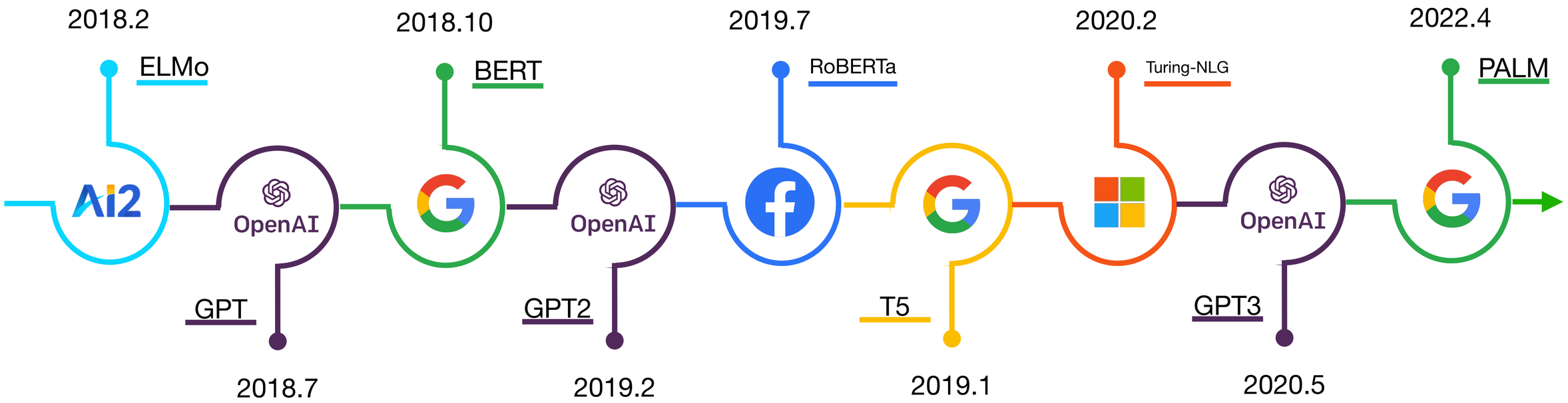
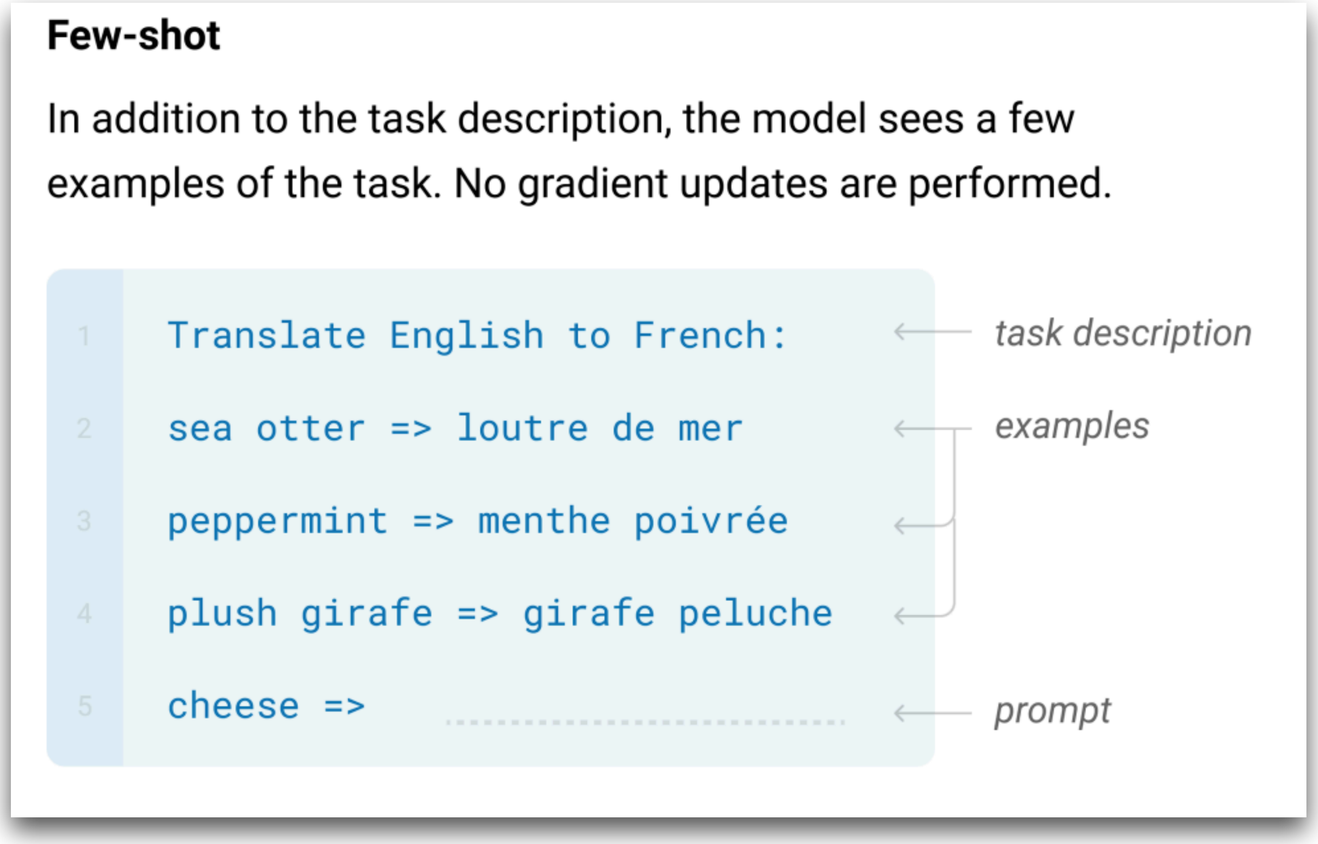
Super-NaturalInstructions: Generalization via Declarative Instructions on 1,600+ NLP Tasks

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi,
Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, Daniel Khashabi, and 31 others

Rapid progress in pre-trained language models

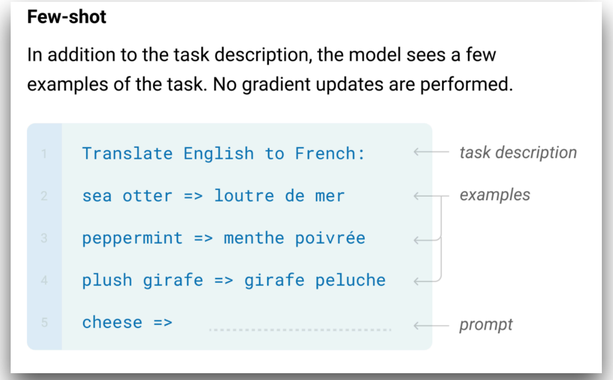


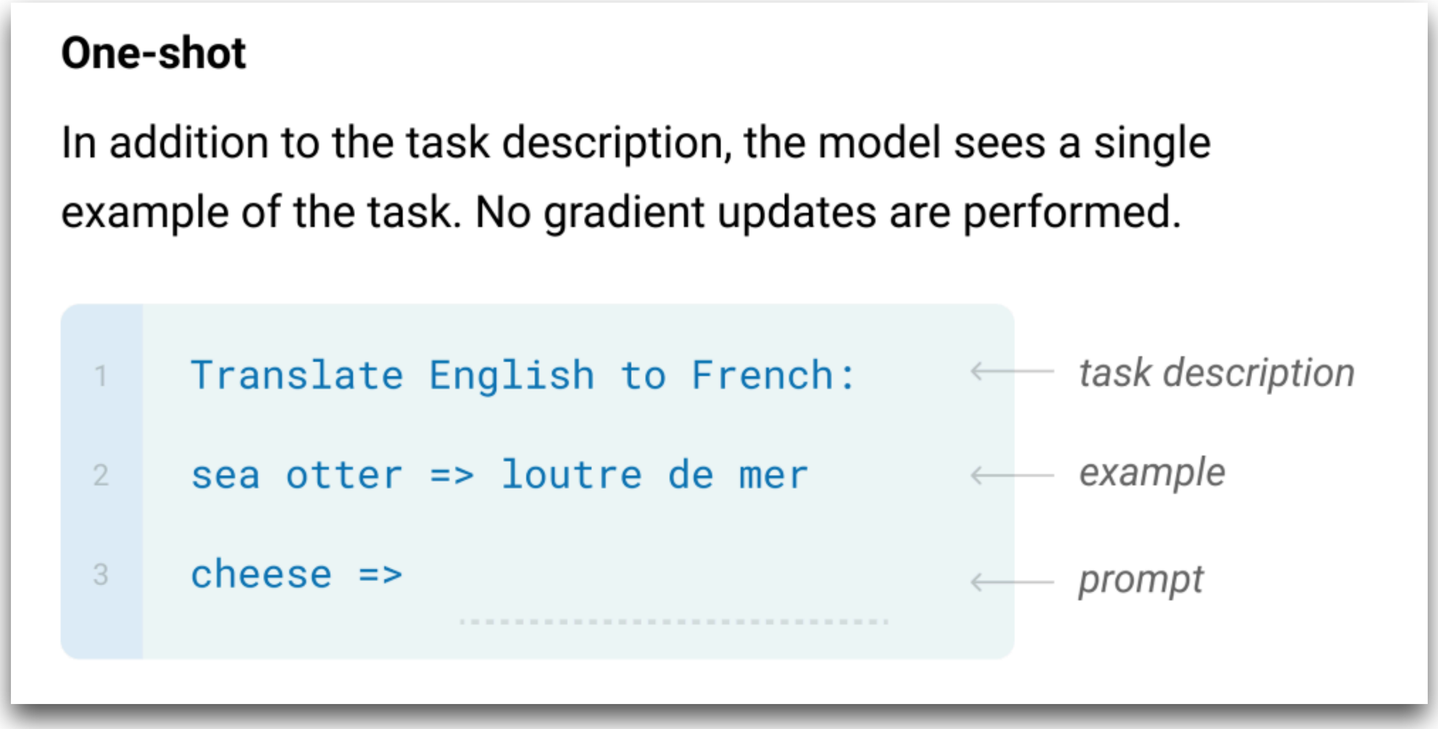
LM’s in-context learning ability



* Pictures are taken from Brown et al., 2020

LM’s in-context learning ability





* Pictures are taken from Brown et al., 2020

LM’s in-context learning ability

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

* Pictures are taken from Brown et al., 2020

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

We are moving towards

an unification era where one model can do many tasks!

* Pictures are taken from Brown et al., 2020

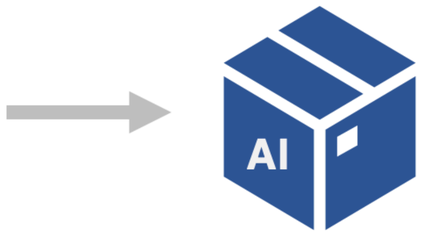
How can we build better model
that can generalize across various tasks?

NLP Before 2018: building task-specific models

Instance-level generalization within one task

Sentiment Analysis

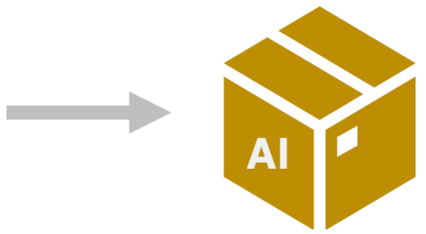
“ My experience has been fantastic! ”



“ Positive ”

Question Answering

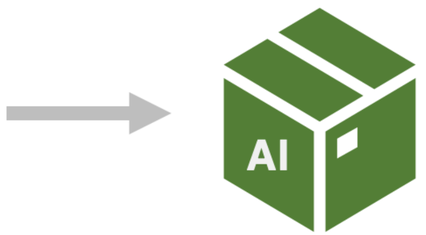
“ Where is World Cup 2022 playing? ”



“ Qatar. ”

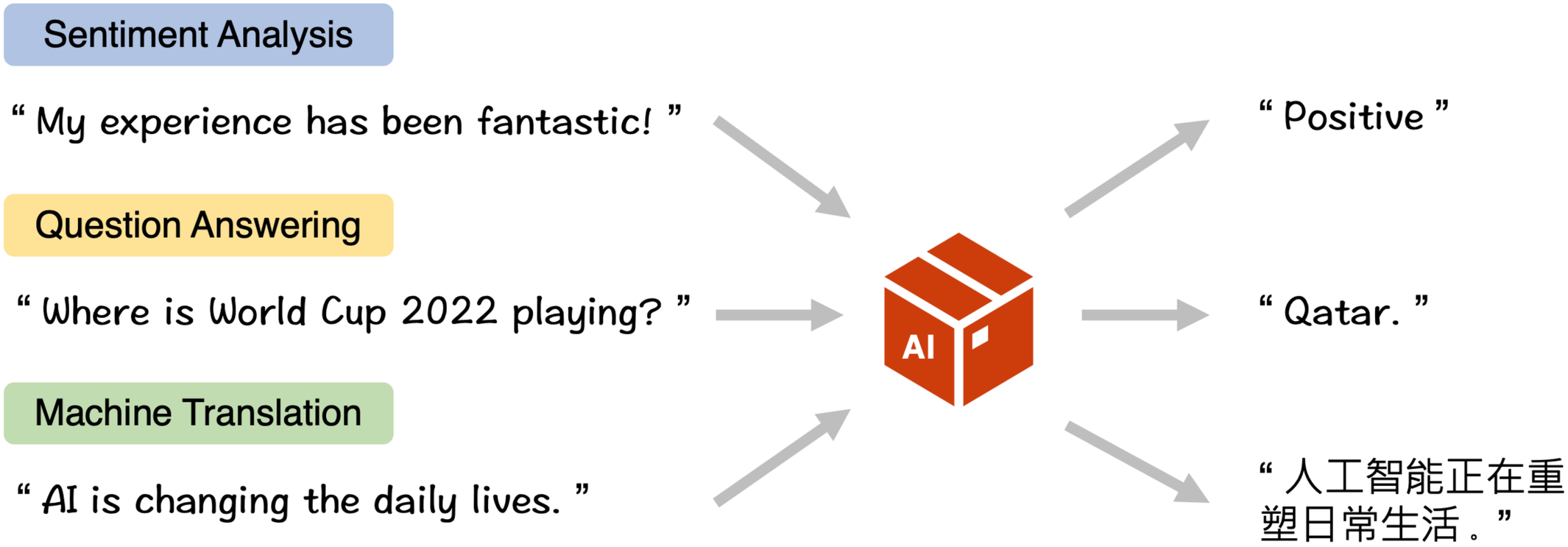
Machine Translation

“ AI is changing the daily lives. ”

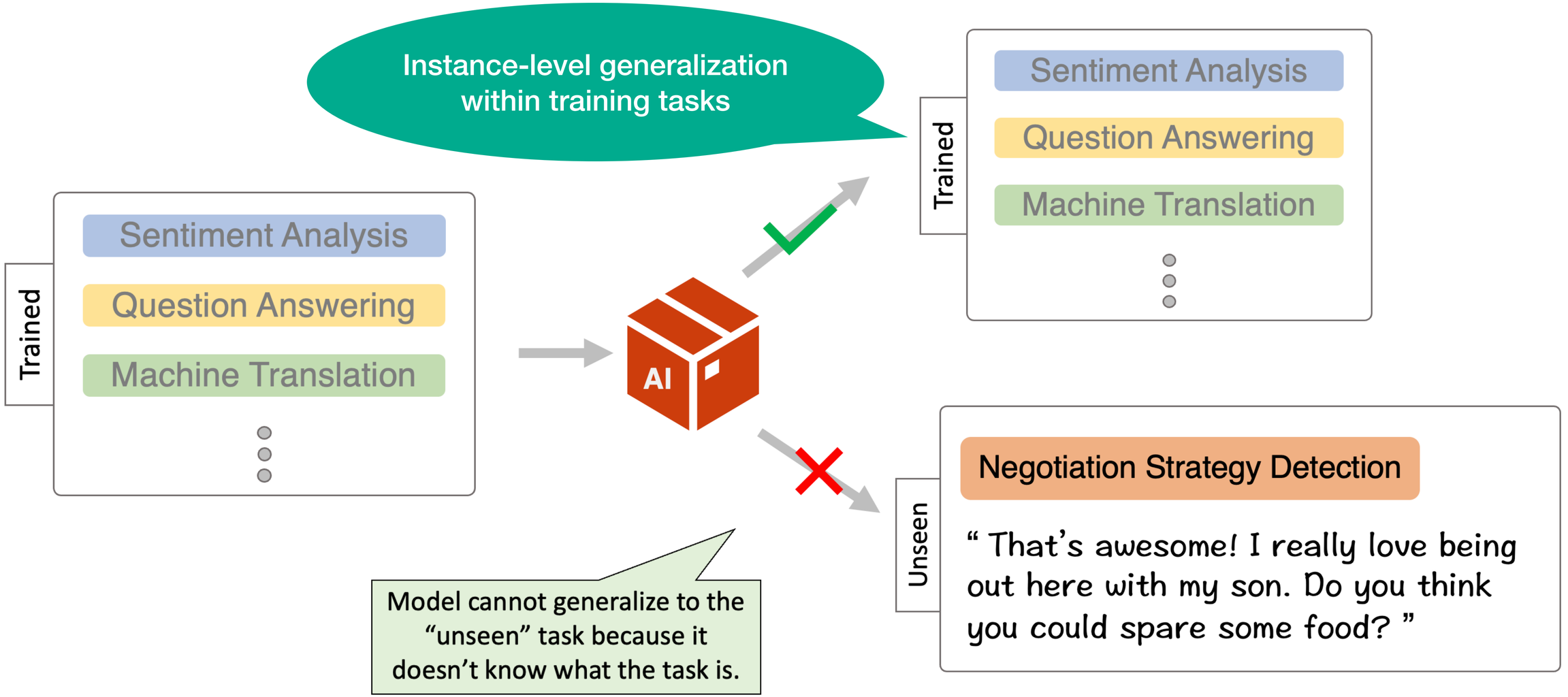


“ 人工智能正在重塑日常生活。”

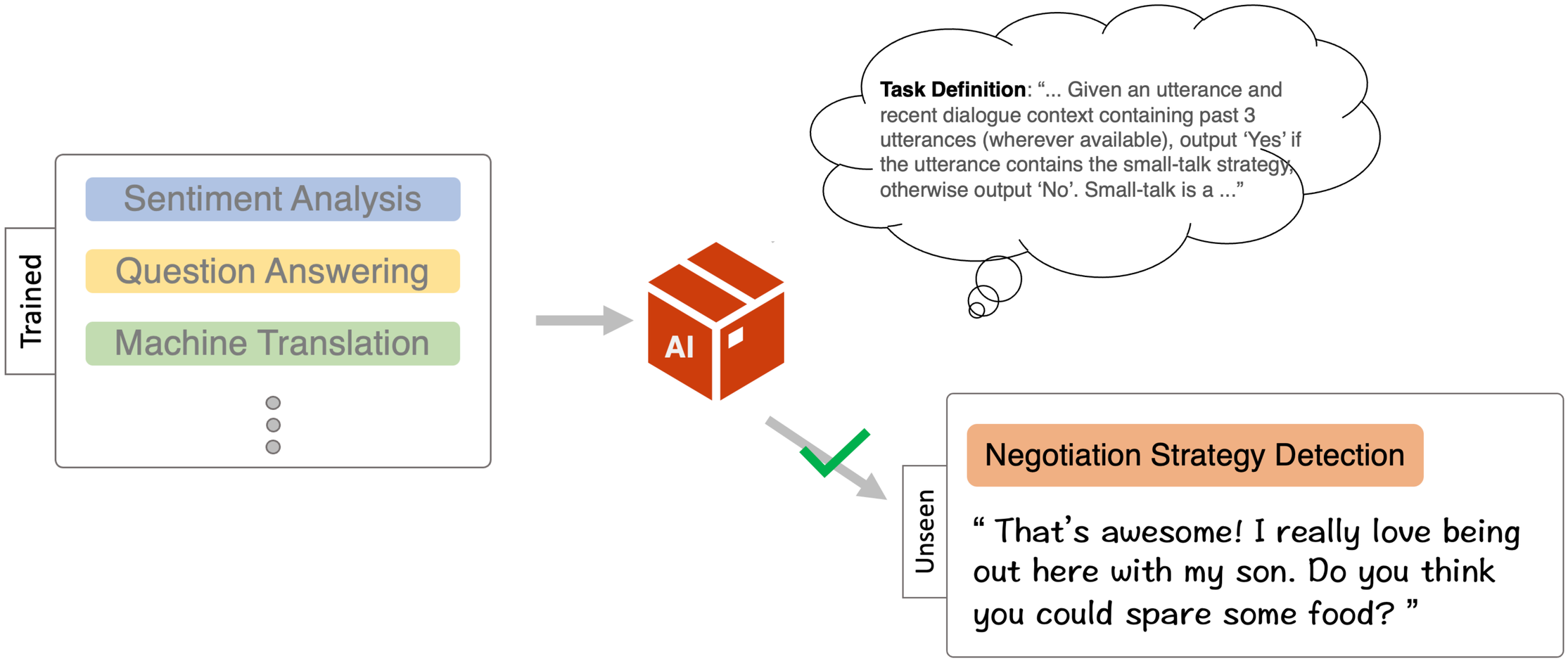
Classical multi-task learning (MTL)



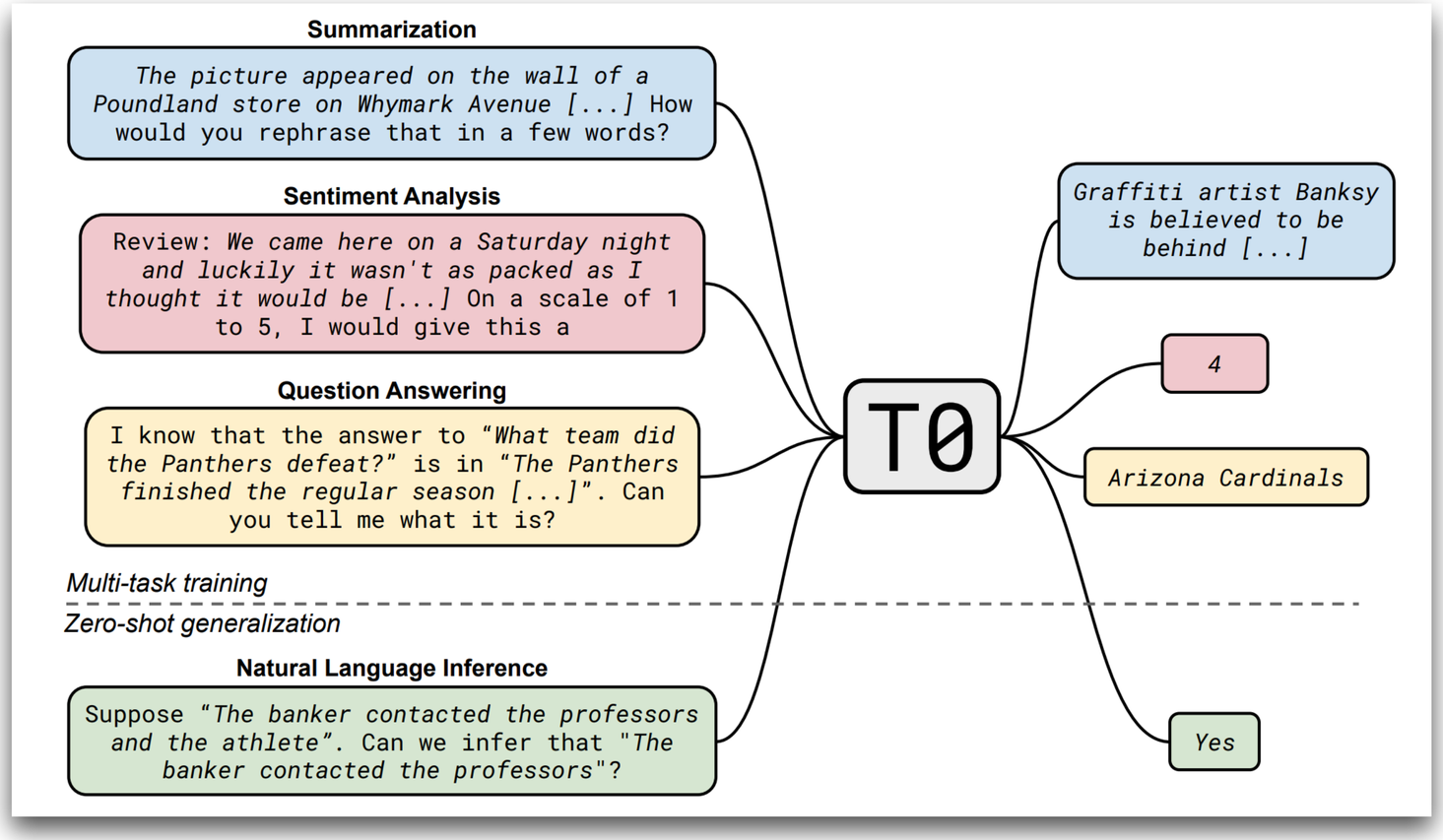
Classical MTL cannot generalize to unseen tasks



Cross-task generalization via instructions



Finetuning the model to follow instructions better (instruction tuning)



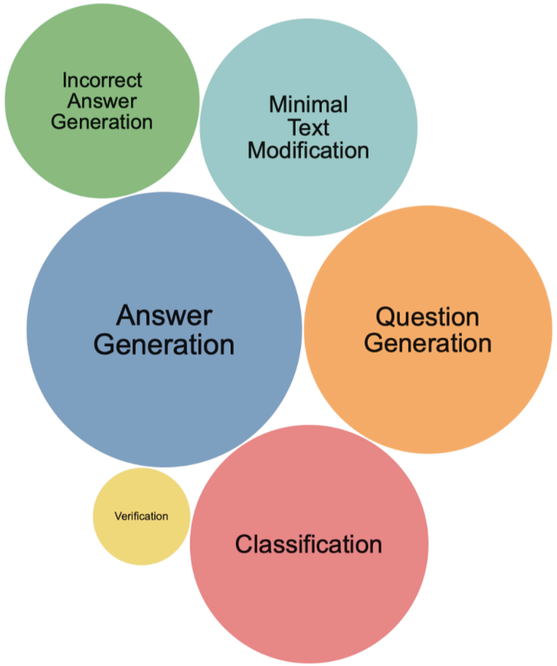
Task-level generalization across different tasks

* Figure taken from Sanh et al., 2022

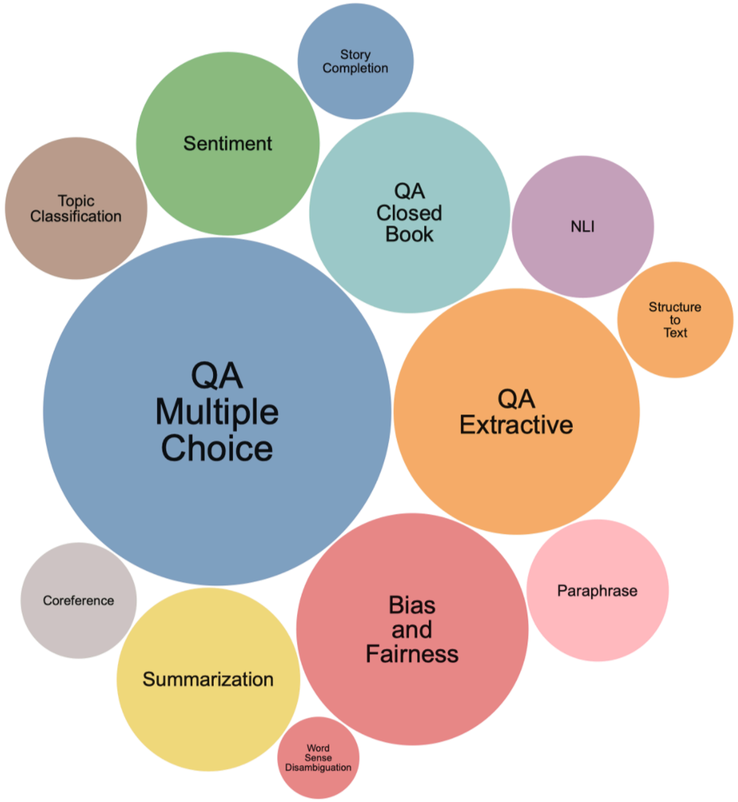
Instruction data (instruction-input-output triples)
is needed for such finetuning!

Existing instruction dataset in the field

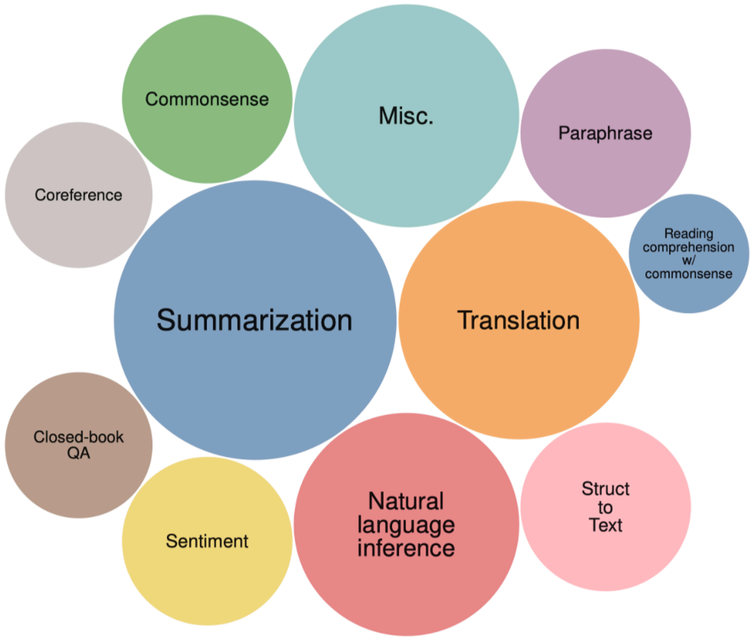
[1] Natural Instruction v1.0
(61 tasks)



[2] PromptSource
(176 tasks)



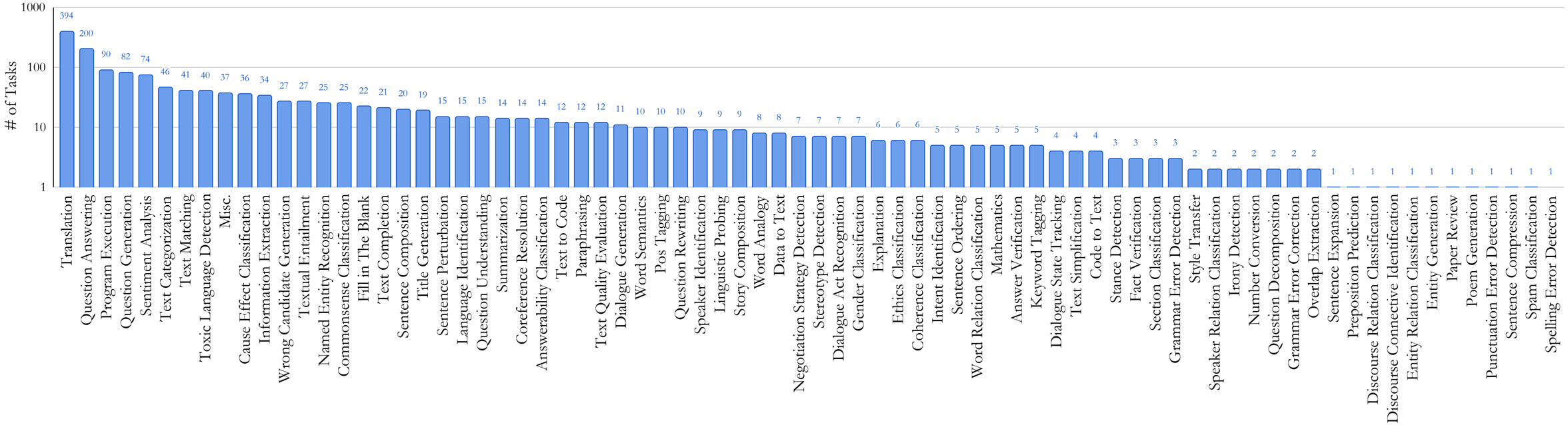
[3] FLAN
(62 tasks)



[1] Mishra et al. “Cross-Task Generalization via Natural Language Crowdsourcing Instructions”. ACL 2022.
[2] Sanh et al. “Multitask Prompted Training Enables Zero-Shot Task Generalization”. ICLR 2022.
[3] Wei et al. “Finetuned Language Models are Zero-Shot Learners.” ICLR 2022.

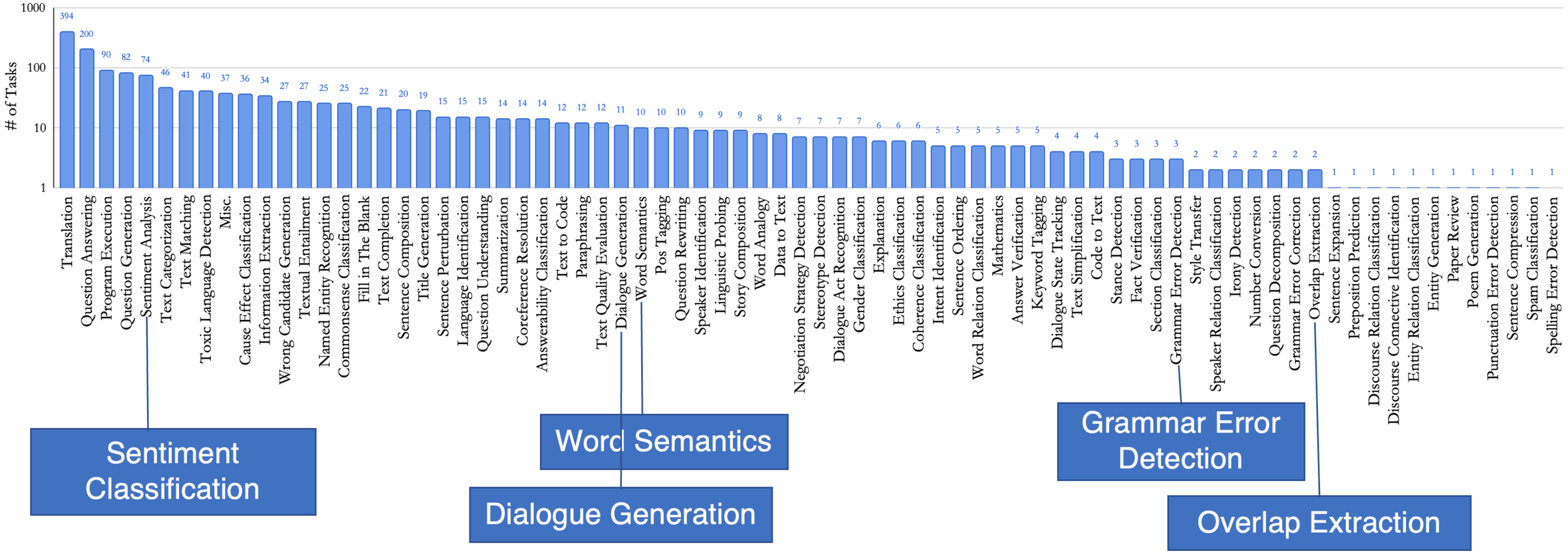
Overview of Super-NatruallInstructions

- 1616 tasks in 76 broad categories



Overview of Super-NatruallInstructions

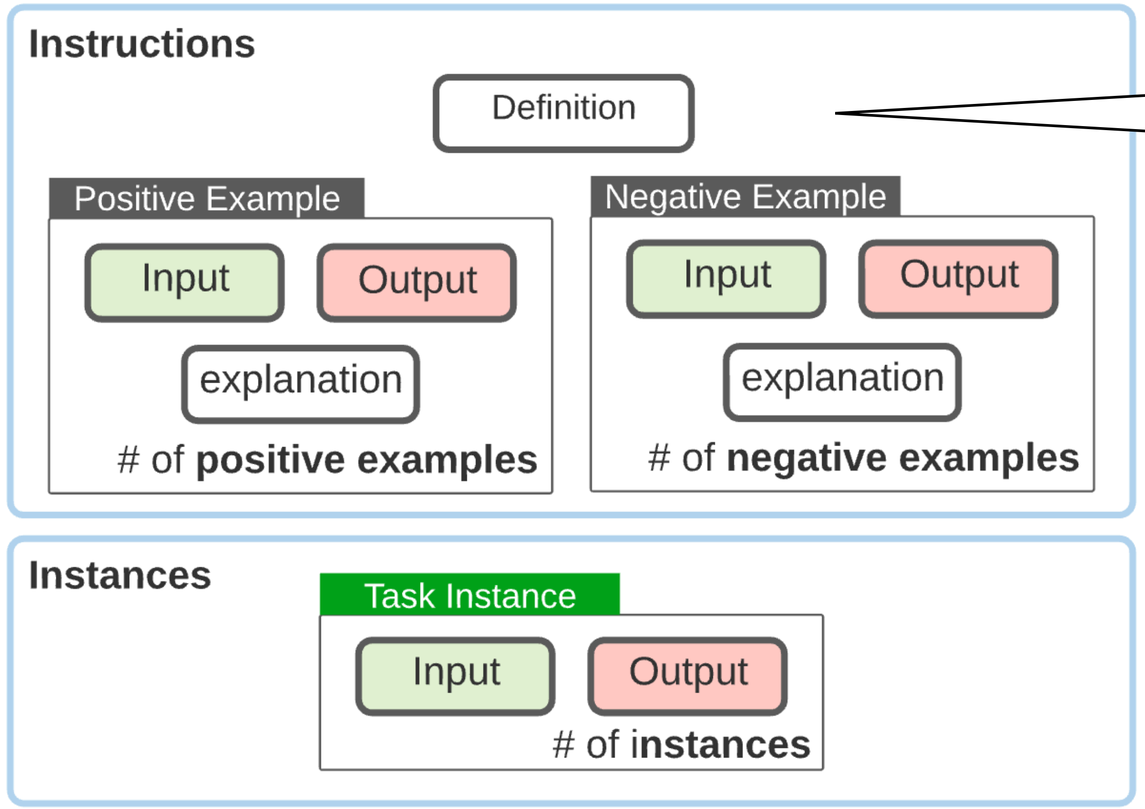
- 1616 tasks in 76 broad categories



Data collection process

- 1616 Tasks are sourced from:
  - existing public NLP datasets
  - available intermediate annotations in crowdsourcing experiments
  - synthetic tasks that can be communicated to an average human
- Contributed by 88 volunteer NLP practitioners.
- Collaborated via GitHub.
- Peer-reviewed by the lead authors to ensure quality (4-6 iterations)

Instruction schema

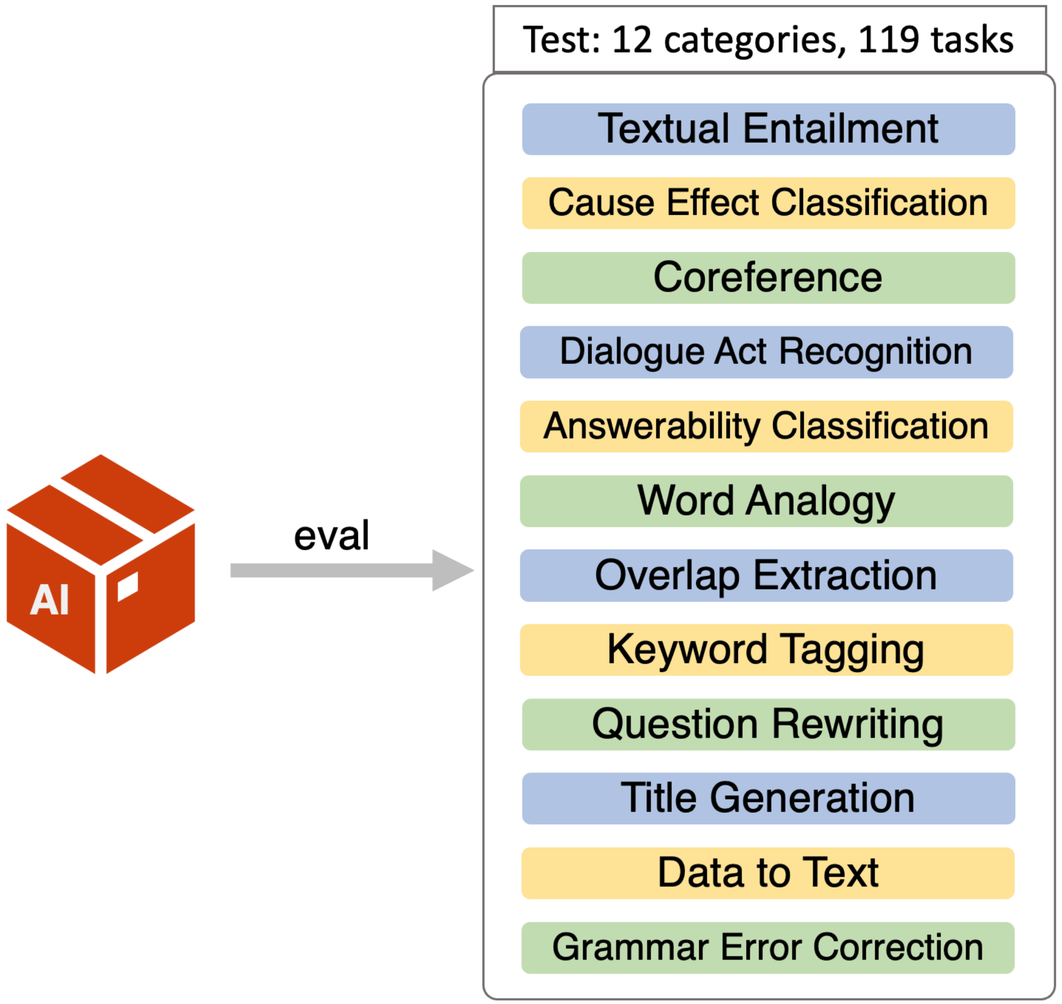


Definition is declarative, and should be sufficient to define the task to average human readers

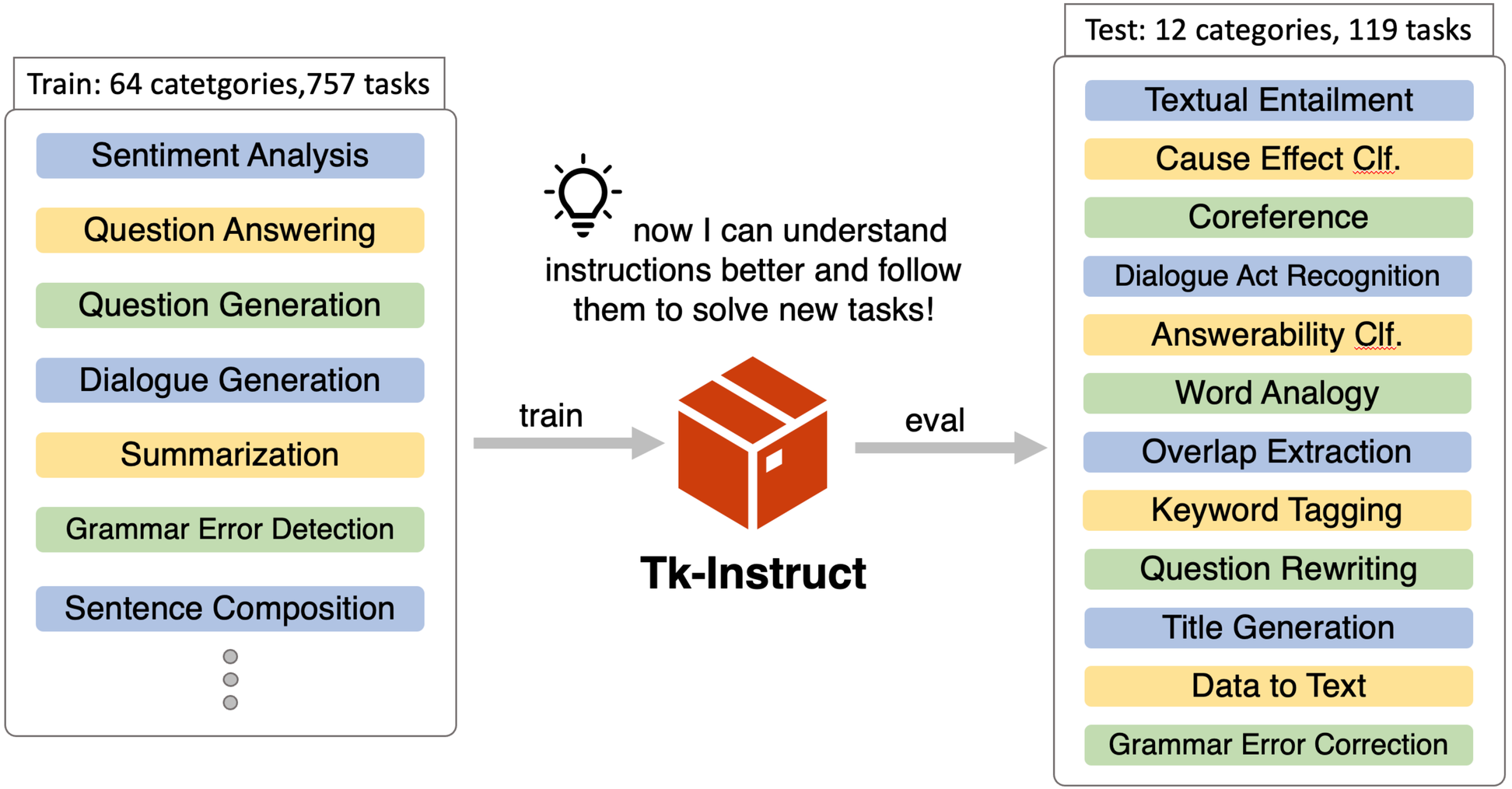
Example definition for SQuAD QA task

This task is about writing a correct answer for the reading comprehension task. Based on the information provided in a given passage, you should identify the shortest continuous text span from the passage that serves as an answer to the given question. Avoid answers that are incorrect or provides incomplete justification for the question.

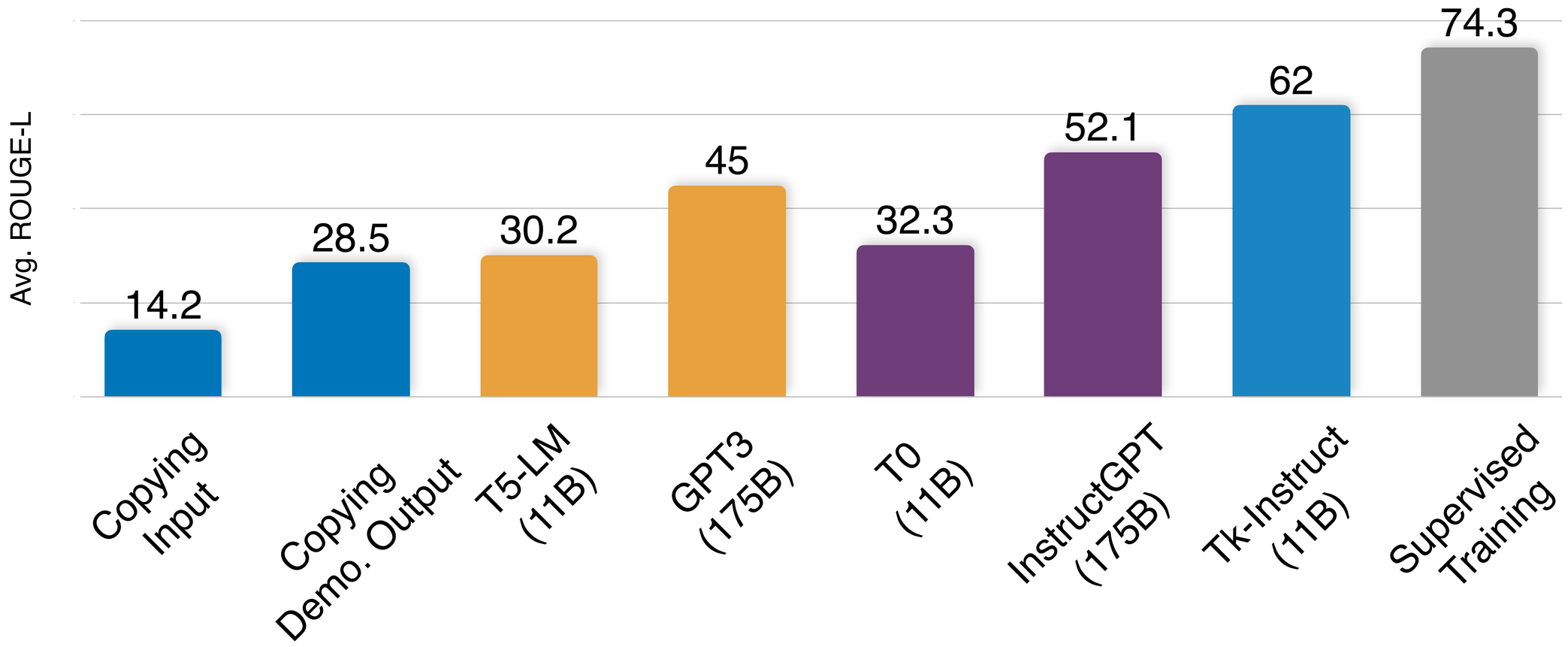
Benchmarking generalization via instructions



Tk-Instruct: an instruction-following model trained on our data

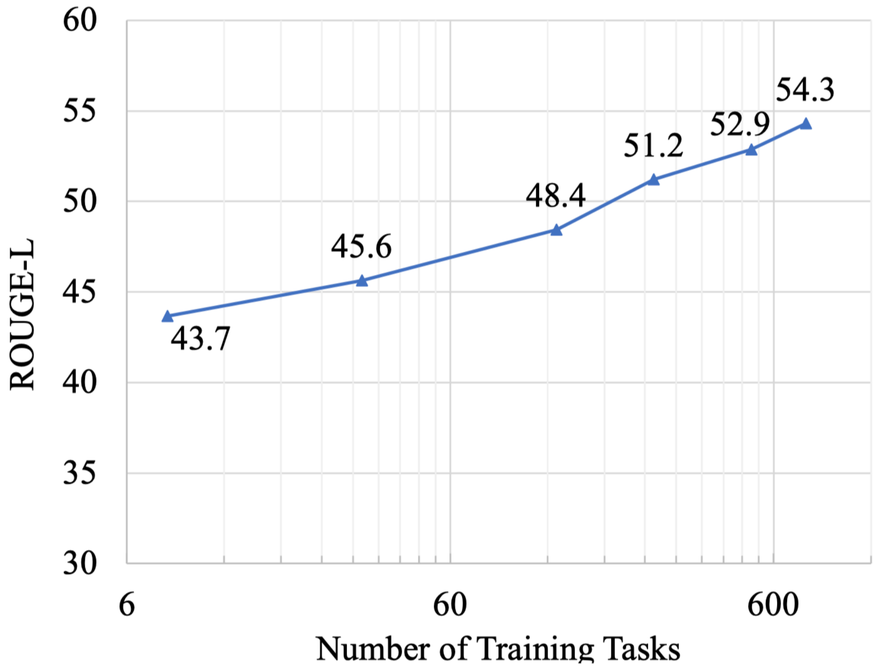


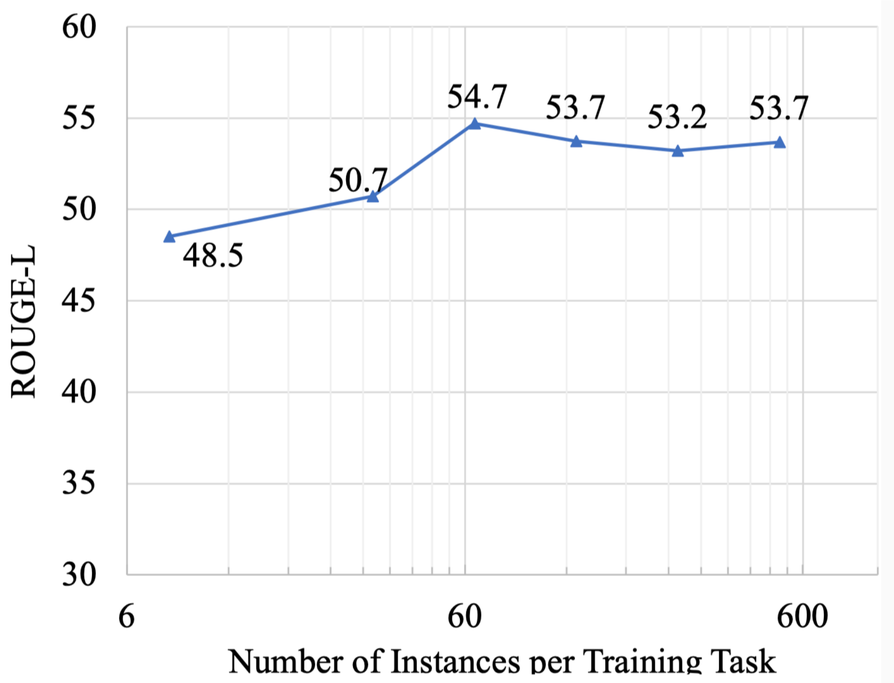
Model performance on the 119 testing tasks



Analysis: does more training data help?

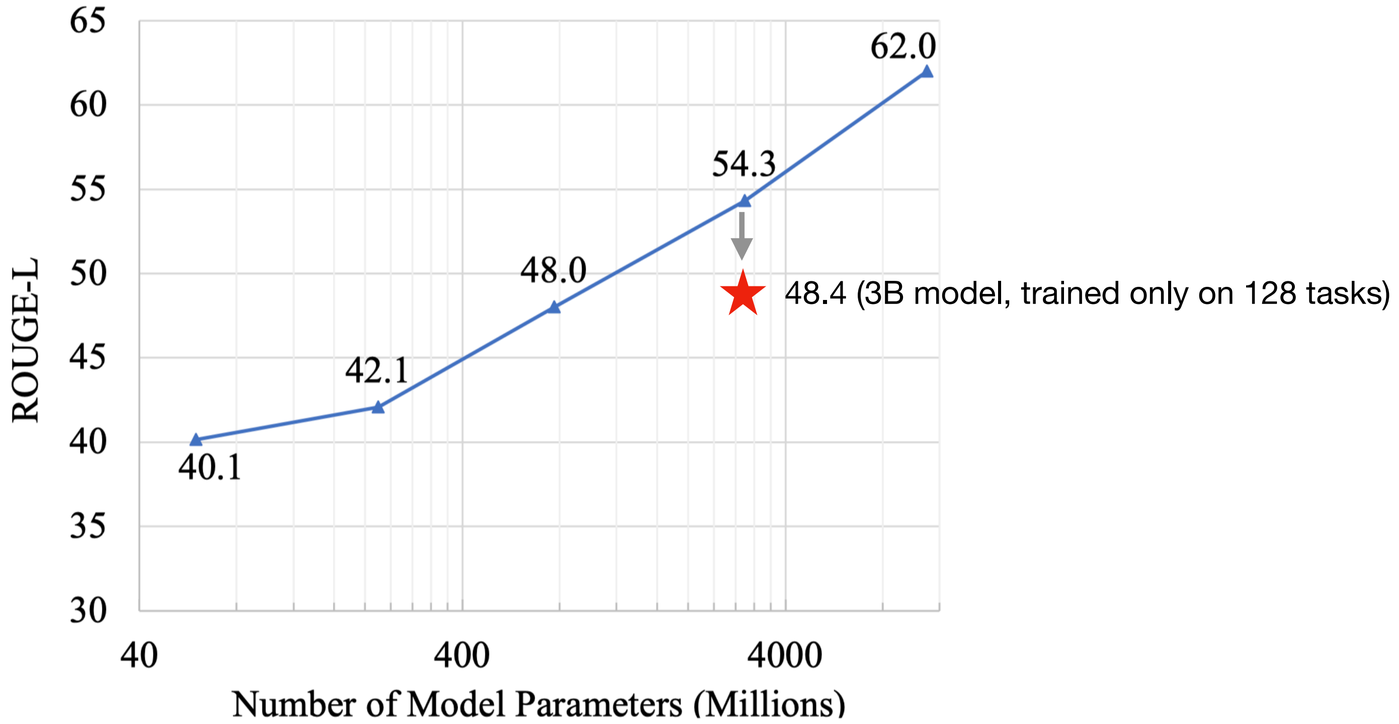
- More tasks lead to better performance.
- This trend slows down as when more tasks are used for training.
- Task (instruction) diversity matters - not the size of each individual dataset!



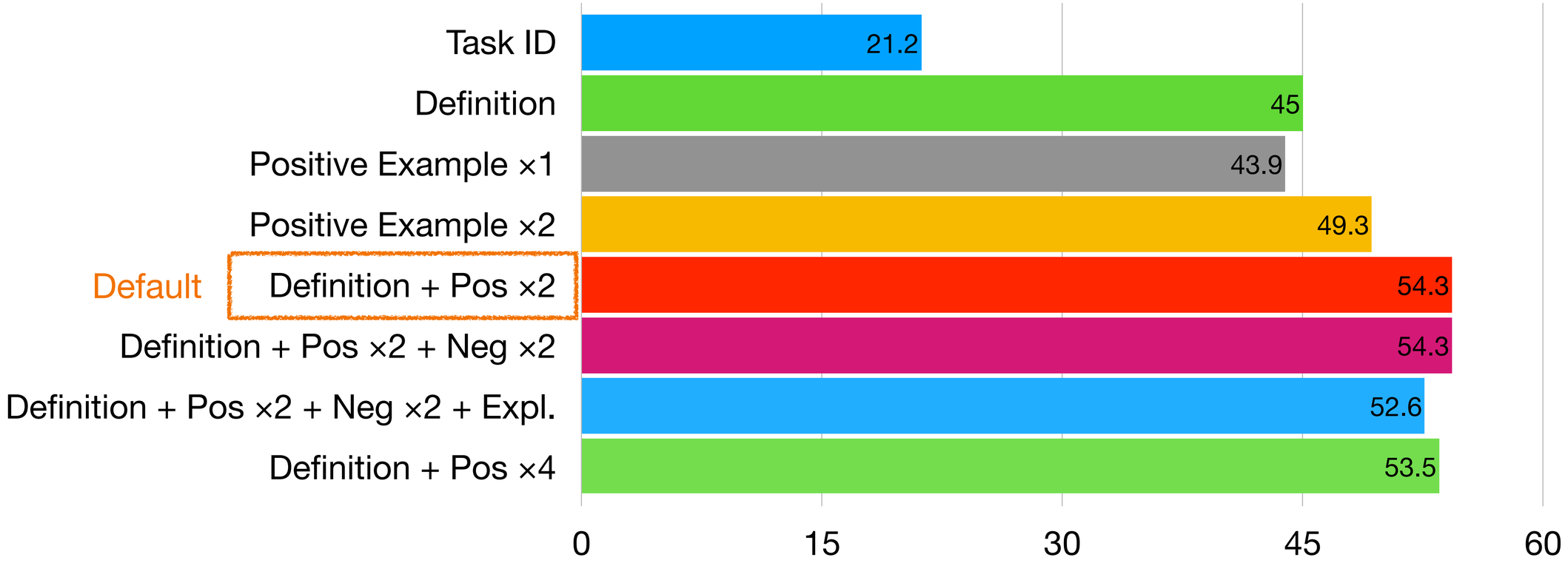


Analysis: model size scaling

- Bigger pre-trained models consistently lead to better final performance!
- Increasing the diversity of training tasks is complementary to scaling the model sizes.



Analysis: which instruction elements are the most helpful?



Takeaways

- Cross-task generalization via instructions is plausible.
- Super-NaturalInstructions provides a rich playground for such study.
- For instruction tuning:
  - Task/Instruction diversity is important!
  - Larger models bring in consistent improvement - not converged yet.
  - Large number of training instances could lead to overfitting to the training task.

Limitations

- Data:
  - Our instruction data is still limited in its style - usually long and wordy.
  - We mostly focus on existing NLP datasets, which are skewed to classification tasks.
  - We only annotated one instruction per task.
- Model:
  - Not robust to the input format (e.g., removing `output:` in the end of the prompt can break the model's generation).
  - Only T5 model series (encoder-decoder architecture) are tested.

Follow-up work

- FLAN-PaLM (Chung et al., 2022): combining instruction data and larger models.
- EditEval (Dwivedi-Yu et al., 2022): instruction-based text improvement.
- HyperTuning (Phang et al., 2022): instruction-based HyperNetwork.
- Coming soon:
  - Instruction-enabled unified text embedding model, SoTA on 70 embedding tasks.
  - Larger scale and more diverse instruction data generated by LM itself.
- Long-term:
  - Instruction following in a multi-modal setup, ideally in real-world scenarios.

Demo?

- https://instructions.apps.allenai.org/demo
- A GPT3 instruction-tuned on SuperNaturalInstructions

Thanks!

@yizhongwyz

yizhongw@cs.washington.edu

https://instructions.apps.allenai.org/

Performance on different task categories