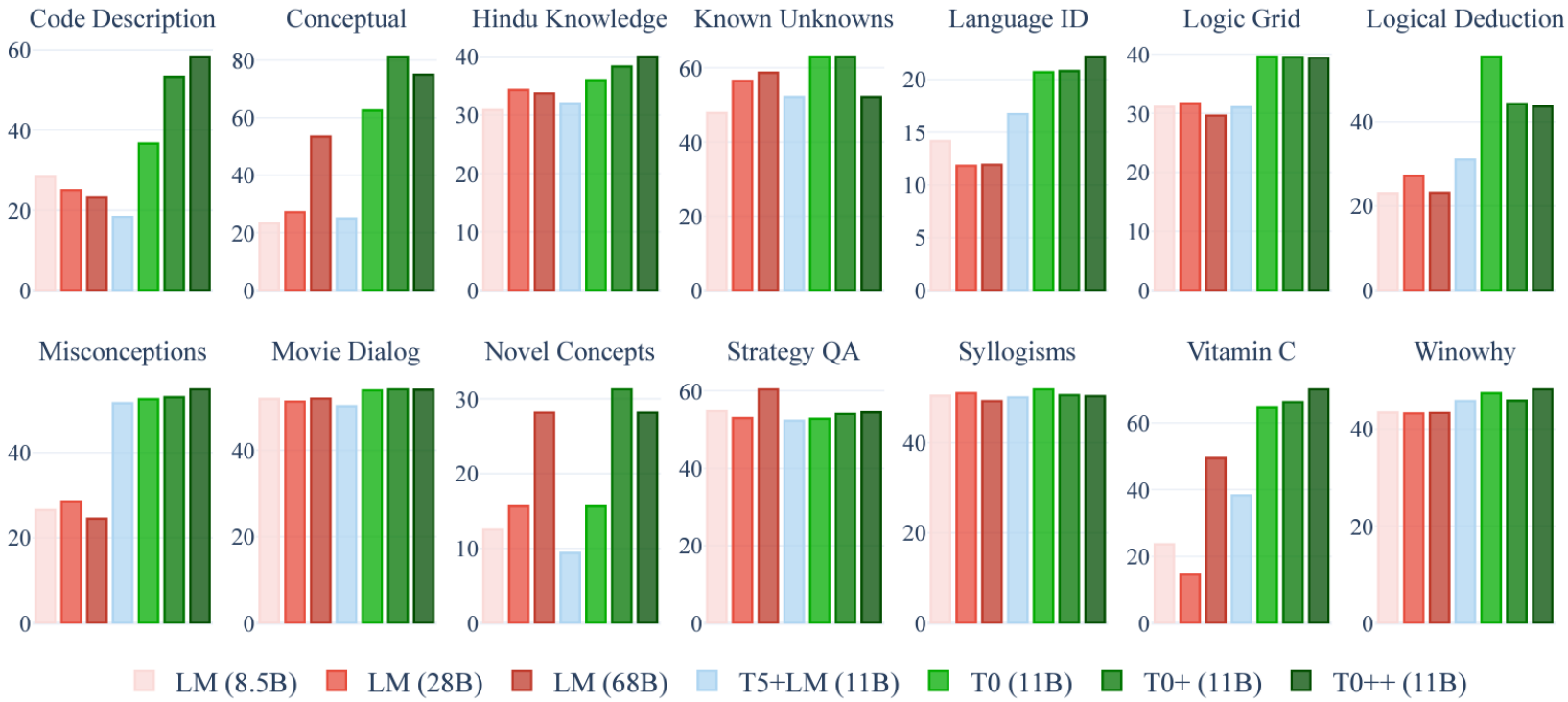
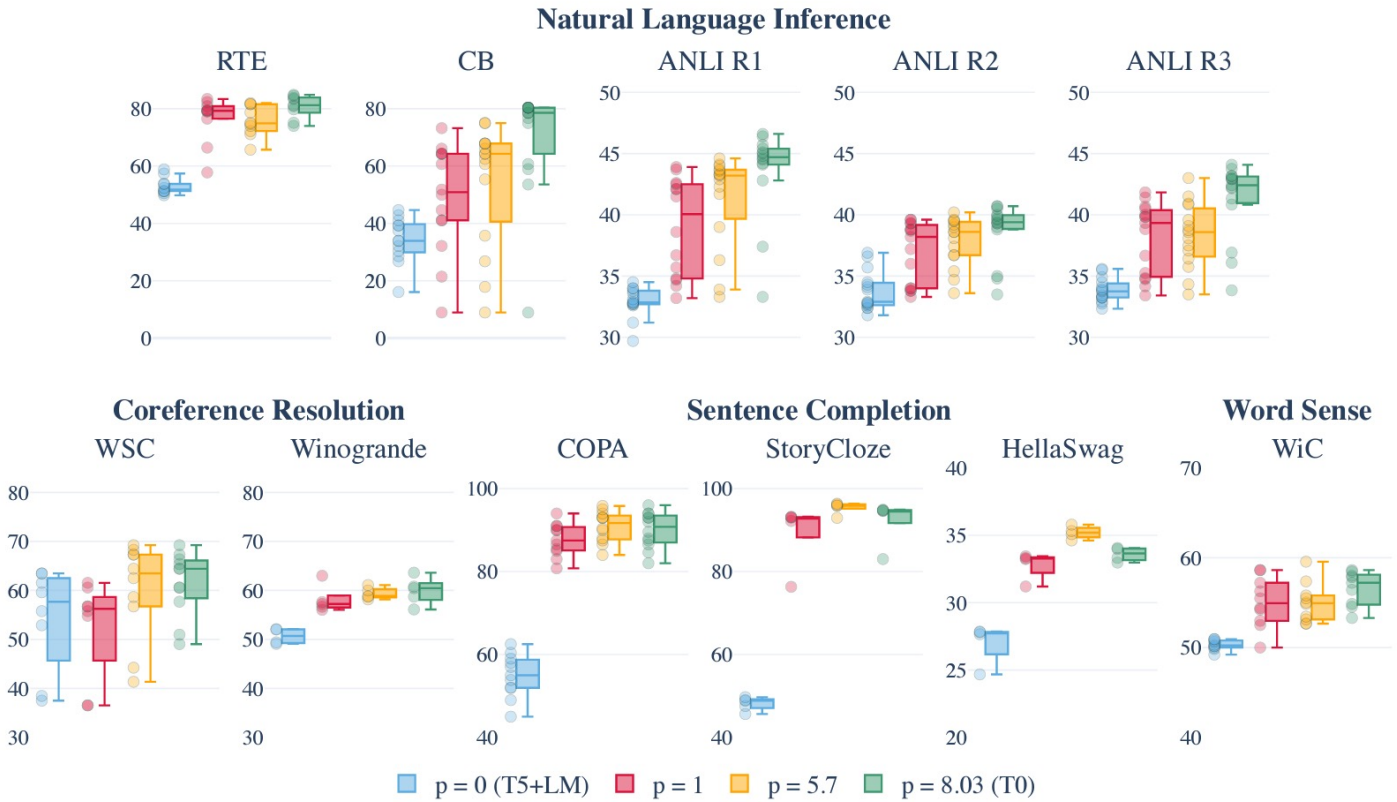


Figure 5: Results for a subset of BIG-bench which has available baselines. The baseline models are Transformer-based language models provided by BIG-bench maintainers, who also provide one prompt per dataset. T0, T0+ and T0++ are identical except for increasing the number of training datasets (§5). BIG-bench Tasks are all zero-shot for all the reported models.

T0 < T0+ < T0++

RQ2 Does training on a wider range of prompts improve robustness to prompt wording?

- **p (#prompt)**



p = 0 : No prompt training / p = 1 : Randomly sampled 1-prompt
p = 5.7 : All orig.-task prompt / p = 8.03 : Orig.-task + non-orig.-task

As the number of prompts (p) increases, performance also improves.

RQ2 Does training on a wider range of prompts improve robustness to prompt wording?

- **d (#dataset)**

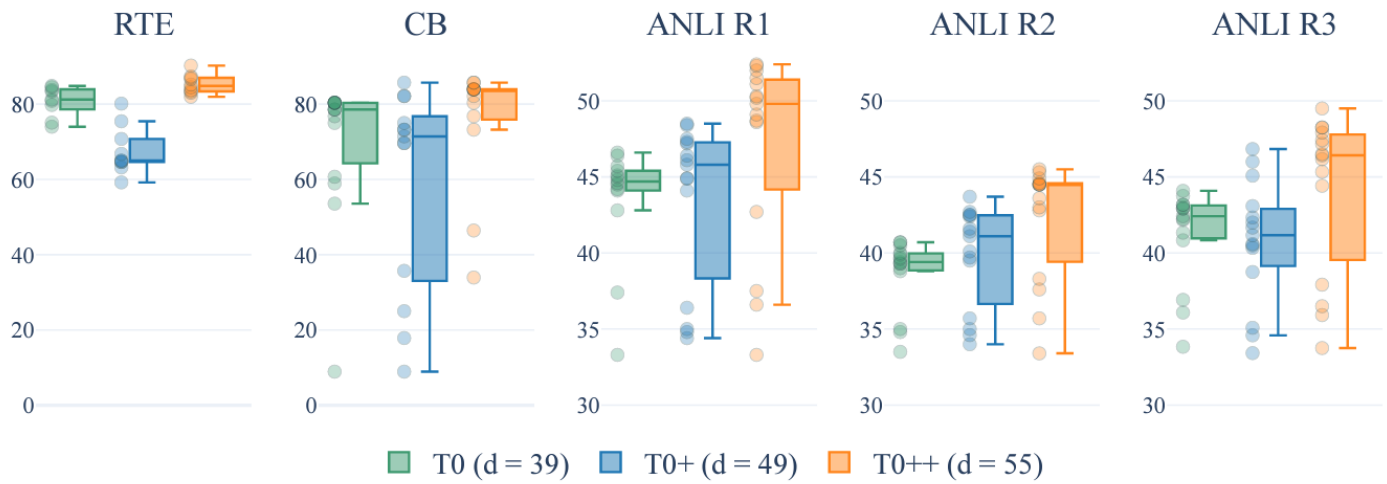


Figure 7: Effect of prompts from more datasets. Zero-shot performance of three models with varying number of datasets (T0, T0+, T0++). Adding more datasets consistently leads to higher median performance but does not always reduce interquartile range for held-out tasks.

As the number of prompts (p) increases, performance also improves.

It is difficult to observe any improvement in performance with the increasing number of datasets (d).

5. Comparing FLAN & T0

- FLAN
- A comparative study

FLAN

Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (2021).

- **FLAN**
  - Improving Zero-shot Generalization through Multi-task Prompted Training

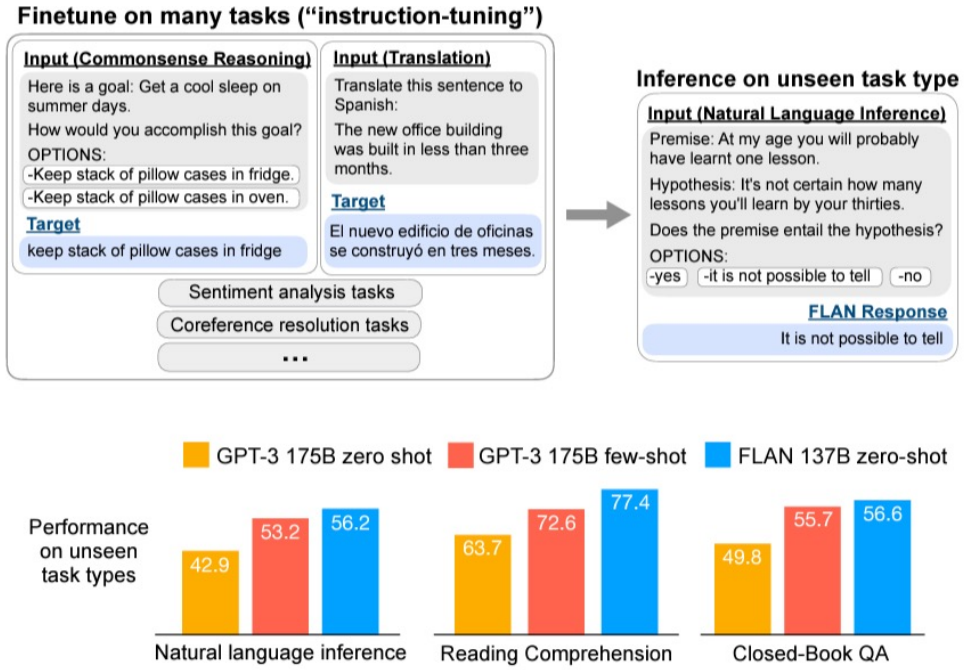


Figure 1: Top: overview of instruction tuning and FLAN. Instruction tuning finetunes a pretrained language model on a mixture of tasks phrased as instructions. At inference time, we evaluate on an unseen task type; for instance, we could evaluate the model on natural language inference (NLI)

FLAN vs. T0

- **FLAN vs. T0**

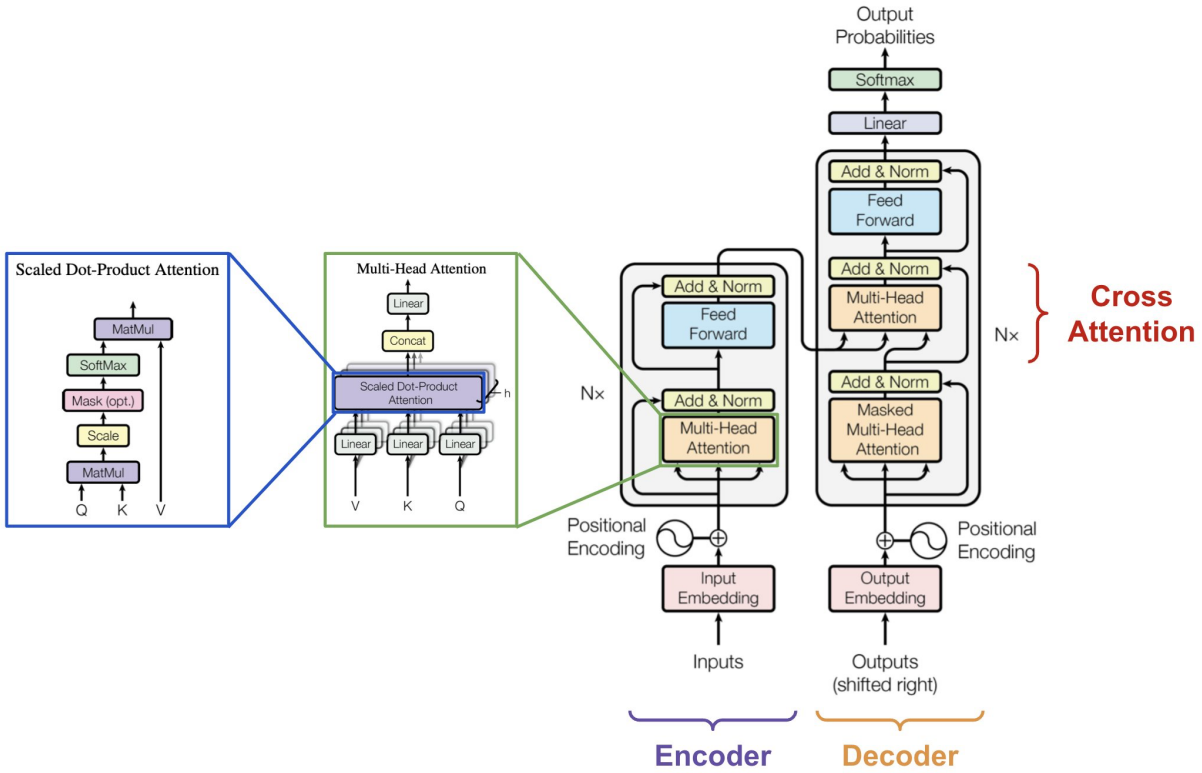
  < Backbone architecture >

  - Transformer-based

> FLAN

> T0

  - Encoder-decoder

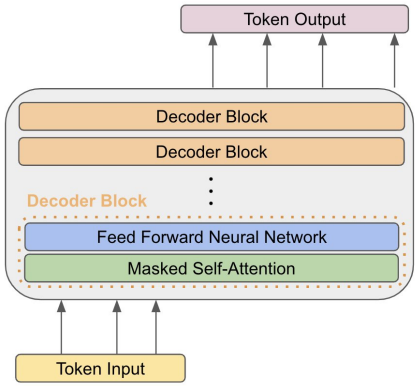


FLAN vs. T0

- **FLAN vs. T0**

  < Backbone architecture >

  - Transformer-based

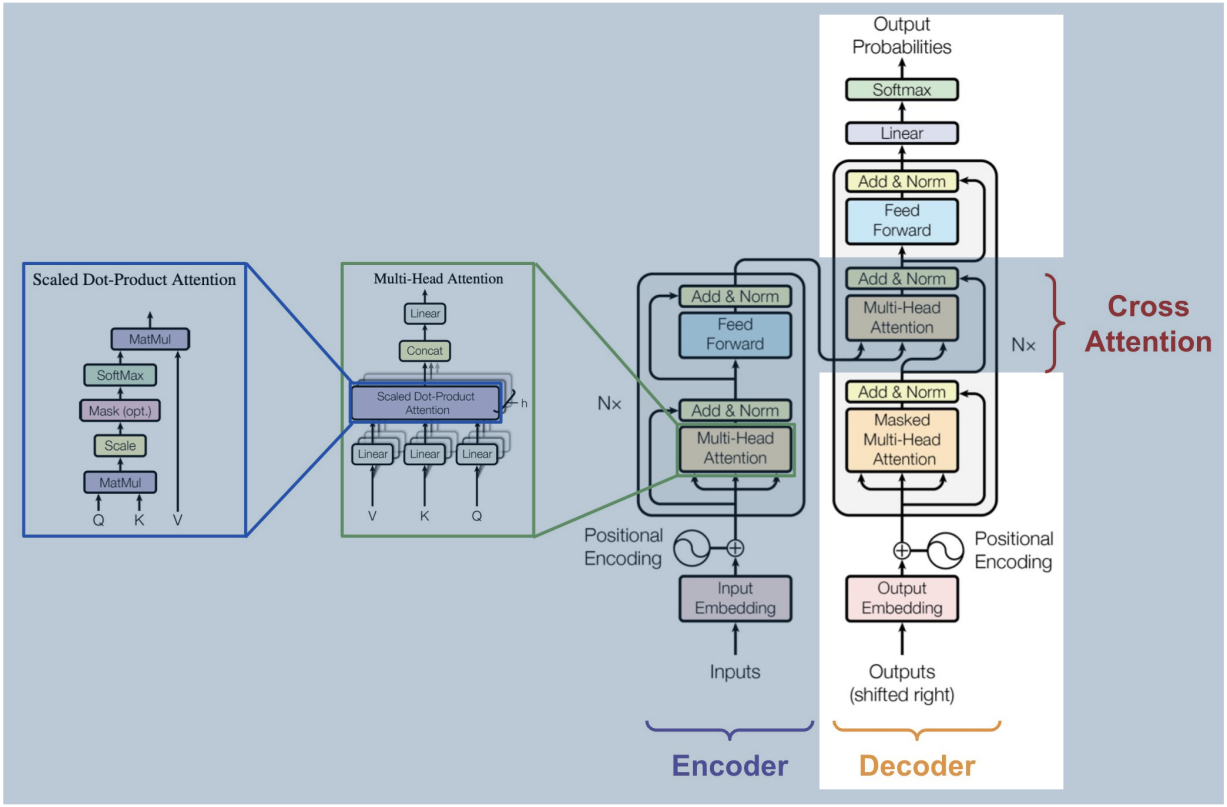


> FLAN

  - Decoder-only

> T0

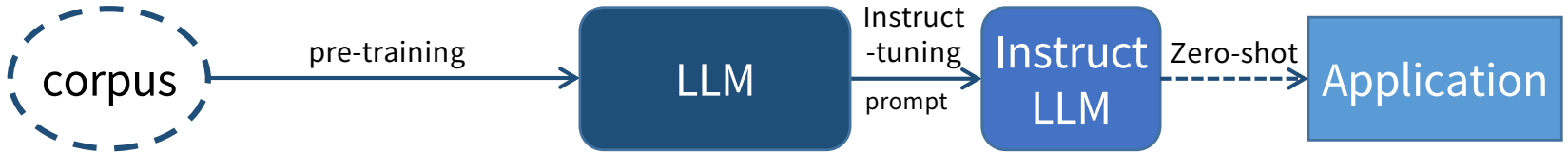
  - Encoder-decoder



FLAN vs. T0

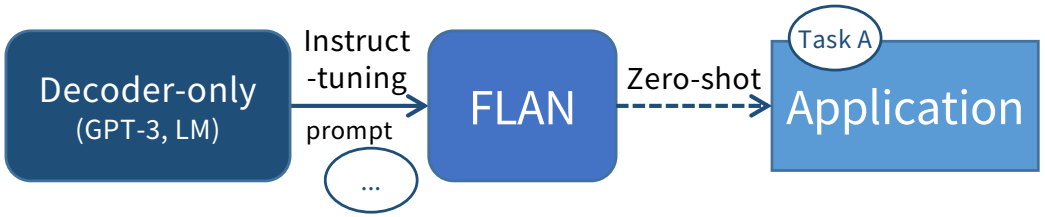
- **FLAN vs. T0**

< Instruction tuning >



> FLAN

- Decoder-only
- Single held-out task



> T0

- Encoder-decoder
- Multi held-out task

FLAN vs. T0

- **FLAN vs. T0**

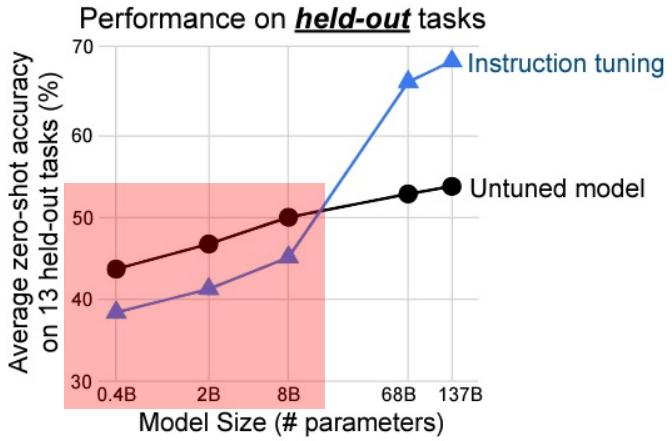
  < Model Scale >

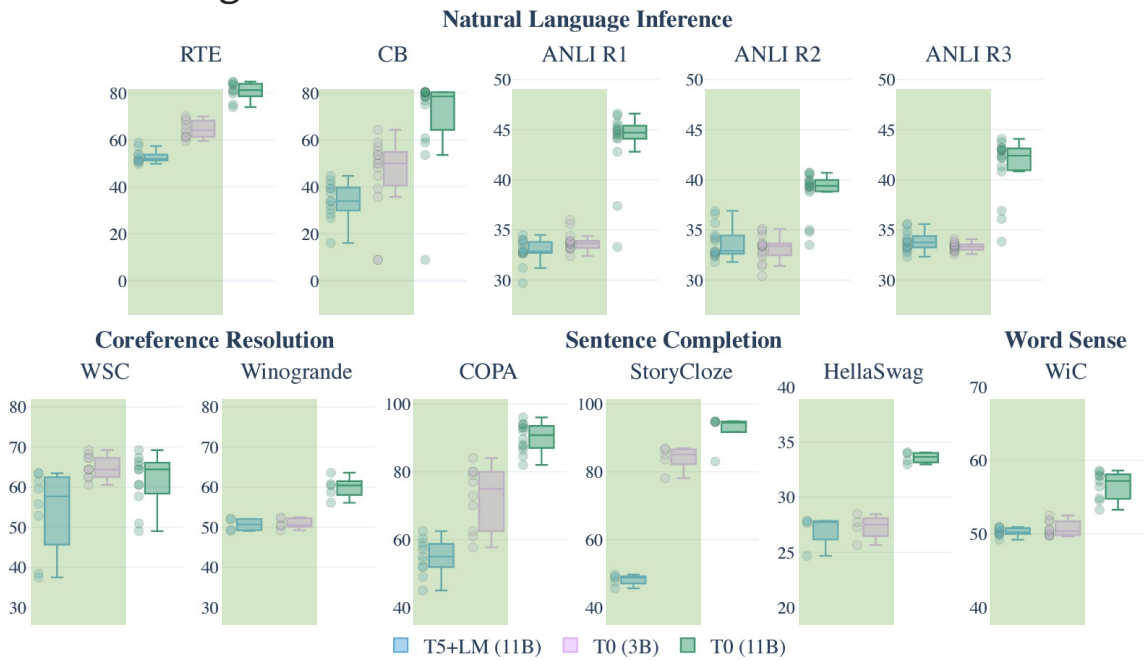
  > FLAN

  - Decoder-only
  - Single held-out task
  - Under 8B model w/ instruction tuning
    < 8B model w/o instruction tuning

  > T0

  - Encoder-decoder
  - Multi held-out task
  - 3B T0 > 11B baseline





FLAN vs. T0

- **FLAN vs. T0**

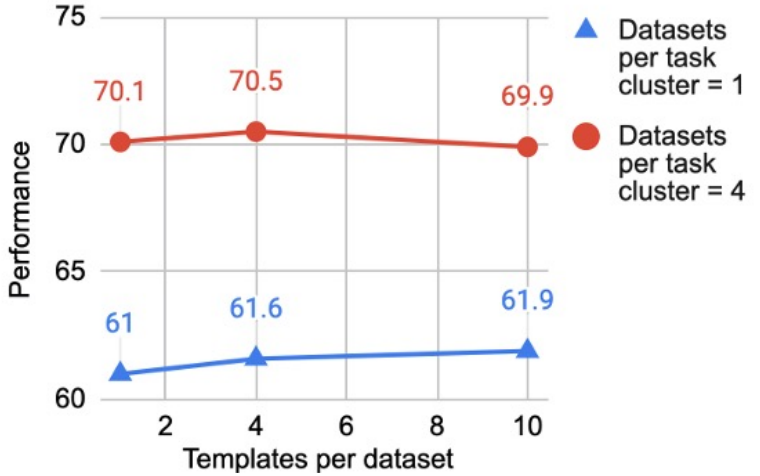
< #prompt >

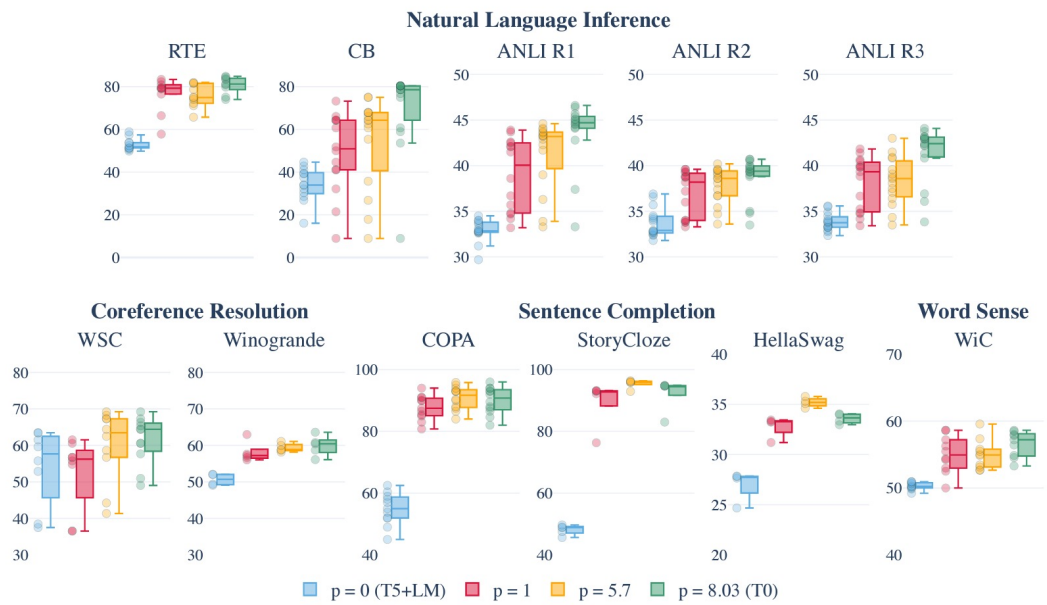
> FLAN

- Decoder-only
- Single held-out task
- Under 8B model w/ instruction tuning
  < 8B model w/o instruction tuning
- 1-line prompt

> T0

- Encoder-decoder
- Multi held-out task
- 3B T0 > 11B baseline
- Various template





Contribution

- Multi-task prompted training can enable strong zero-shot generalization abilities in LM
- Demonstrating ablation study for robustness of prompt wording
- Releasing all prompt template and model

{ End Page }

Thank you :D

Yejin Yoon

HYU NLP Lab.
Dept. of Artificial Intelligence Application,
Hanyang University

stillwithyou@hanyang.ac.kr