Multitask Training and Instruction-Based Fine-Tuning

Flan T5/ T0

Groupe_5

Abdelaziz Guelfane
Imane Meziany
Malek Bouhadida
Mohammed El Barhichi
Yousra Yakhou

Timeline - Paradigm Change

CentraleSupélec

- Pre-2018 : **Task-Specific Training**
  - RNNs, LSTMs, GRUs were trained on specific tasks.
  - Even Seq2Seq models with Attention (Transformer architecture) involved training a separate model for each task.

Timeline - Paradigm Change

CentraleSupélec

- Pre-2018 : **Task-Specific Training**
  - RNNs, LSTMs, GRUs were trained on specific tasks.
  - Even Seq2Seq models with Attention (Transformer architecture) involved training a separate model for each task.

- 2019 : **Text-To-Text framework**
  - T5 reformulates any NLP task to a text generation problem.
  - Example : Instead of predicting a class label, the model generates the label as text (e.g., "positive sentiment").
  - However, fine-tuning on specific tasks was still needed for optimal performance.

Timeline - Paradigm Change

CentraleSupélec

- Pre-2018 : **Task-Specific Training**
  - RNNs, LSTMs, GRUs were trained on specific tasks
  - Even Seq2Seq models with Attention (Transformer architecture) involved training a separate model for each task.

- 2019 : **Text-To-Text framework**
  - T5 reformulates any NLP task to a text generation problem.
  - Example : Instead of predicting a class label, the model generates the label as text (e.g., "positive sentiment").
  - However, fine-tuning on specific tasks was still needed for optimal performance.

- 2020-2022 : **Instruction Tuning on Multi-tasks**
  - **T0** : Sanh, V., A. Webson, C. Raffel, et al. "**Multitask Prompted Training Enables Zero-Shot Task Generalization.**" In ICLR, 2022.
  - **Flan-T5** : Chung, H.W., L. Hou, S. Longpre, et al. "**Scaling Instruction-Finetuned Language Models.**" In CoRR, 2022.

Timeline - Paradigm Change

CentraleSupélec

- Pre-2018 : **Task-Specific Training**
  - RNNs, LSTMs, GRUs were trained on specific tasks
  - Even Seq2Seq models with Attention (Transformer architecture) involved training a separate model for each task.

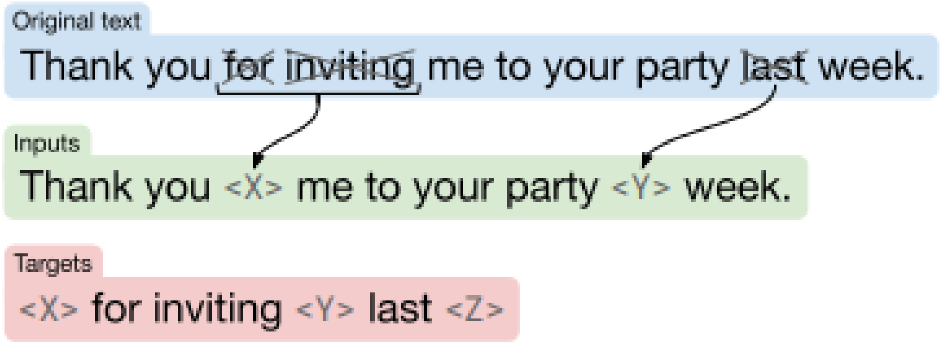
- 2019 : **Text-To-Text framework**
  - T5 reformulates any NLP task to a text generation problem.
  - Example : Instead of predicting a class label, the model generates the label as text (e.g., "positive sentiment").
  - However, fine-tuning on specific tasks was still needed for optimal performance.

Better Generalization ?

- 2020-2022 : **Instruction Tuning on Multi-tasks**
  - **T0** : Sanh, V., A. Webson, C. Raffel, et al. "**Multitask Prompted Training Enables Zero-Shot Task Generalization.**" In ICLR, 2022.
  - **Flan-T5** : Chung, H.W., L. Hou, S. Longpre, et al. "**Scaling Instruction-Finetuned Language Models.**" In CoRR, 2022.

T5: Text-to-Text Transfer Transformer

- **T5** reformulates any NLP task as a sequence-to-sequence generation problem.
- It is based on an **Encoder-Decoder** architecture and trained using **span corruption.**
- It was pretained on **C4 (Colossal Clean Crawled Corpus).**



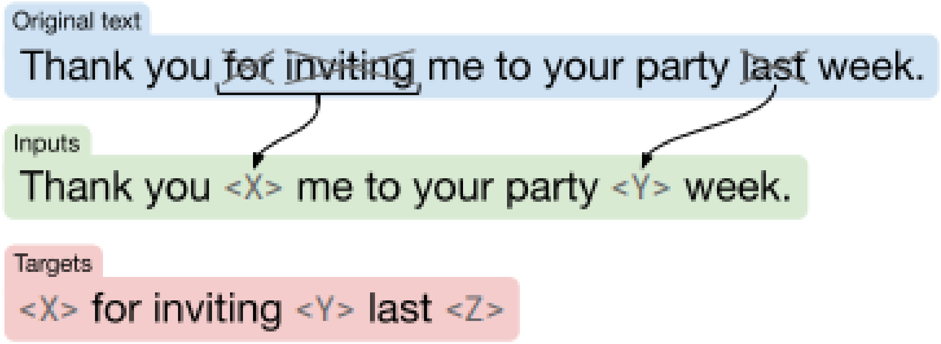
- **T5+LM** is the T5 version designed for language modeling tasks.
- It is T5 trained on 100B additional tokens from C4 on a **causal language modeling** objective.

The model predicts the next token in a sequence, given all prior tokens.

T5: Text-to-Text Transfer Transformer

CentraleSupélec

- **T5** reformulates any NLP task as a sequence-to-sequence generation problem.
- It is based on an **Encoder-Decoder** architecture and trained using **span corruption.**
- It was pretained on **C4 (Colossal Clean Crawled Corpus).**



- **T5+LM** is the T5 version designed for language modeling tasks.
- It is T5 trained on 100B additional tokens from C4 on a **causal language modeling** objective.

The model predicts the next token in a sequence, given all prior tokens.

- T5-LM, and LLMs in general, attain reasonable zero-shot generalization on a diverse set of tasks.

Hypothesis :

Consequence of implicit multitask learning in language models' pretraining.

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

Implicit multitask learning

- Many websites contain lists of trivia Q&As

››››› **Supervised training data** for the task of closedbook question answering.

1. What does "www" stand for in a website browser?

Answer: World Wide Web

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

Answer: 50 meters

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

Answer: Germany, Italy, and Japan

https://www.quizbreaker.com/trivia-questions

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

Implicit multitask learning

Limitations..

- Many websites contain lists of trivia Q&As

››››› **Supervised training data** for the task of closedbook question answering.

1. What does "www" stand for in a website browser?

Answer: World Wide Web

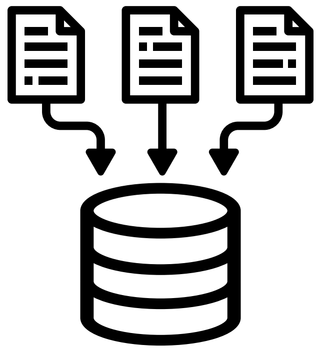
2. How long is an Olympic swimming pool (in meters)?

Answer: 50 meters

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

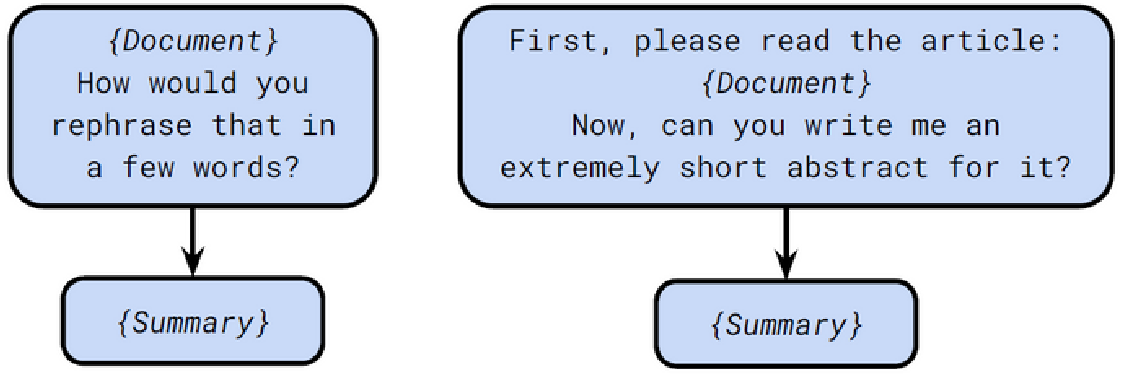
Answer: Germany, Italy, and Japan

1



Requires a sufficiently large model + larger corpus of data

2



Sensitive to the wording of prompts

https://www.quizbreaker.com/trivia-questions

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

Implicit multitask learning

- Many websites contain lists of trivia Q&As

››››› **Supervised training data** for the task of closedbook question answering.

1. What does "www" stand for in a website browser?

Answer: World Wide Web

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

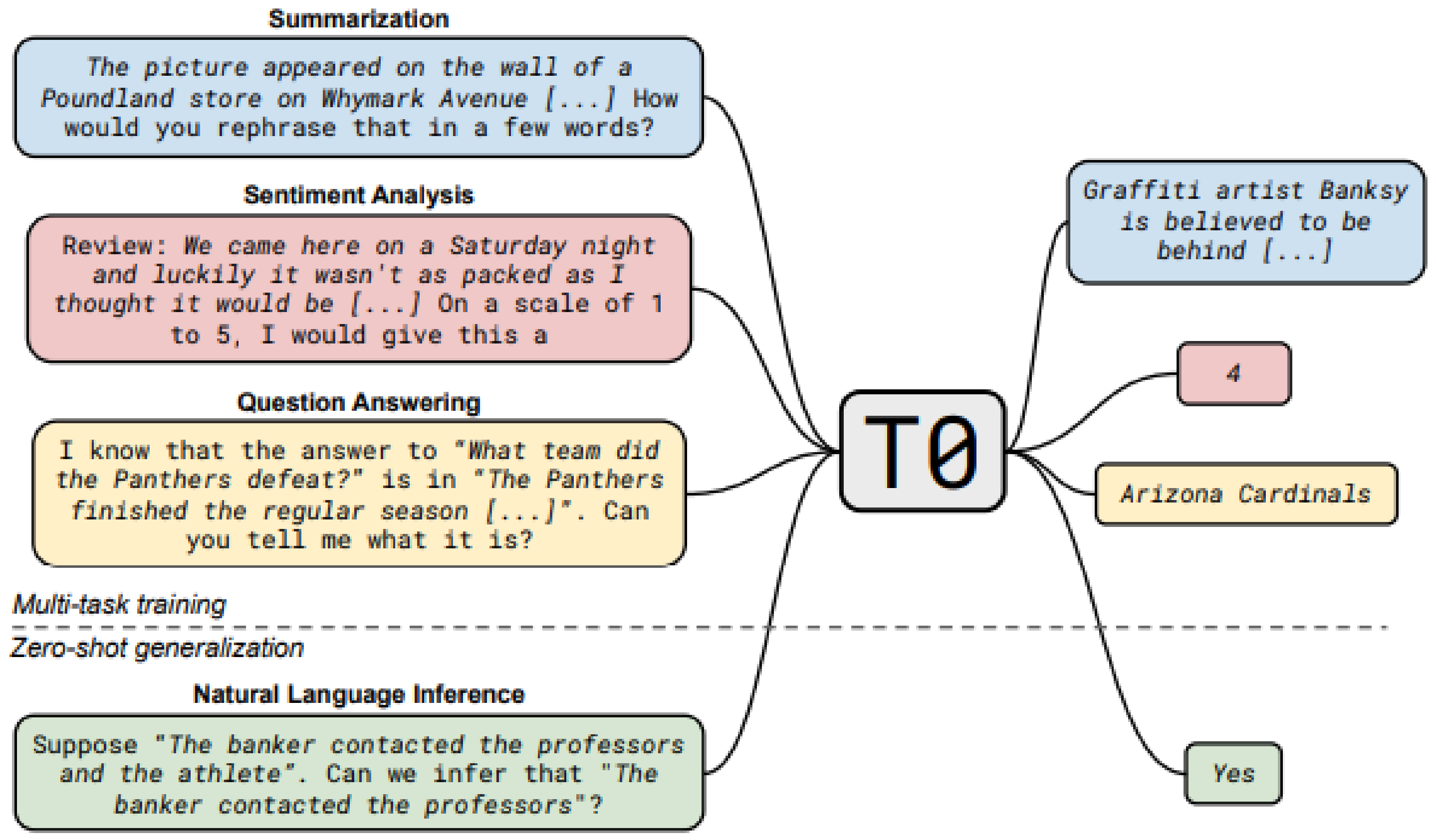
Answer: 50 meters

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

Answer: Germany, Italy, and Japan

VS

Explicit multitask learning

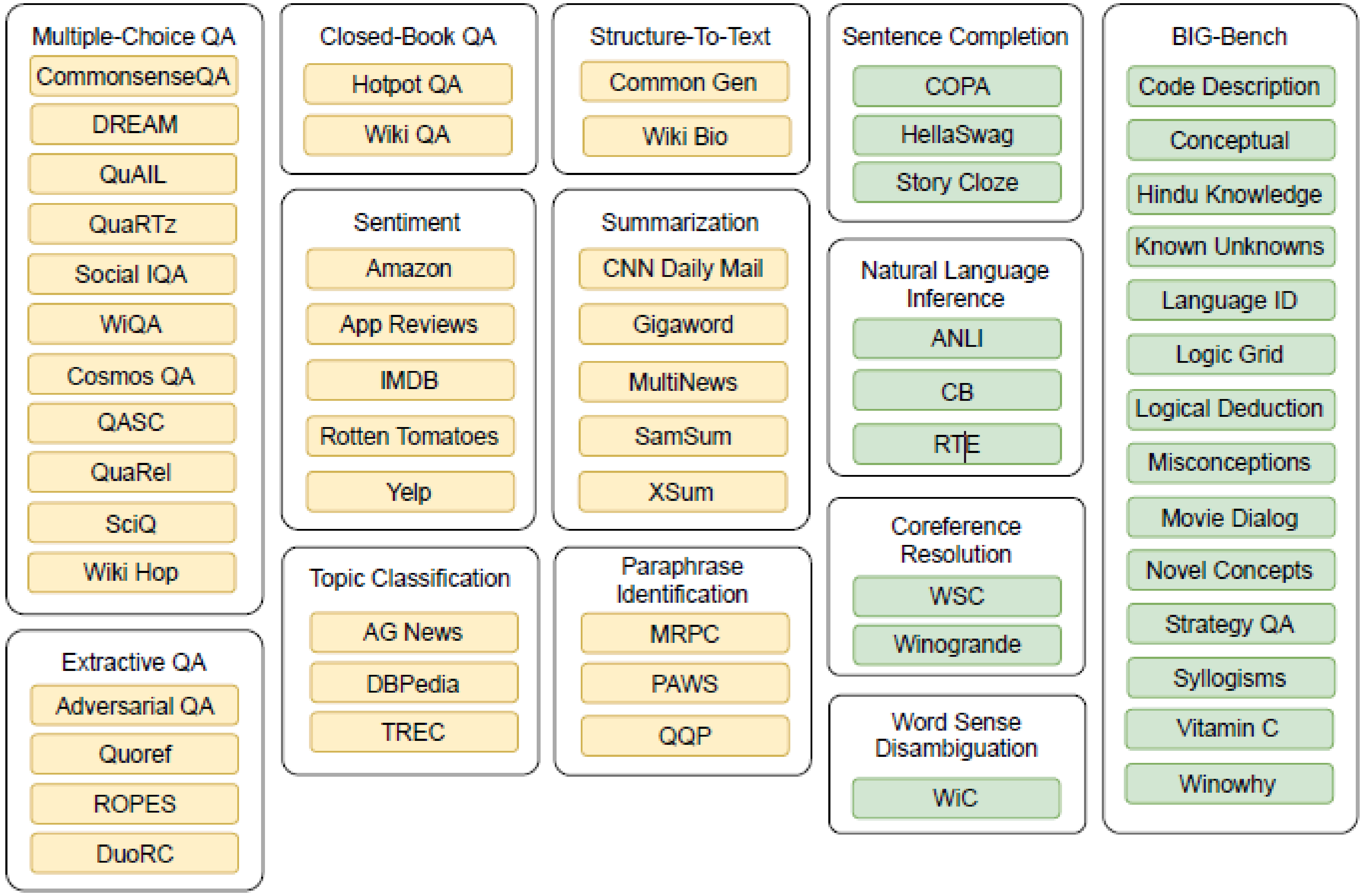


T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

- **T0** datasets and task taxonomy (T0+ and T0++ are trained on additional datasets).

- **12 tasks and 62 datasets** with publicly contributed prompts.
- Yellow datasets are in the training mixture.
- Green datasets are held out and represent tasks that were not seen during training.



T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

- A quick overview of the models :

T5+LM — **T5 trained on 100B additional tokens from C4 on a standard language modeling objective.**

T0 — **Fine-tuning T5+LM on multitask datasets reformulated as prompts.**

T0+ — **T0 Trained on a mixture that adds GPT-3's evaluation datasets.**

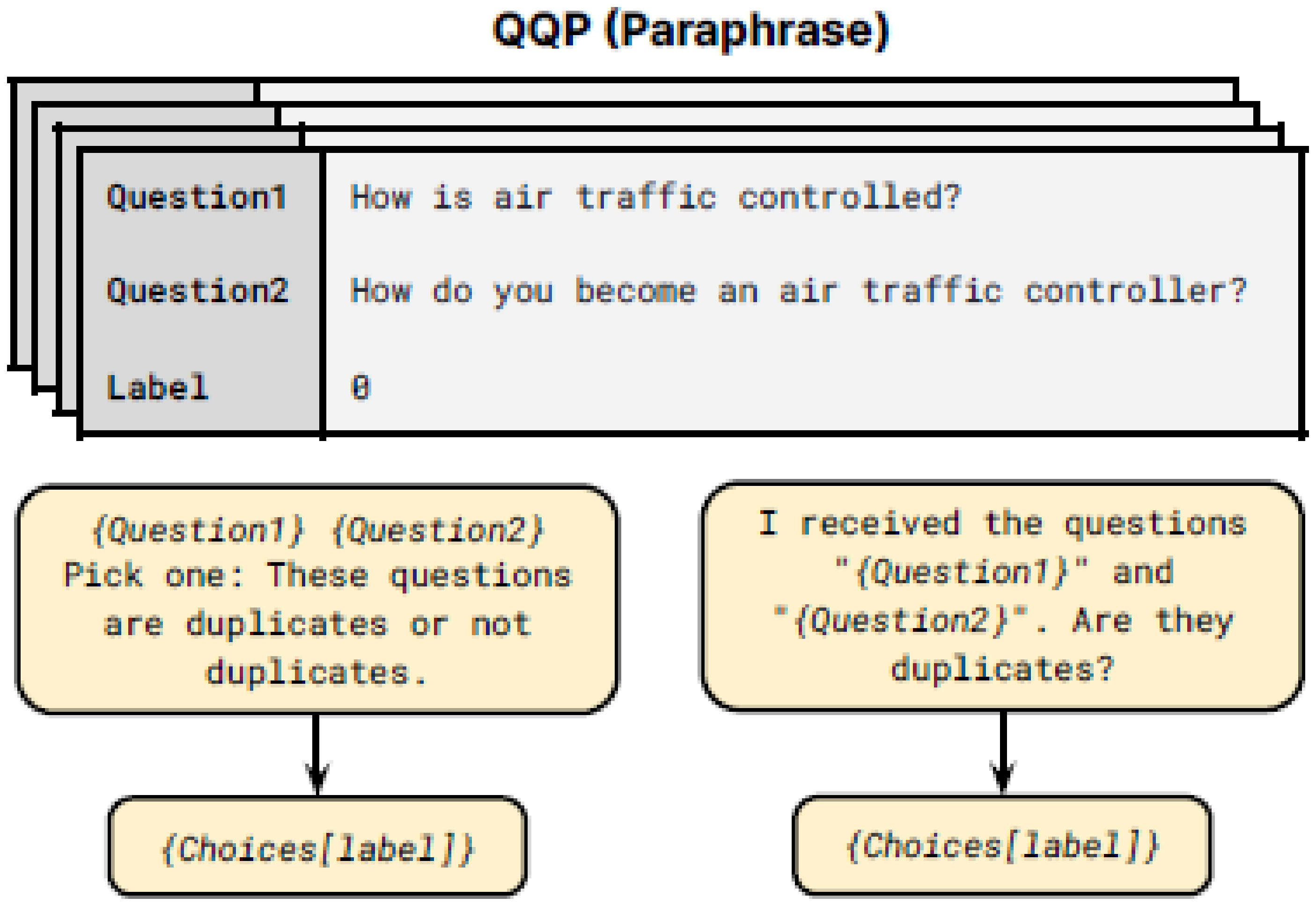
T0++ — **T0+ Trained on a mixture that adds SuperGLUE's evaluation datasets.**

The above **T0 variants** are all initialized from the **11B parameters** version of **T5+LM**.

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

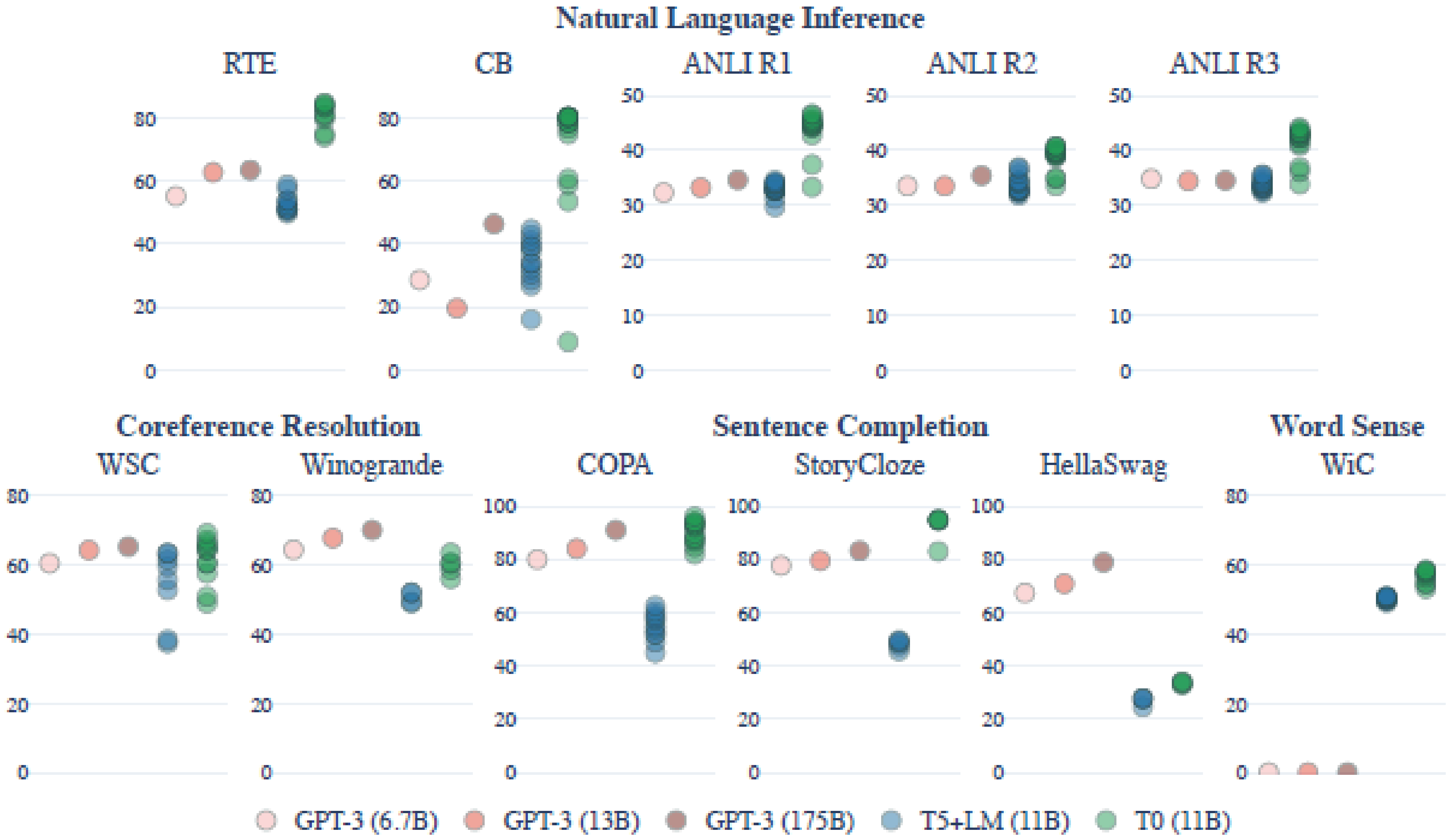
Models learn to understand the prompts as task instructions which help them generalize to held-out tasks



How did generalization to held-out tasks improve?

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

- Results for **T0** task generalization experiments compared to **GPT-3**.
- T5+LM (baseline model) is the same as T0 except without multitask prompted training.



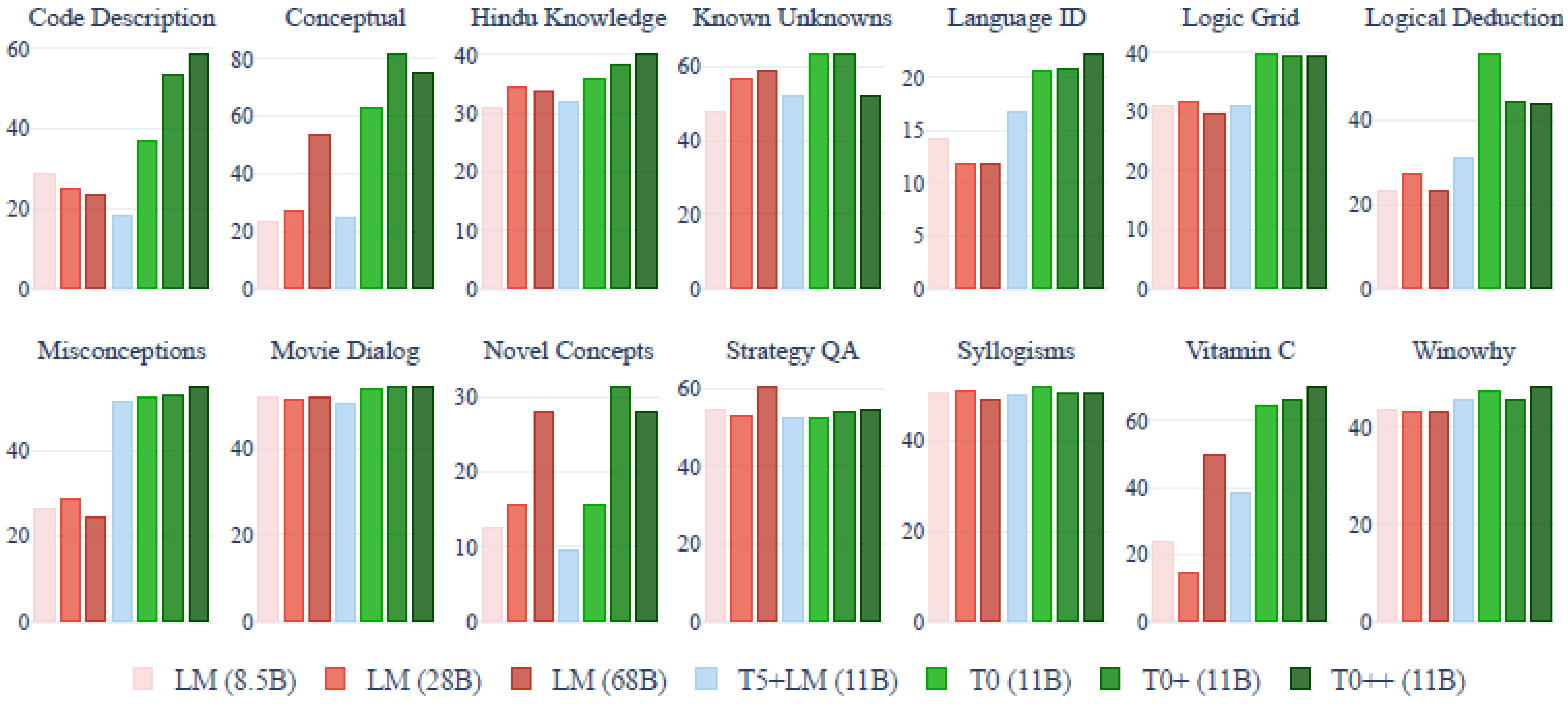
Better Generalization !

Each dot is the performance of one evaluation prompt.

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

- Results for a subset of BIG-bench models compared to T0, T0+ and T0++.

Better Generalization !



T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

- So far, we have seen that..

››››› **Models learn to understand the prompts as task instructions which help them generalize to held-out tasks.**

››››› **Multitask prompted training improve generalization to held-out tasks**

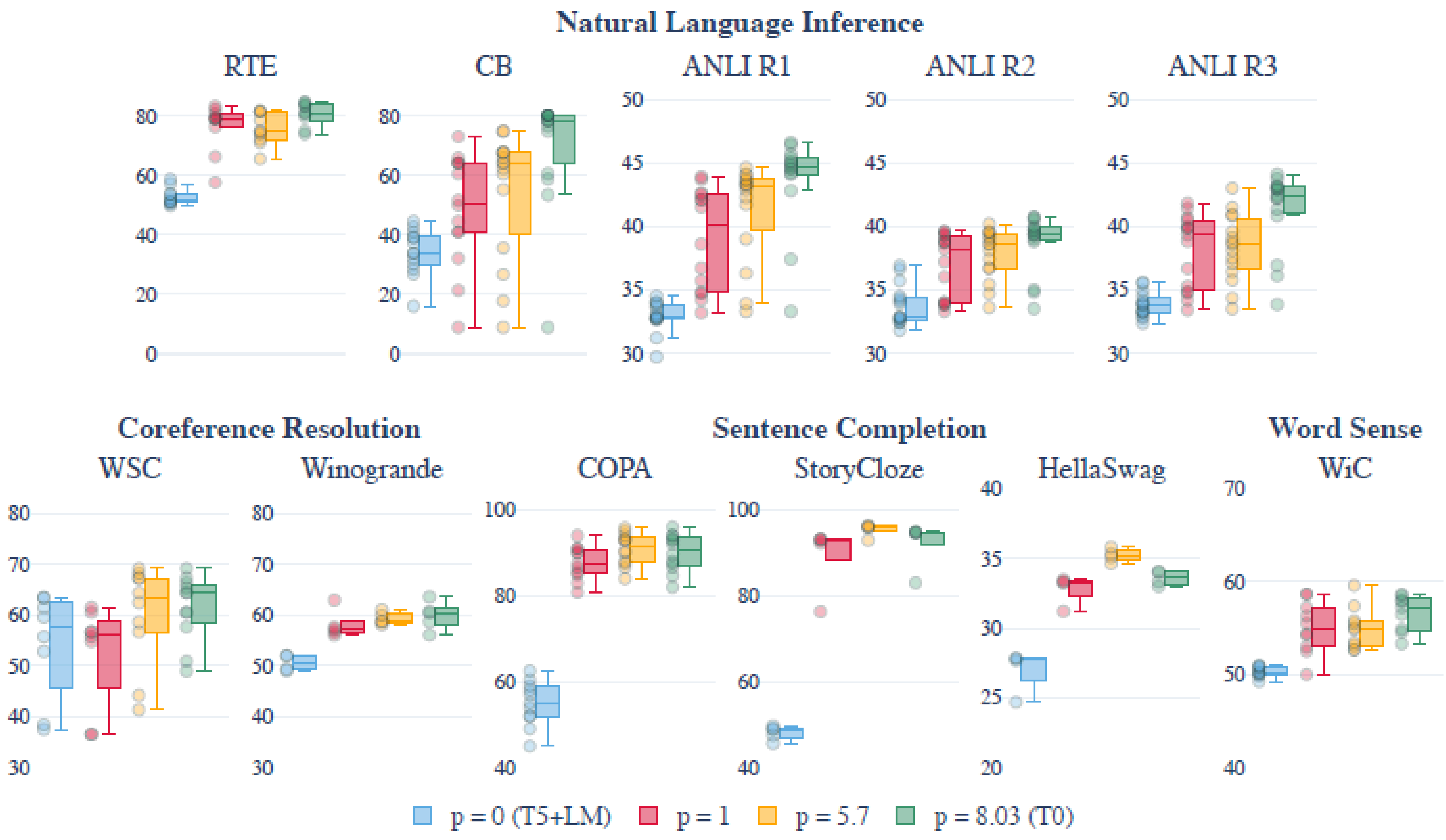
- Now,

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

T0: MultiTask Prompted Training Enables Zero-Shot Task Generalization

CentraleSupélec

- Zero-shot performance of T0 and T5+LM when increasing number of training prompts per dataset.



Each dot is the performance of one evaluation prompt.
Adding more training prompts consistently leads to higher median performance and generally lower interquartile range for held-out tasks.

Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

What happens if we scale even further?

- Increasing the **number of tasks** and **data diversity**?
- Leveraging **larger models**?
- Incorporating **reasoning capabilities** like Chain-of-Thought (CoT)?

Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

What happens if we scale even further?

- Increasing the **number of tasks** and **data diversity**?
- Leveraging **larger models**?
- Incorporating **reasoning capabilities** like Chain-of-Thought (CoT)?

arXiv:2210.11416v5 [cs.LG] 6 Dec 2022

Scaling Instruction-Finetuned Language Models

Hyung Won Chung* Le Hou* Shayne Longpre* Barret Zoph† Yi Tay†
William Fedus† Yunxuan Li Xuezhi Wang Mostafa Dehghani Siddhartha Brahma
Albert Webson Shixiang Shane Gu Zhuyun Dai Mirac Suzgun Xinyun Chen
Aakanksha Chowdhery Alex Castro-Ros Marie Pellat Kevin Robinson
Dasha Valter Sharan Narang Gaurav Mishra Adams Yu Vincent Zhao
Yanping Huang Andrew Dai Hongkun Yu Slav Petrov Ed H. Chi
Jeff Dean Jacob Devlin Adam Roberts Denny Zhou Quoc V. Le
Jason Wei*

Google

Abstract

Finetuning language models on a collection of datasets phrased as instructions has been shown to improve model performance and generalization to unseen tasks. In this paper we explore instruction finetuning with a particular focus on (1) scaling the number of tasks, (2) scaling the model size, and (3) finetuning on chain-of-thought data. We find that instruction finetuning with the above aspects dramatically improves performance on a variety of model classes (PaLM, T5, U-PaLM), prompting setups (zero-shot, few-shot, CoT), and evaluation benchmarks (MMLU, BBH, TyDiQA, MGSM, open-ended generation, RealToxicityPrompts). For instance, Flan-PaLM 540B instruction-finetuned on 1.8K tasks outperforms PaLM 540B by a large margin (+9.4% on average). Flan-PaLM 540B achieves state-of-the-art performance on several benchmarks, such as 75.2% on five-shot MMLU. We also publicly release Flan-T5 checkpoints,[1] which achieve strong few-shot performance even compared to much larger models, such as PaLM 62B. Overall, instruction finetuning is a general method for improving the performance and usability of pretrained language models.

Figure 1: We finetune various language models on 1.8K tasks phrased as instructions, and evaluate them on unseen tasks. We finetune both with and without exemplars (i.e., zero-shot and few-shot) and with and without chain-of-thought, enabling generalization across a range of evaluation scenarios.

*Equal contribution. Correspondence: lehou@google.com.
†Core contributor.
[1]Public checkpoints: https://github.com/google-research/t5x/blob/main/docs/models.md#flan-t5-checkpoints.

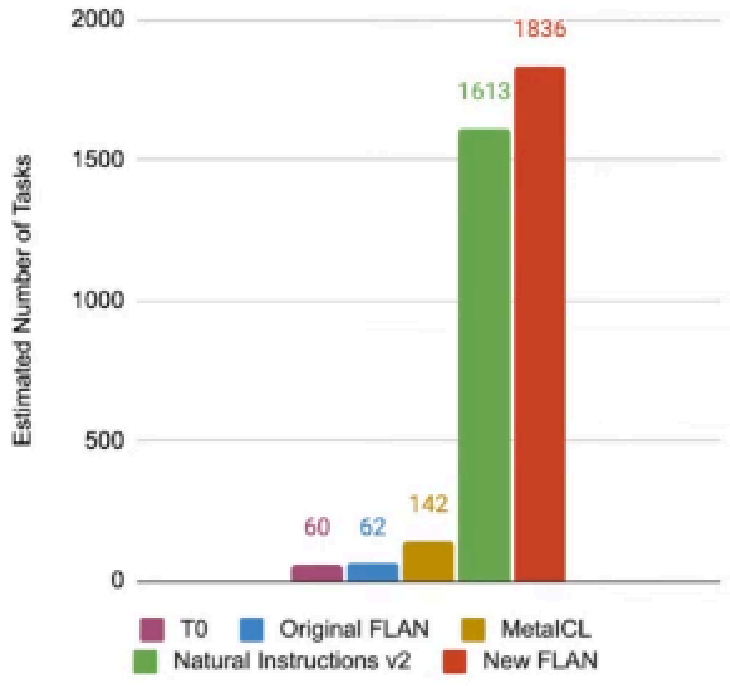
1

Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

1

Scaling Tasks and Datasets



Flan-T5 : Scaling Instruction-Finetuned Language Models

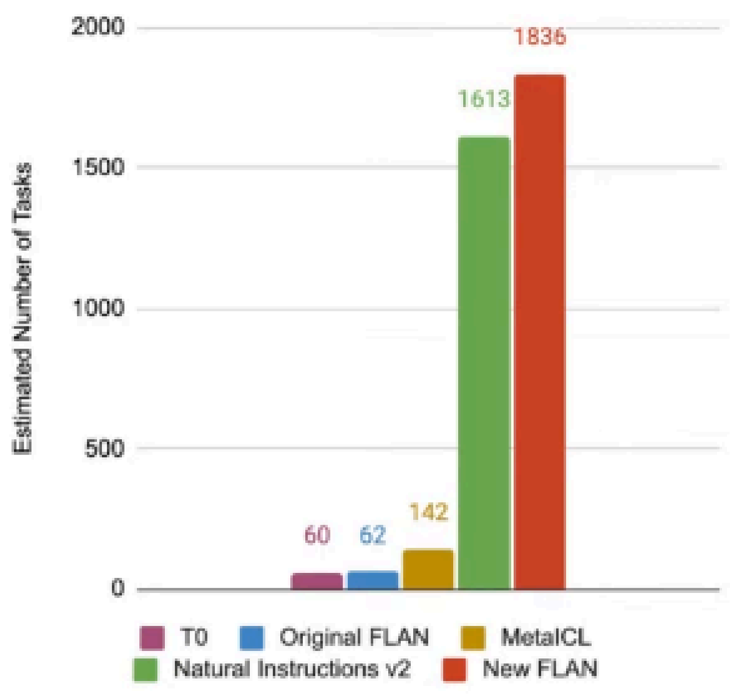
1

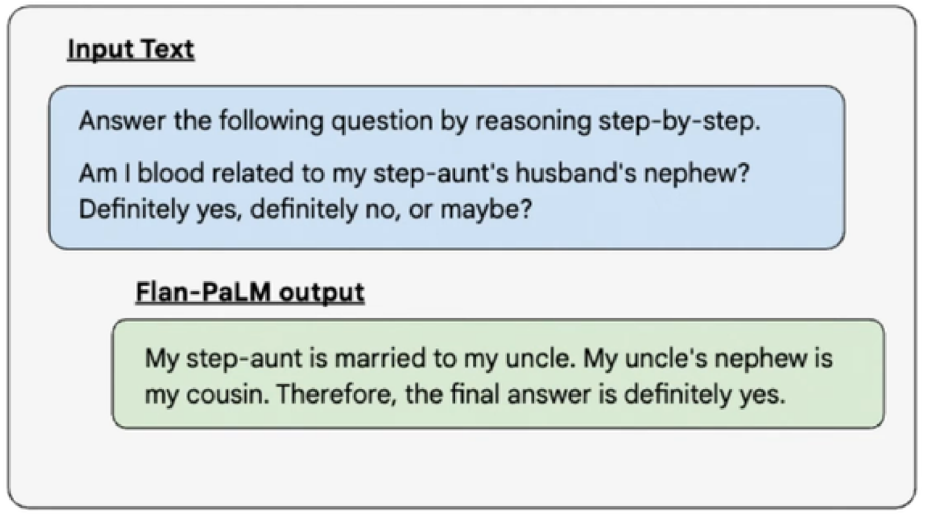
Scaling Tasks and Datasets

+

2

Adding Chain-of-Thought Reasoning





Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

1

Scaling Tasks and Datasets

+

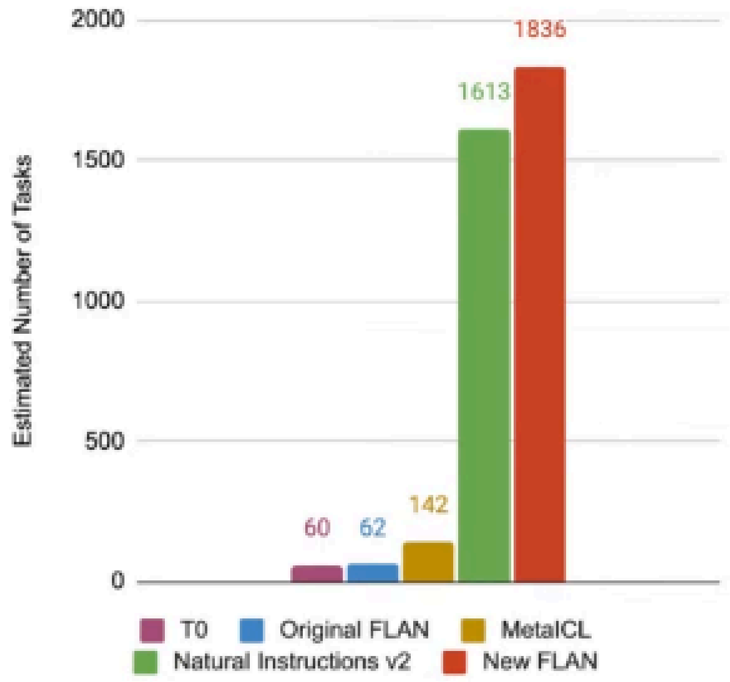
2

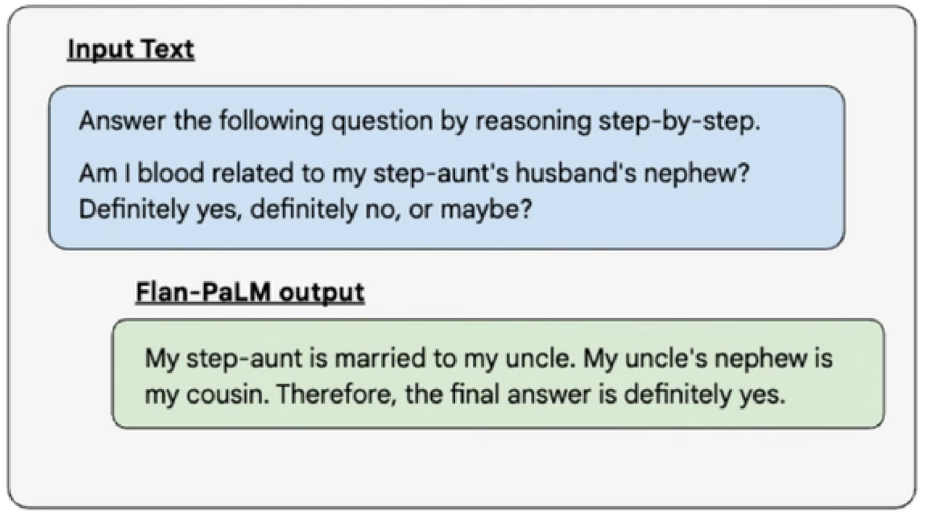
Adding Chain-of-Thought Reasoning

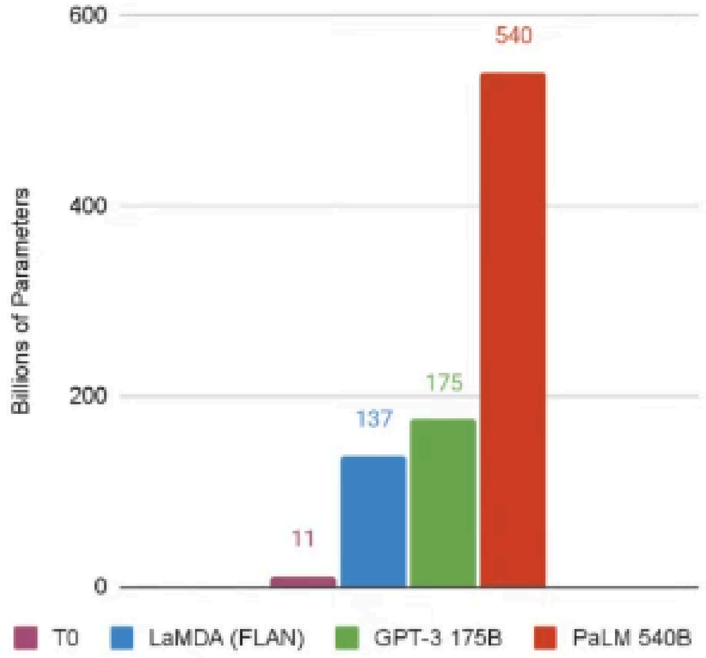
+

3

Using Larger Language Models



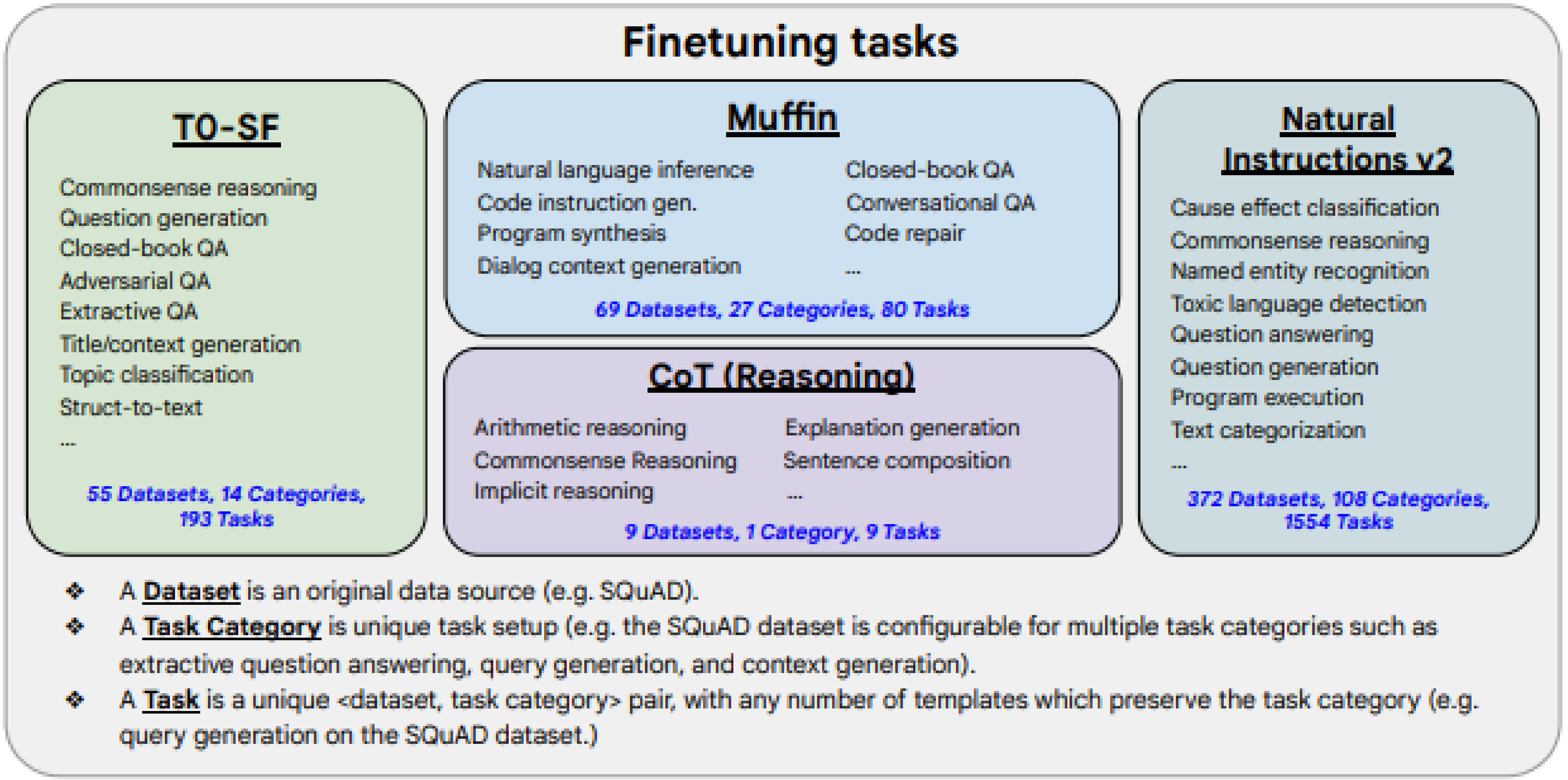




Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

- Dataset used :



- Scaling Finetuning Tasks :

473 datasets

146 task categories

1 836 Tasks !

Flan-T5 : Scaling Instruction-Finetuned Language Models

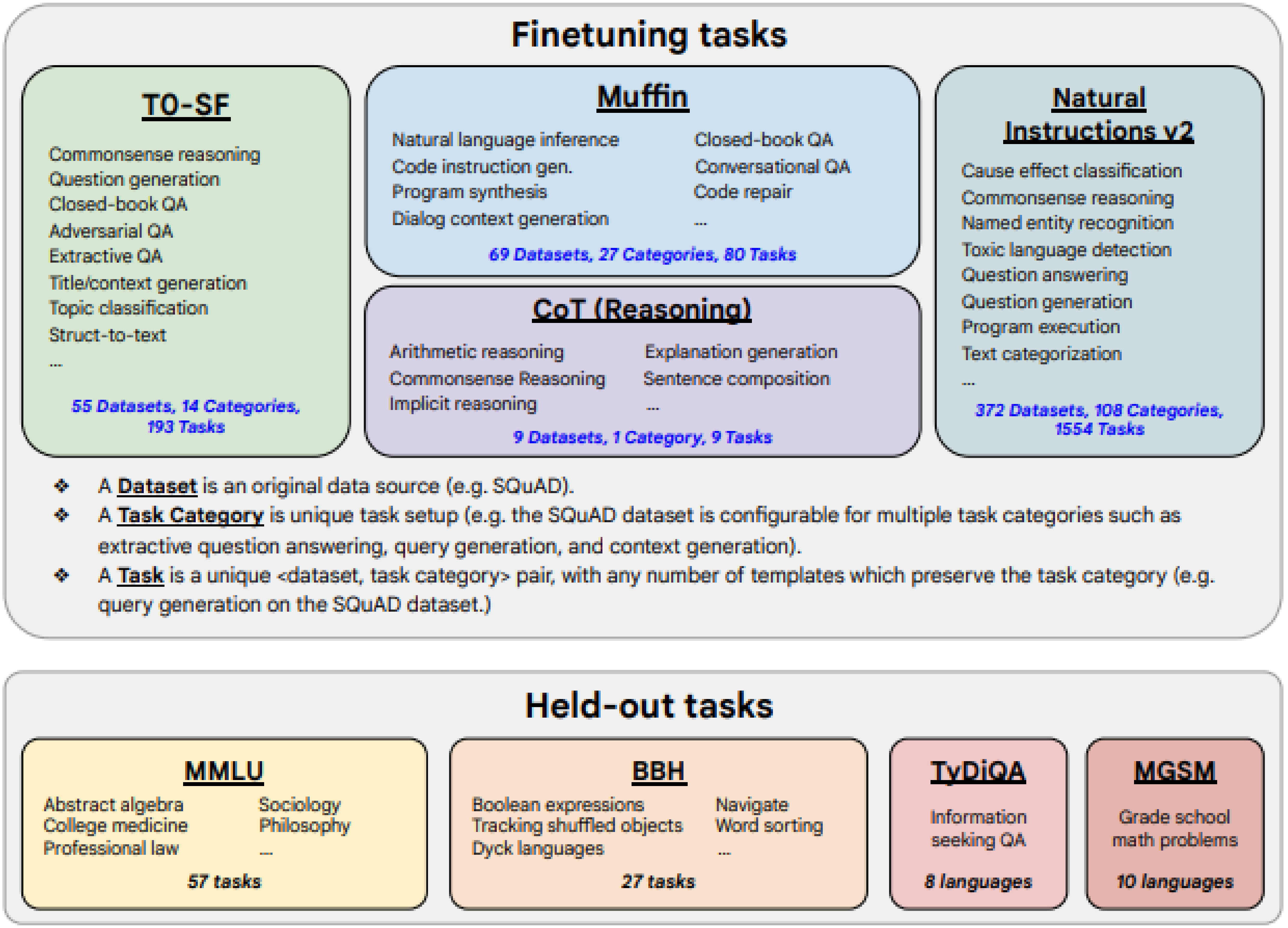
CentraleSupélec

- Dataset used :

  - Scaling Finetuning Tasks :

    473 datasets
    146 task categories
    1 836 Tasks !

  - Evaluation :



Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

- Tasks & Instructions variety :

Finetuning includes more diverse tasks :

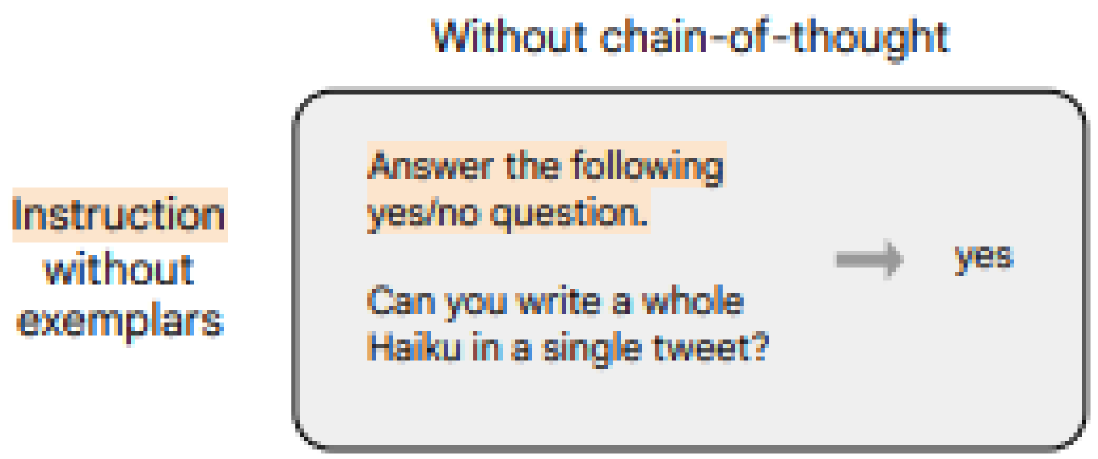
- zero-shot
- few-shot
- chain-of-thought

Flan-T5 : Scaling Instruction-Finetuned Language Models

- Tasks & Instructions variety :

Finetuning includes more diverse tasks :

- zero-shot
- few-shot
- chain-of-thought

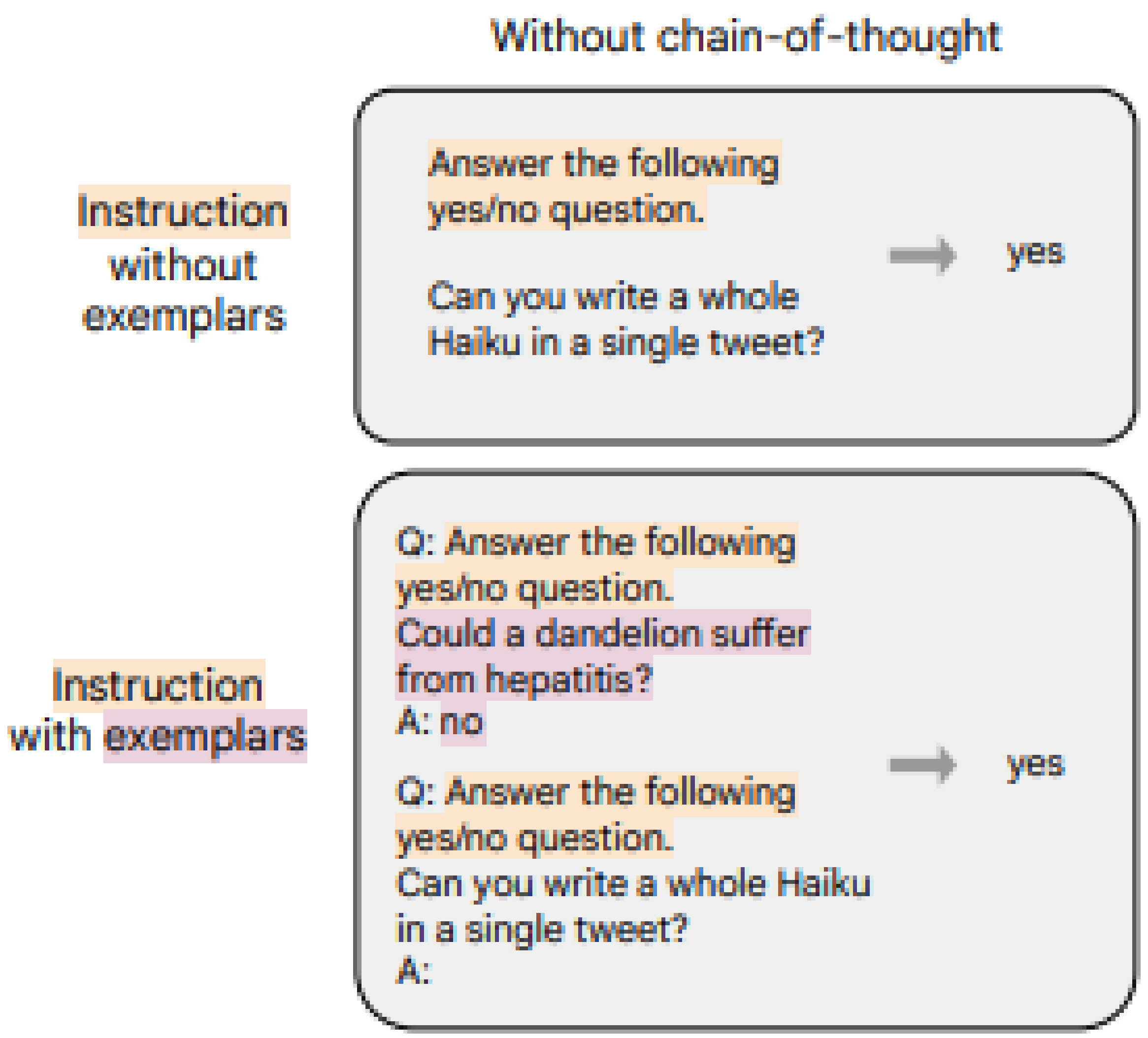


Flan-T5 : Scaling Instruction-Finetuned Language Models

- Tasks & Instructions variety :

Finetuning includes more diverse tasks :

- zero-shot
- few-shot
- chain-of-thought

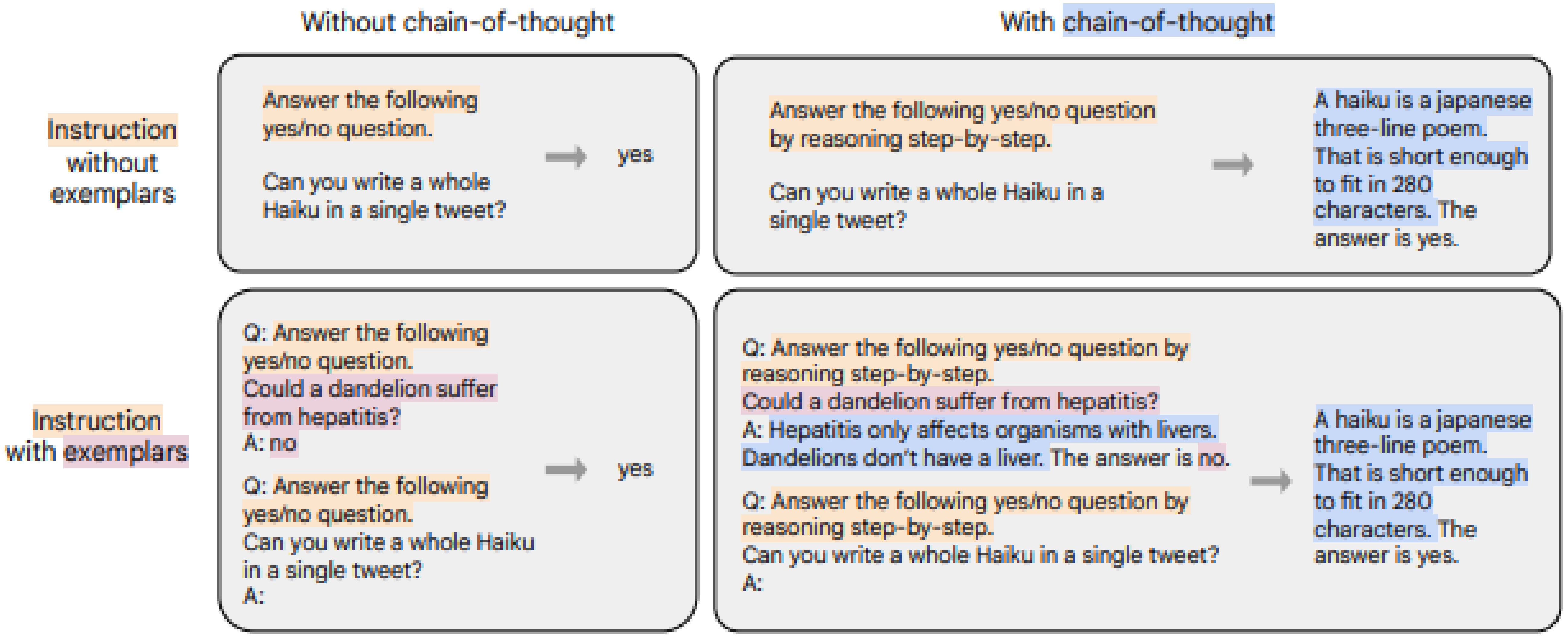


Flan-T5 : Scaling Instruction-Finetuned Language Models

- Tasks & Instructions variety :

Finetuning includes more diverse tasks :

- zero-shot
- few-shot
- chain-of-thought



Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

- Models' Sizes :

Params	Model	Architecture	Pre-training Objective
80M	Flan-T5-Small	encoder-decoder	span corruption
250M	Flan-T5-Base	encoder-decoder	span corruption
780M	Flan-T5-Large	encoder-decoder	span corruption
3B	Flan-T5-XL	encoder-decoder	span corruption
11B	Flan-T5-XXL	encoder-decoder	span corruption
8B	Flan-PaLM	decoder-only	causal LM
62B	Flan-PaLM	decoder-only	causal LM
540B	Flan-PaLM	decoder-only	causal LM
62B	Flan-cont-PaLM	decoder-only	causal LM
540B	Flan-U-PaLM	decoder-only	prefix LM + span corruption

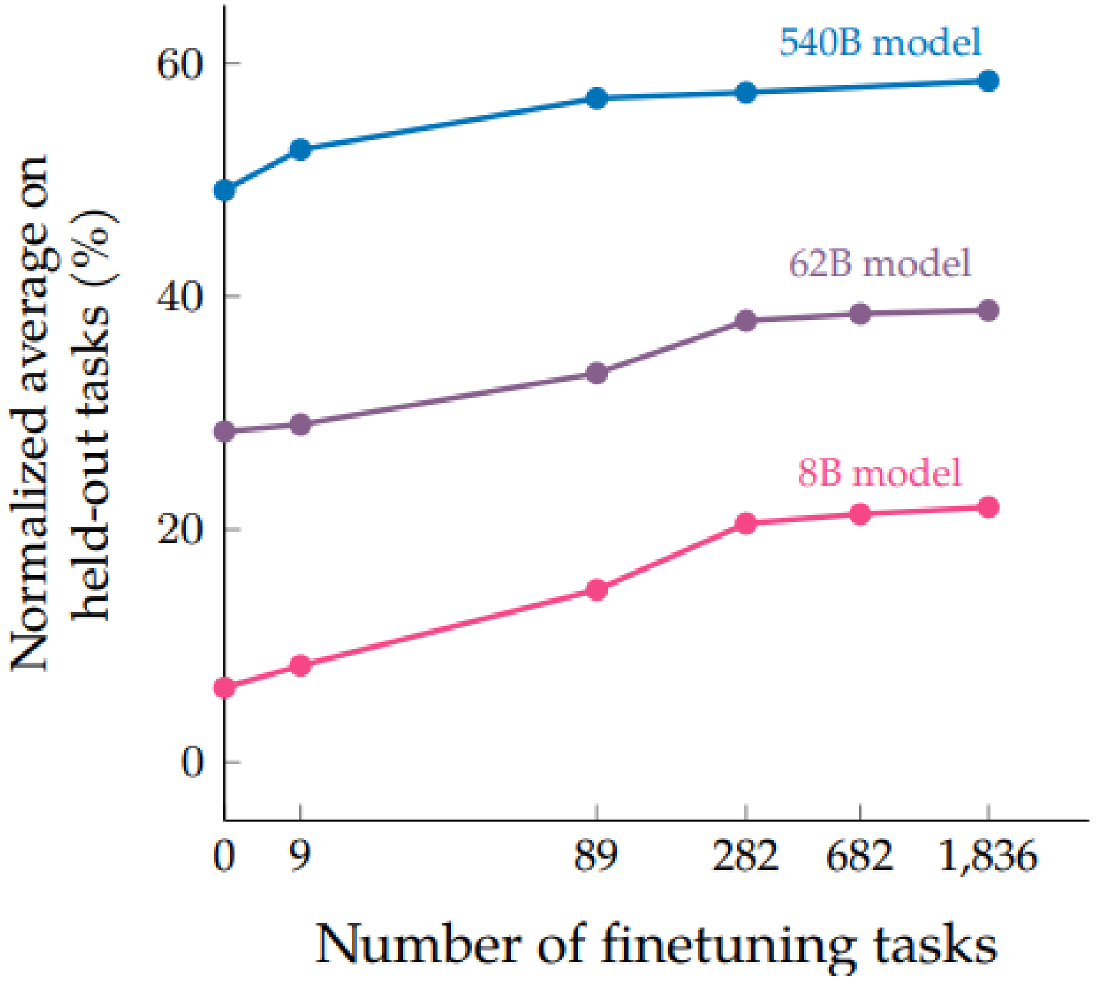
Flan-T5 : Scaling Instruction-Finetuned Language Models

CentraleSupélec

- Results:

Performance improves on :

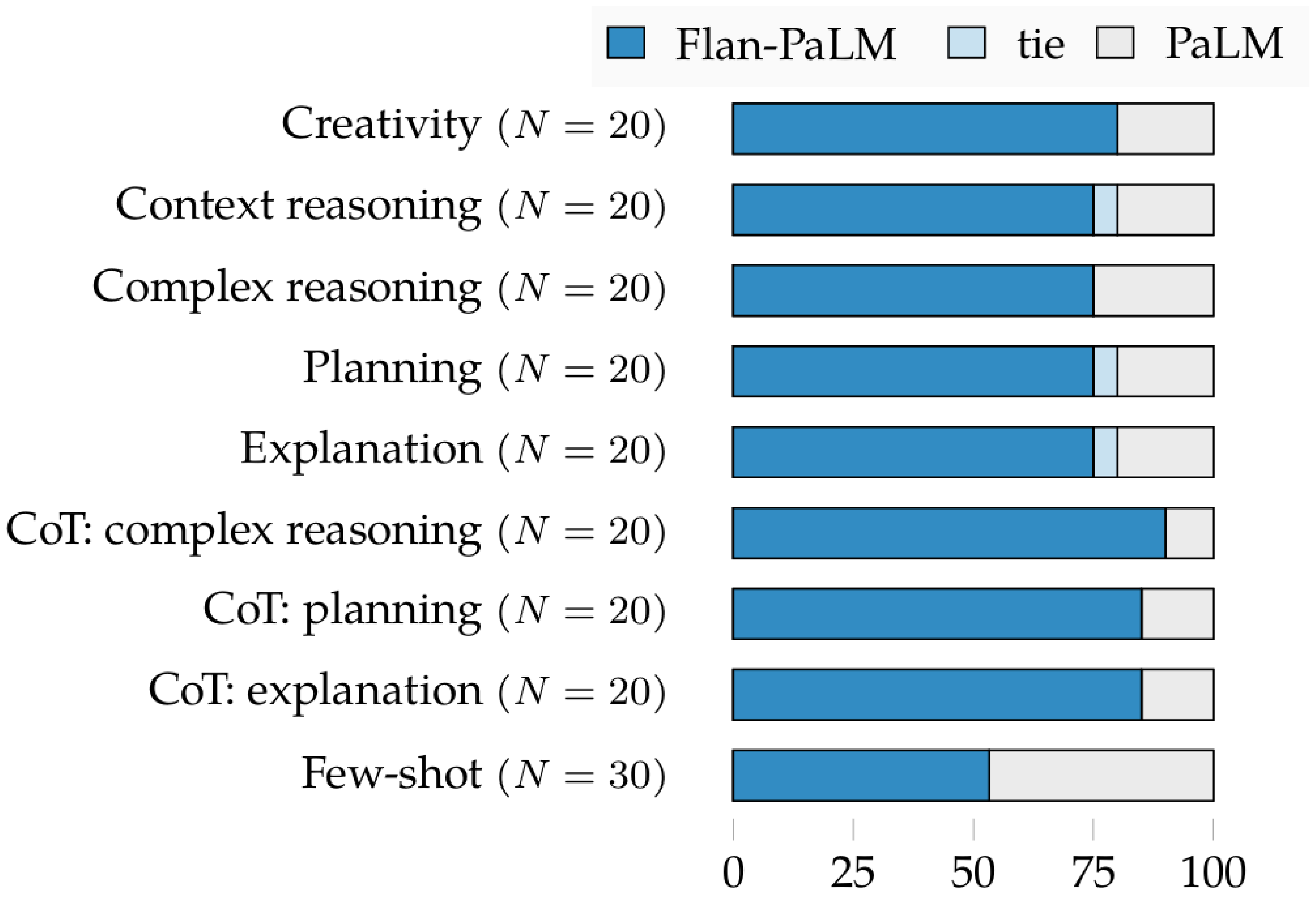
- **different model classes**
- **prompting setups**
- **evaluation benchmarks**



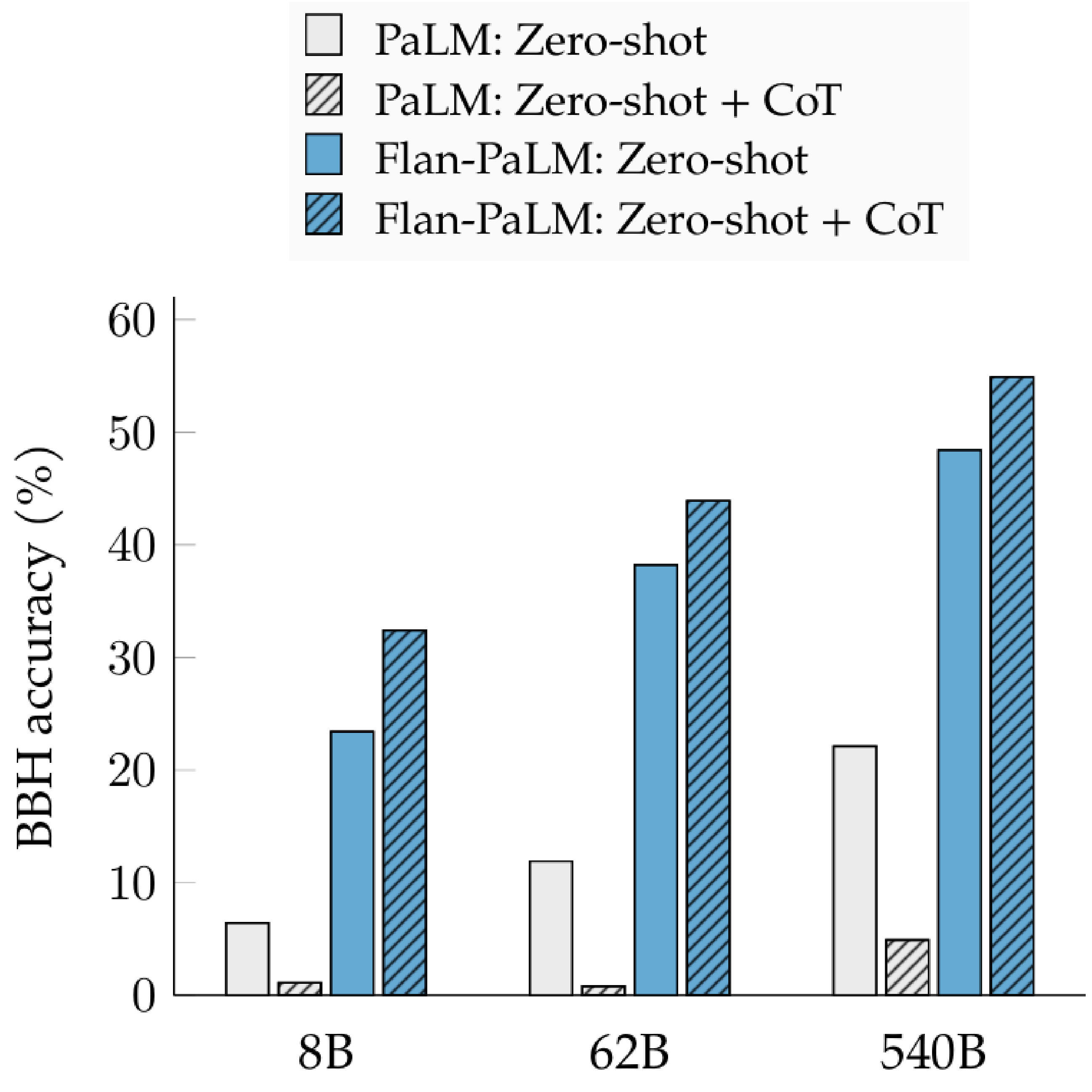
			MMLU		BBH		TyDiQA	MGSM
Params	Model	Norm. avg.	Direct	CoT	Direct	CoT	Direct	CoT
80M	T5-Small	-9.2	26.7	5.6	27.0	7.2	0.0	0.4
	Flan-T5-Small	-3.1 (+6.1)	28.7	12.1	29.1	19.2	1.1	0.2
250M	T5-Base	-5.1	25.7	14.5	27.8	14.6	0.0	0.5
	Flan-T5-Base	6.5 (+11.6)	35.9	33.7	31.3	27.9	4.1	0.4
780M	T5-Large	-5.0	25.1	15.0	27.7	16.1	0.0	0.3
	Flan-T5-Large	13.8 (+18.8)	45.1	40.5	37.5	31.5	12.3	0.7
3B	T5-XL	-4.1	25.7	14.5	27.4	19.2	0.0	0.8
	Flan-T5-XL	19.1 (+23.2)	52.4	45.5	41.0	35.2	16.6	1.9
11B	T5-XXL	-2.9	25.9	18.7	29.5	19.3	0.0	1.0
	Flan-T5-XXL	23.7 (+26.6)	55.1	48.6	45.3	41.4	19.0	4.9
8B	PaLM	6.4	24.3	24.1	30.8	30.1	25.0	3.4
	Flan-PaLM	21.9 (+15.5)	49.3	41.3	36.4	31.1	47.5	8.2
62B	PaLM	28.4	55.1	49.0	37.4	43.0	40.5	18.2
	Flan-PaLM	38.8 (+10.4)	59.6	56.9	47.5	44.9	58.7	28.5
540B	PaLM	49.1	71.3	62.9	49.1	63.7	52.9	45.9
	Flan-PaLM	58.4 (+9.3)	73.5	70.9	57.9	66.3	67.8	57.0
62B	cont-PaLM	38.1	61.2	57.6	41.7	53.1	45.7	32.0
	Flan-cont-PaLM	46.7 (+8.6)	66.1	62.0	51.0	53.3	62.7	40.3
540B	U-PaLM	50.2	71.5	64.0	49.2	62.4	54.6	49.9
	Flan-U-PaLM	59.1 (+8.9)	74.1	69.8	59.3	64.9	68.3	60.4

Flan-T5 : Scaling Instruction-Finetuned Language Models

- Results:



Instruction Tuning Improves Human Usability



Finetuning with Chain-of-Thought Data Unlocks Zero-Shot Reasoning

References

- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K.,. . . Wei, J. (2022, 20 octobre). Scaling Instruction-Finetuned Language Models. arXiv.org. https://arxiv.org/abs/2210.11416

- Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N.,. . . Rush, A. M. (2021, 15 octobre). Multitask prompted training enables Zero-Shot task generalization. arXiv.org. https://arxiv.org/abs/2110.08207

CentraleSupélec

THANK YOU

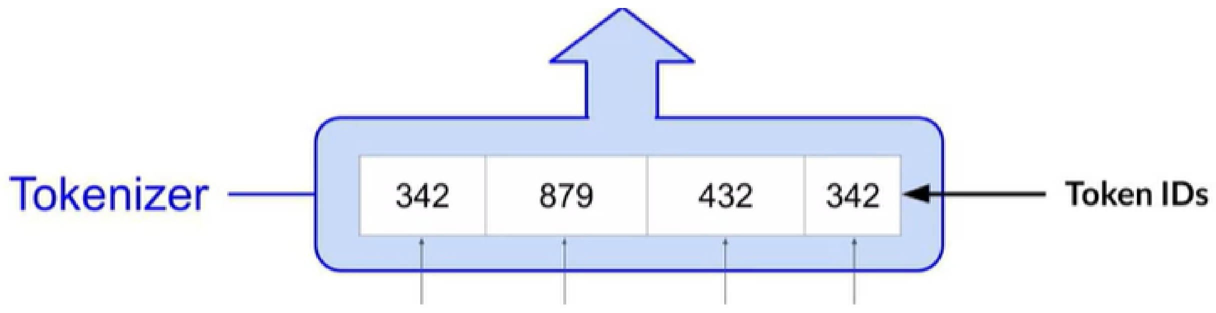
Appendix

CentraleSupélec

sequence vectors:

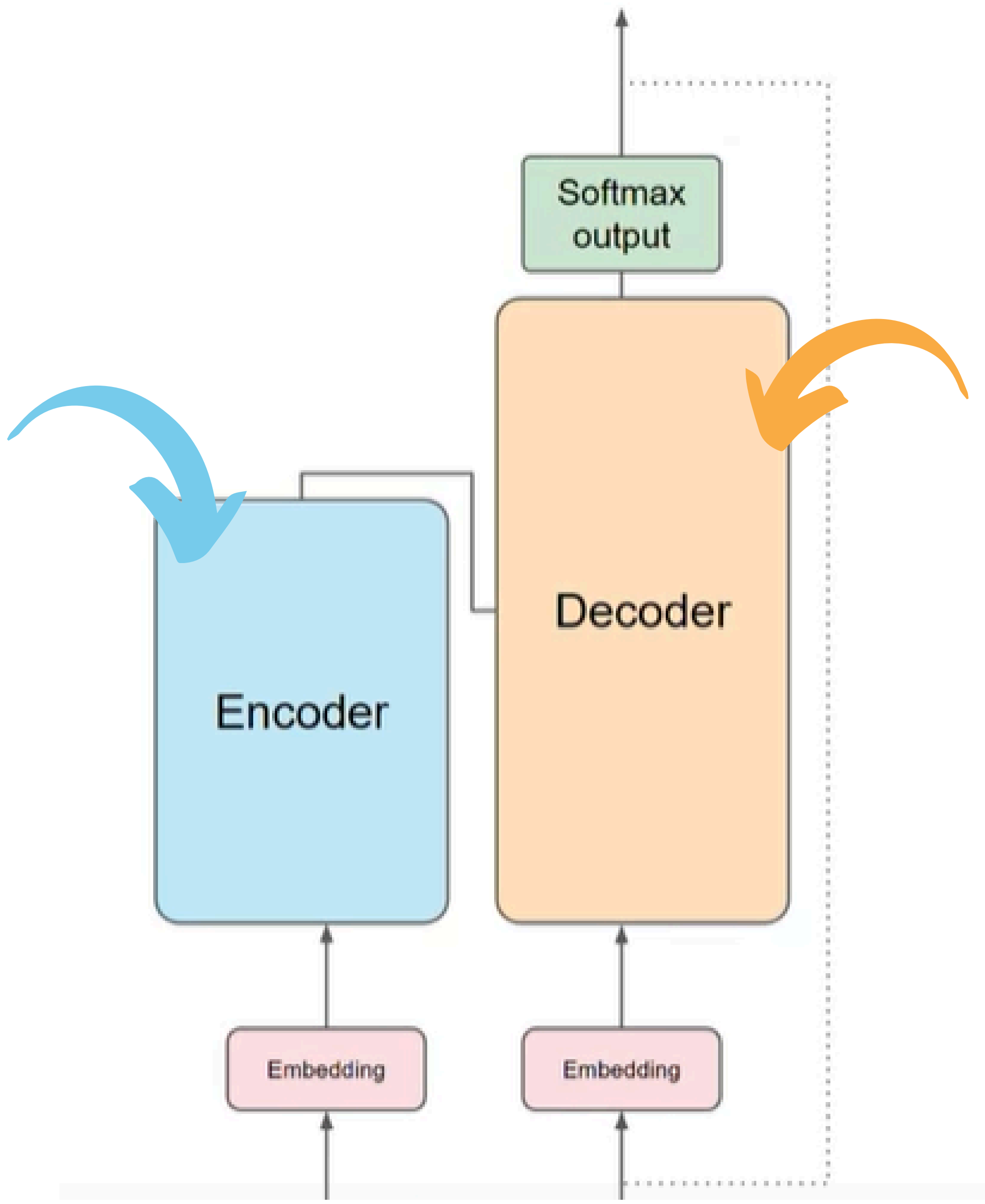
encodes information about both the element itself and its context within the sequence.

process the input sequence and create contextualized representations for each element (word or token) in the sequence



Input

Output



Inputs

the final representation vectors

generate an output sequence based on the contextualized representations provided by the encoder