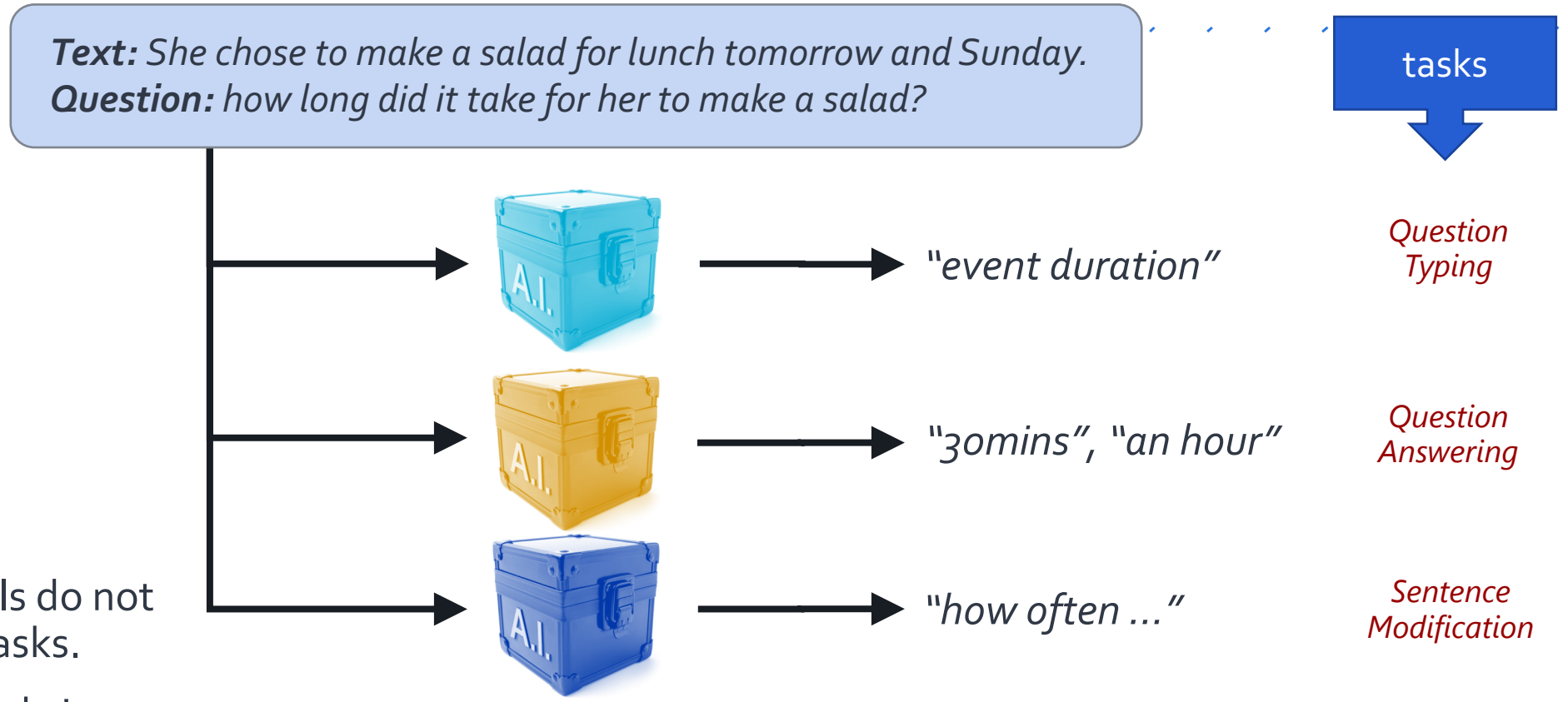
Cross-Task Generalization via Natural Language Instructions

Daniel Khashabi
Allen Institute for AI

Joint w/ Swaroop Mishra and Hanna Hajishirzi

Task-Specific Models



- Task-specific models do not generalize across tasks.
- There are MANY tasks!

Beyond Task-Specific Models

Text: *She chose to make a salad for lunch tomorrow and Sunday.*
Question: *how long did it take for her to make a salad?*

tasks

"event duration" — *Question Typing*

"30mins", "an hour" — *Question Answering*

"how often ..." — *Sentence Modification*

Beyond Task-Specific Models

Text: *She chose to make a salad for lunch tomorrow and Sunday.*
Question: *how long did it take for her to make a salad?*

tasks

`question-typing`

`question-answering`

`sentence-modification`

"event duration" — *Question Typing*

"30mins", "an hour" — *Question Answering*

"how often ..." — *Sentence Modification*

Do not generalize to "unseen" tasks.

[Raffel et al. 2020]

Beyond Task-Specific Models

human readable definitions; fully define the task

Text: *She chose to make a salad for lunch tomorrow and Sunday.*
Question: *how long did it take for her to make a salad?*

tasks

Question Typing Instructions

Indicate the type of temporal phenomenon in the question …

Question Answering Instructions

In this task we ask you to write answer to the given question. …

Sentence Mod. Instructions

Label a question that is free of any grammatical/logical errors w/ 'yes' …

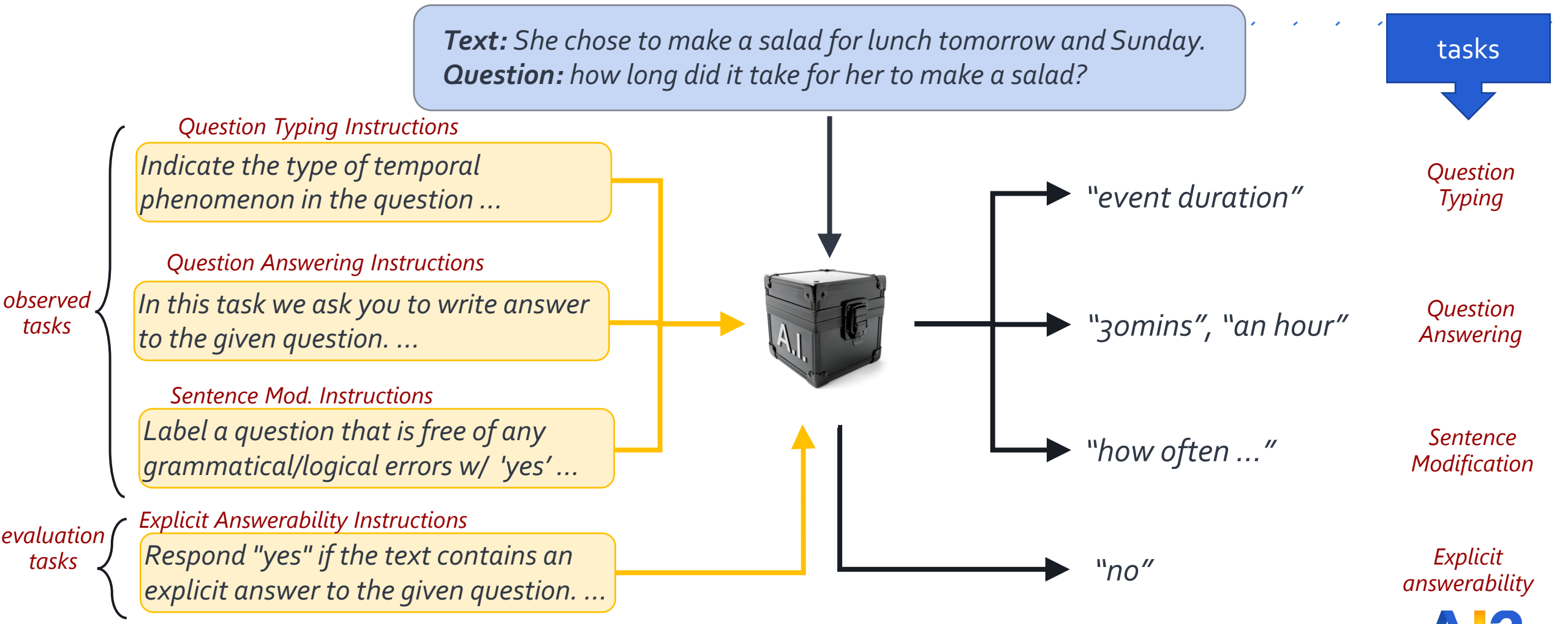
"event duration" — *Question Typing*

"30mins", "an hour" — *Question Answering*

"how often …" — *Sentence Modification*

- Instruction define tasks explicitly in natural language.

Cross-Task Generalization



Instructions Paradigm: Challenges

1. There is no benchmarks containing natural language instructions for a diverse range of of tasks.

We present a dataset of **natural instructions** for a **wide variety** of tasks!

2. Unclear whether models benefit from task "instructions".

We show empirical evidence of their **benefits**!

Natural-Instructions: Overview

- Natural Instructions:
  - 61 tasks and instructions
  - 160k instances (input -> outputs)

Input: *She chose to make a salad for lunch tomorrow and Sunday.*

Instructions: generating "duration" questions task

In this task, we ask you to write a question that involves "event duration", based on a given sentence. Here, event duration is defined as the understanding of how long events typically last. For example, "brushing teeth", usually takes few minutes. ....

Output: *"how long did it take for her to make a salad?*

Natural-Instructions: Construction (1)

- Use existing datasets and the instructions used to crowdsource them

(Efrat & Levy, 2020)

A typical data construction pipeline

collecting existing datasets

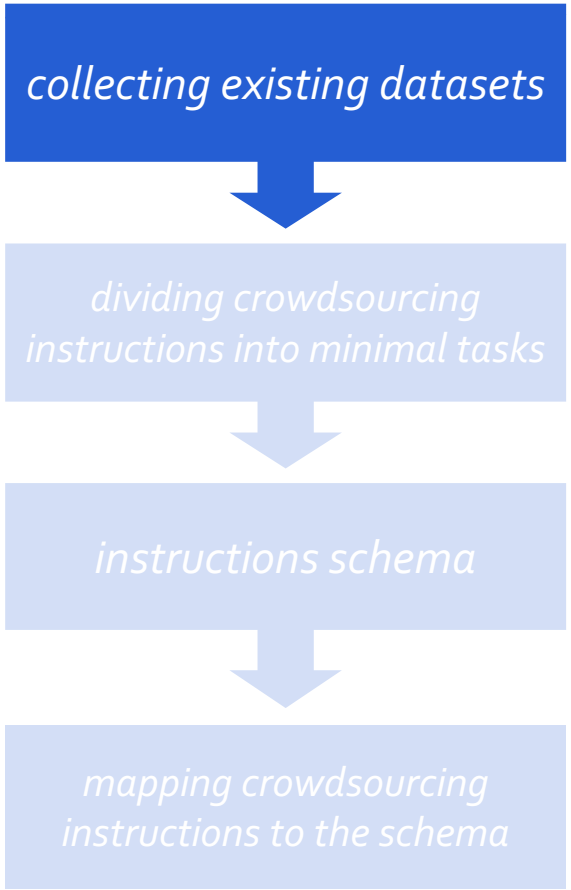
dividing crowdsourcing instructions into minimal tasks

instructions schema

mapping crowdsourcing instructions to the schema

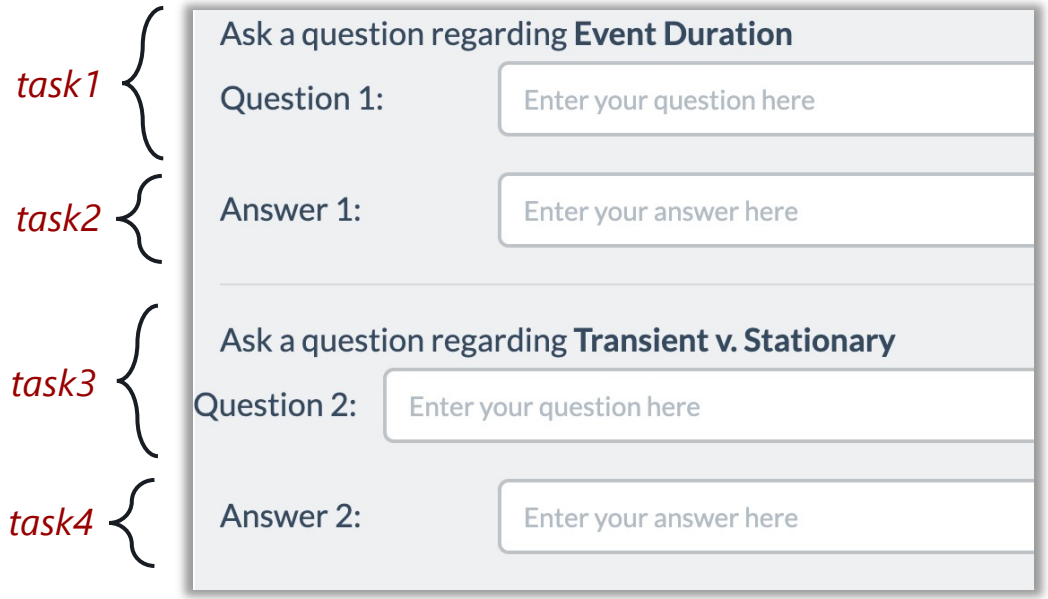
Natural-Instructions: Construction (1)

- Contacted dataset authors to access their crowdsourcing instructions and the associated annotations
  1. CosmosQA [Huang et al. 2019]
  2. DROP [Dua et al. 2019]
  3. Essential-Terms [Khashabi et al. 2017]
  4. MCTACO [Zhou et al. 2019]
  5. MultiRC [Khashabi et al. 2018]
  6. QASC [Khot et al. 2020]
  7. Quoref [Dasigi et al. 2019]
  8. ROPES [Lin et al. 2019]
  9. Winogrande [Sakaguchi et al. 2020]



Natural-Instructions: Construction (2)

- Crowdsourcing instructions tend to involve multiple annotation steps.
- Split them to self-contained tasks.



MC-TACO
[Zhou et al. 2019]

collecting existing datasets

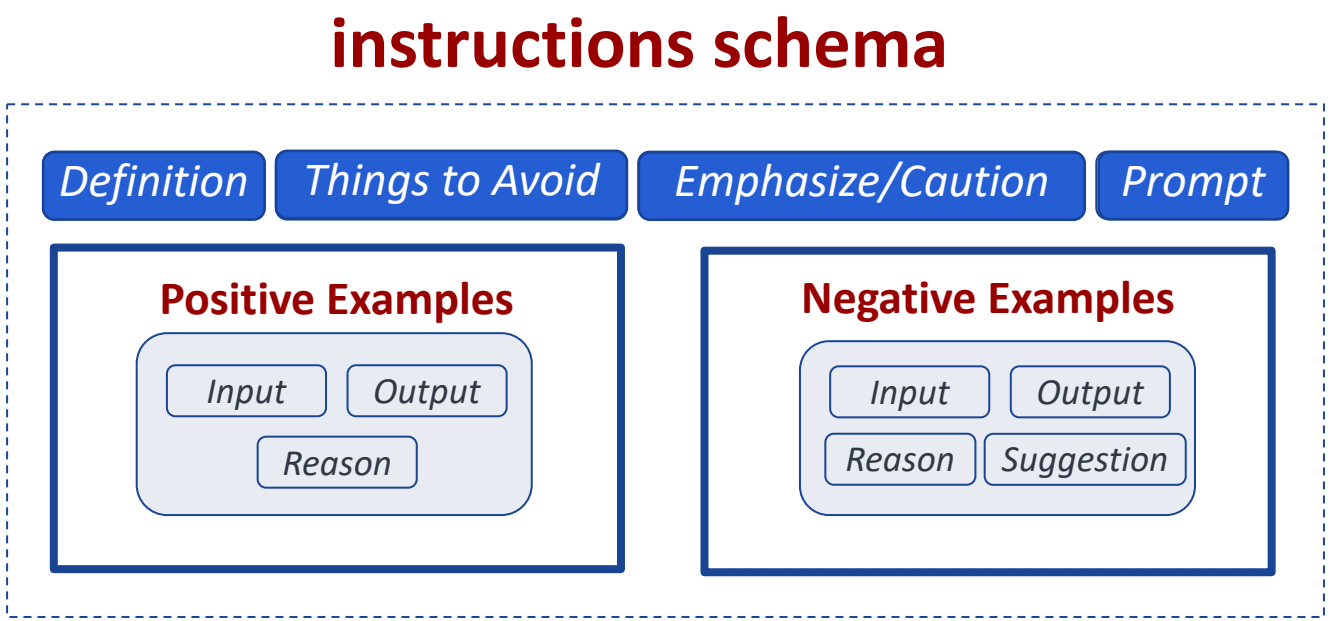
dividing crowdsourcing instructions into minimal tasks

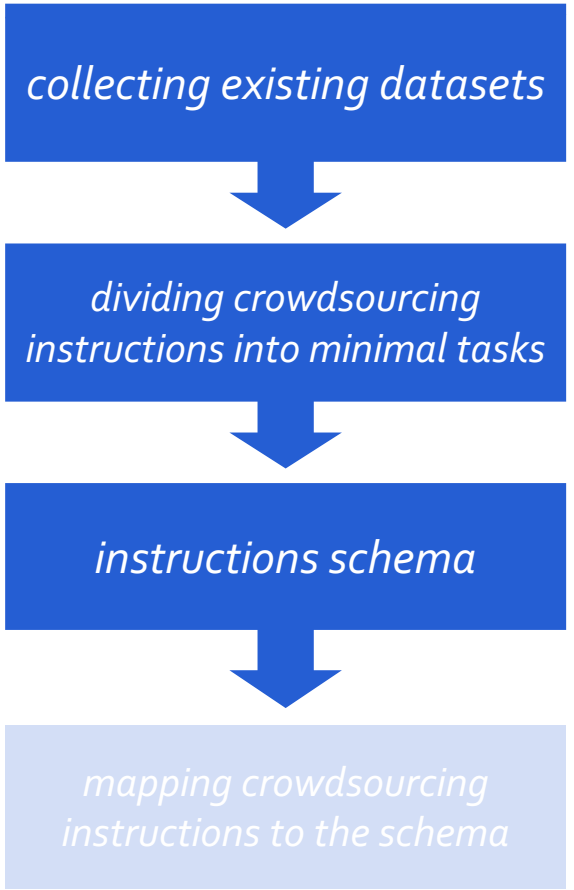
instructions schema

mapping crowdsourcing instructions to the schema

Natural-Instructions: Construction (3)

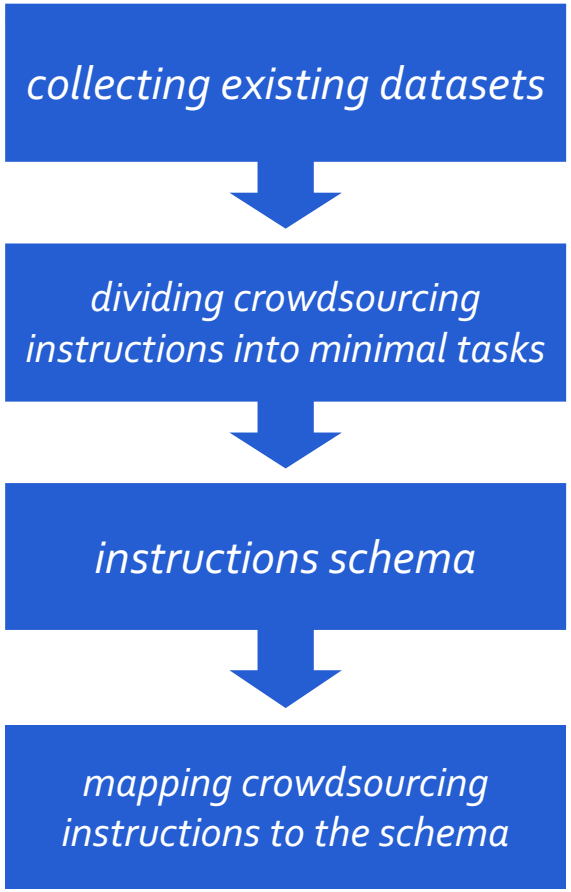
- Crowdsourcing instructions are written in a variety of ways.
- A unified schema for consistent representation across tasks.





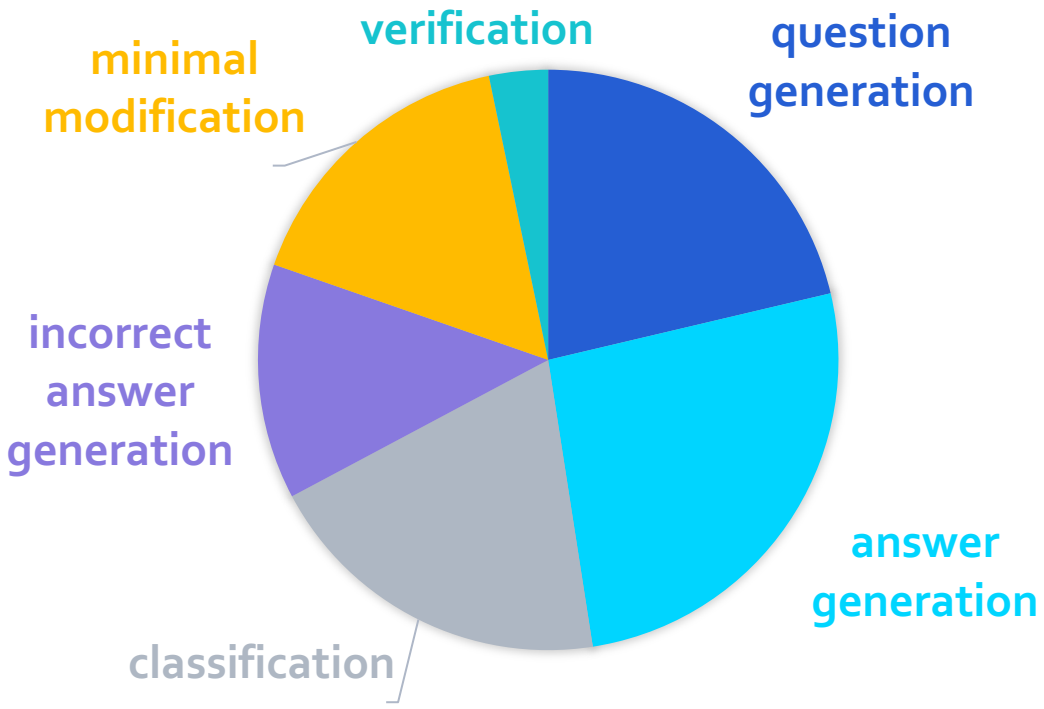
Natural-Instructions: Construction (4)

- This process was done by an expert annotator and verified by another.
- Mapping crowdsourcing instructions to our schema:
  - Retained the original phrasing.
  - Redacted verbose/repetitive content.
  - Created negative examples wherever they were absent.
- Took ~10 hours for each task.

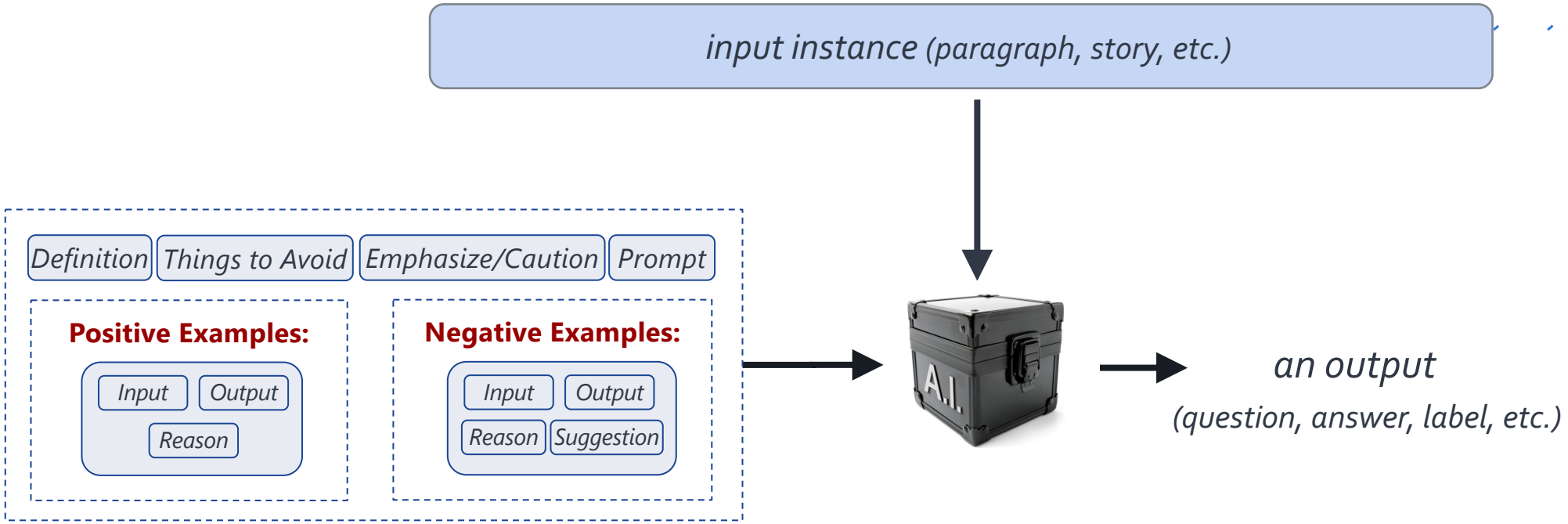


Natural Instructions: Statistics

- 61 tasks



Natural-Instructions: Example



Natural-Instructions: Example

input instance (paragraph, story, etc.)

Generating "duration" questions task

Definition: *In this task, we ask you to write a question that involves "event duration", based on a given sentence. Here, event duration is defined as the understanding of how long events typically last. For example, "brushing teeth", usually takes few minutes.*

Things to Avoid: *Don't create questions which have explicit mentions of answers in text. Instead, it has to be implied from what is given. In other words, we want you to use "instinct" or "common sense".*

Emphasize/Caution: *The written questions are not required to have a single correct answer.*

Positive example 1
Input: *Sentence: Jack played basketball after school, after which he was very tired.*
Output: *How long did Jack play basketball?*
Reason: *The question asks about the duration of an event; therefore it's a temporal event duration question.*

...

an output

(question, answer, label, etc.)

Natural-Instructions: Example

Input: *She chose to make a salad for lunch tomorrow and Sunday.*

Generating "duration" questions task

Definition: *In this task, we ask you to write a question that involves "event duration", based on a given sentence. Here, event duration is defined as the understanding of how long events typically last. For example, "brushing teeth", usually takes few minutes.*

Things to Avoid: *Don't create questions which have explicit mentions of answers in text. Instead, it has to be implied from what is given. In other words, we want you to use "instinct" or "common sense".*

Emphasize/Caution: *The written questions are not required to have a single correct answer.*

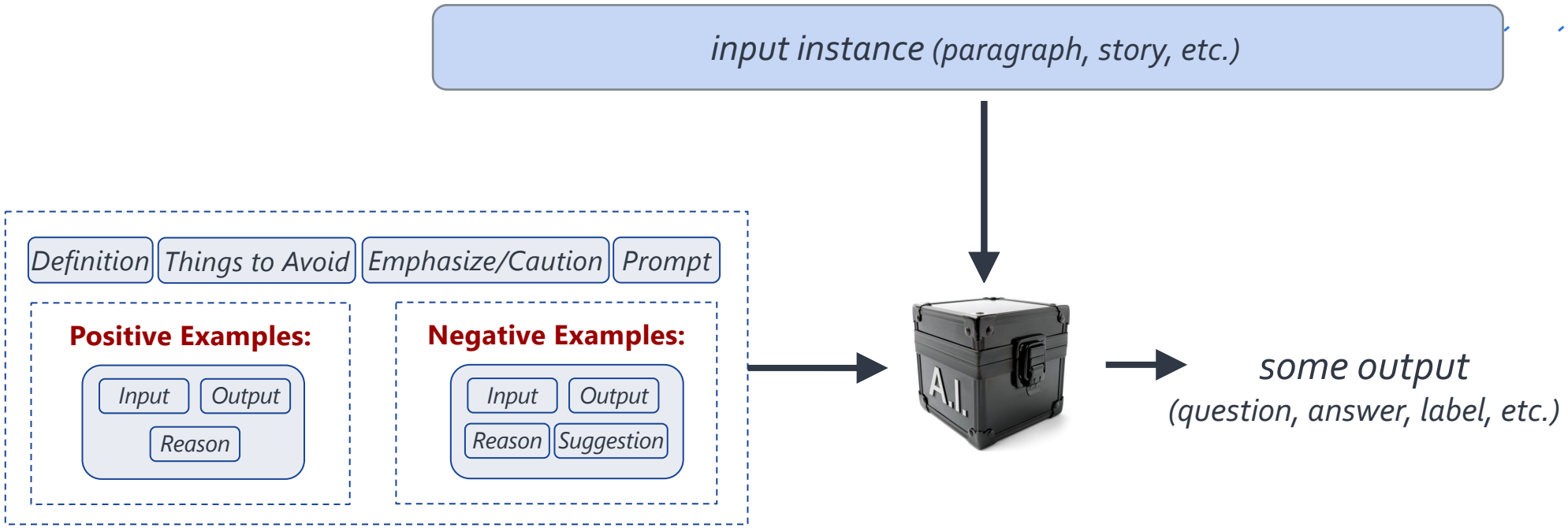
Positive example 1
Input: *Sentence: Jack played basketball after school, after which he was very tired.*
Output: *How long did Jack play basketball?*
Reason: *The question asks about the duration of an event; therefore it's a temporal event duration question.*

...

"how long did it take for her to make a salad?

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

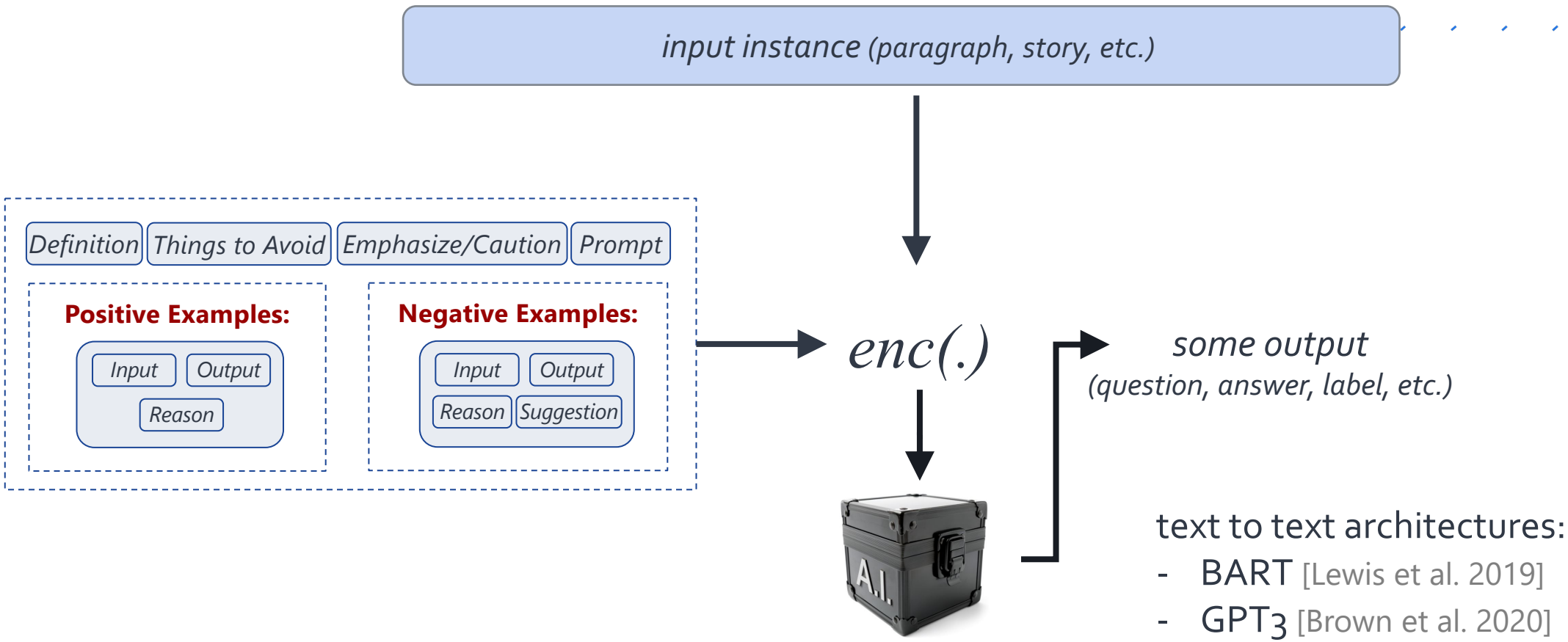
Encoding Instructions



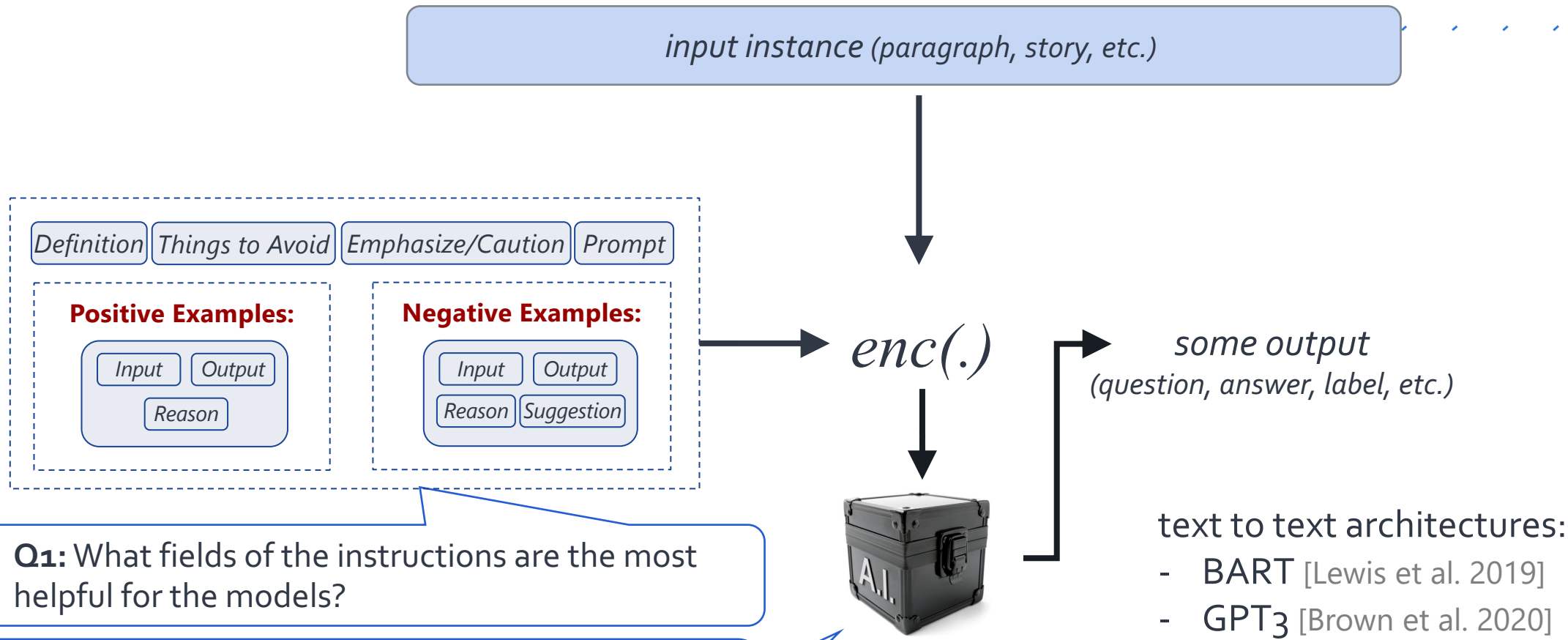
text to text architectures:

- BART [Lewis et al. 2019]
- GPT3 [Brown et al. 2020]

Encoding Instructions



Empirical Questions

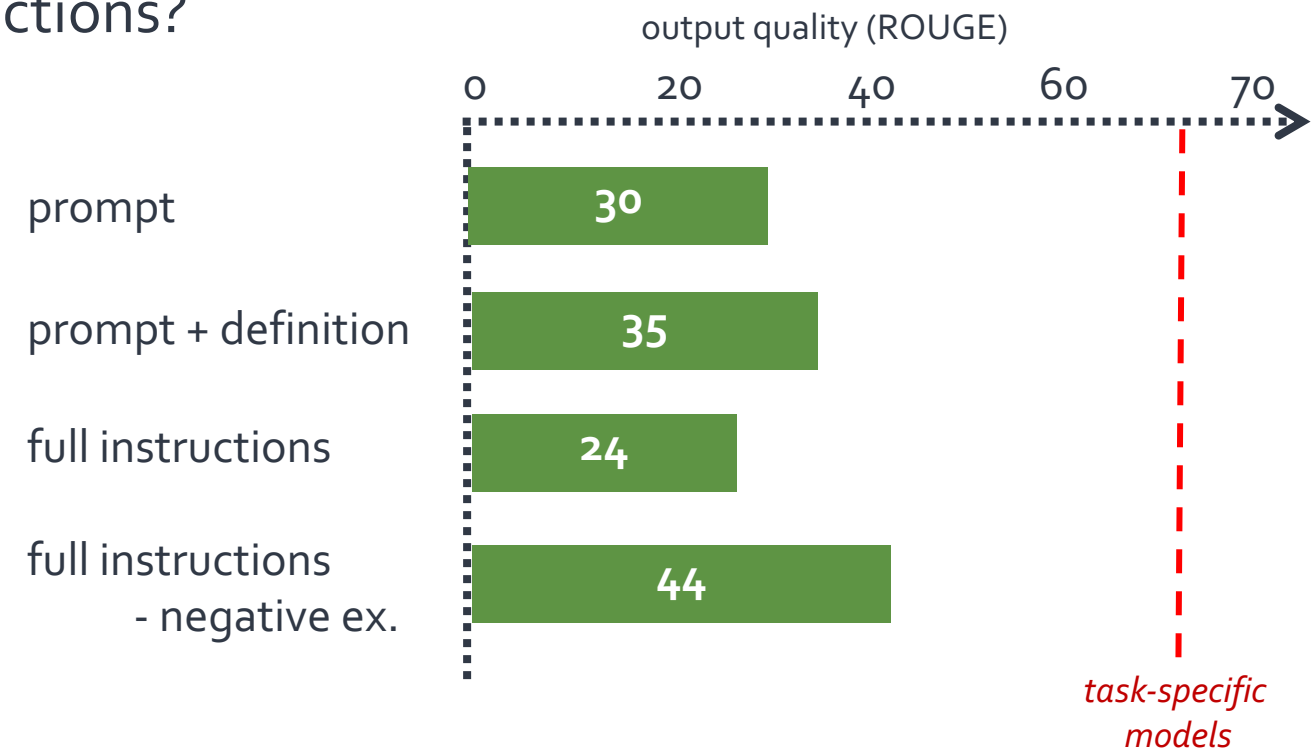


- **Q1:** What fields of the instructions are the most helpful for the models?

- **Q2:** Does GPT3 understand instructions?
- **Q3:** Can smaller model (such as BART) be tuned to follow instructions?

Experiment: Evaluating GPT3

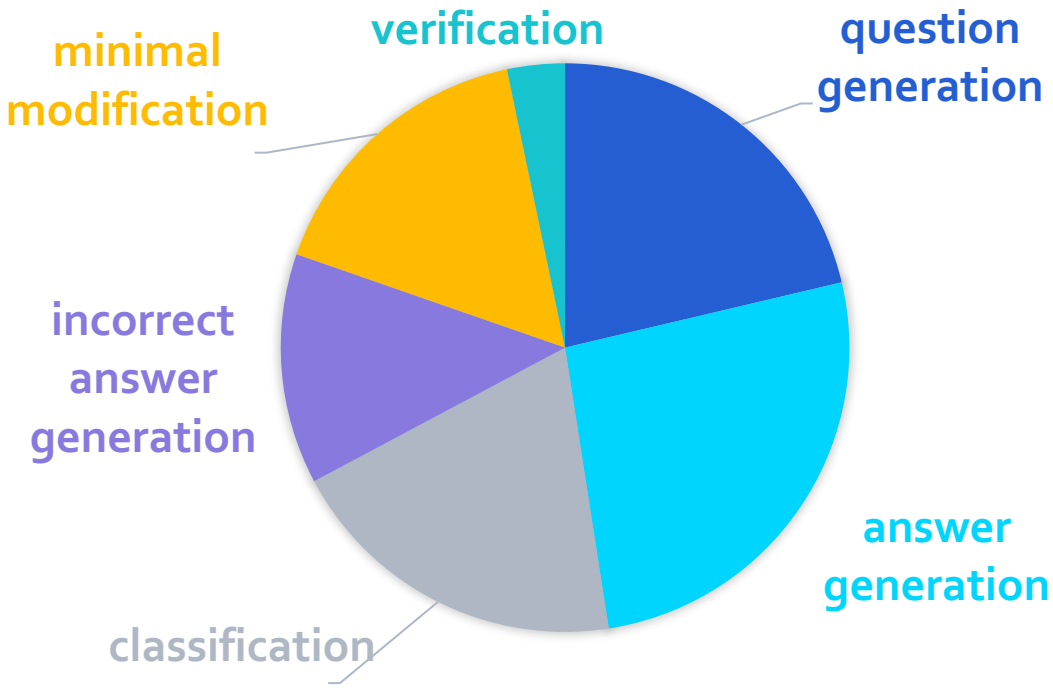
- Does GPT3 understand task instructions?
- Instructions **improve** GPT3's performance! 🎉
- All instruction elements (except negative examples) help!
- A wide margin to be solved 🤔
  - A task-specific BART scores ~70%



Evaluating Fine-tuned Models: Setup

- Splitting the data for fine-tuning a smaller model:

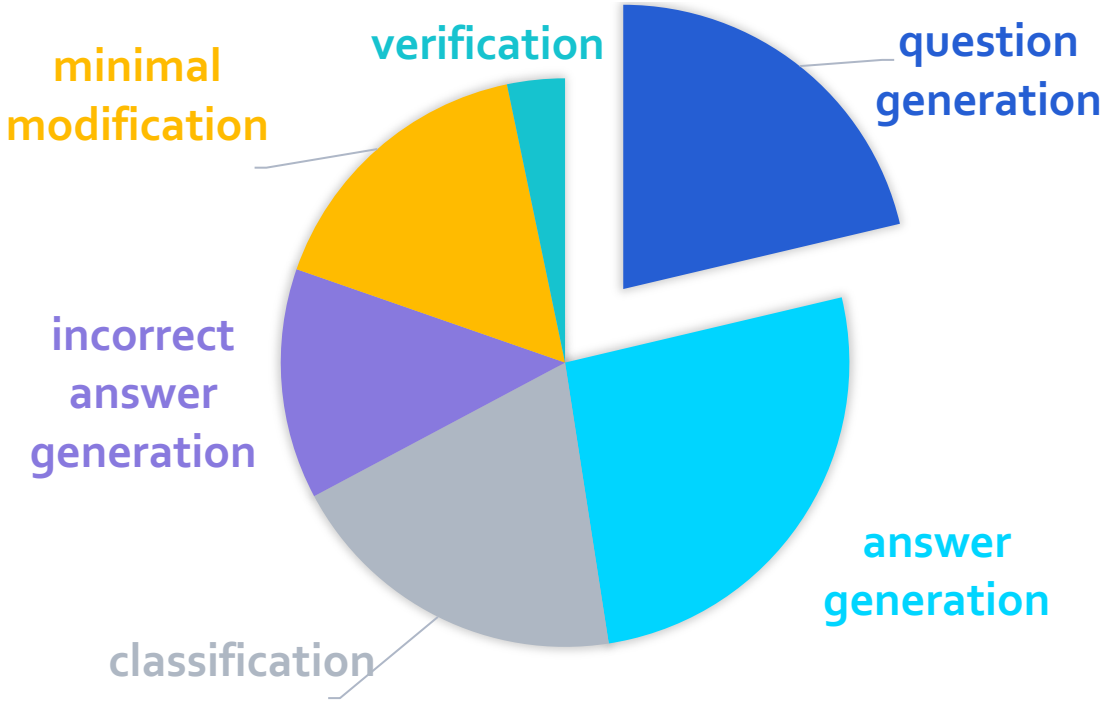
1. Randomly split the tasks
   - 12 **evaluation** tasks
   - 49 **supervision** tasks
2. Leave-one-**category**-out



Evaluating Fine-tuned Models: Setup

- Splitting the data for fine-tuning a smaller model:

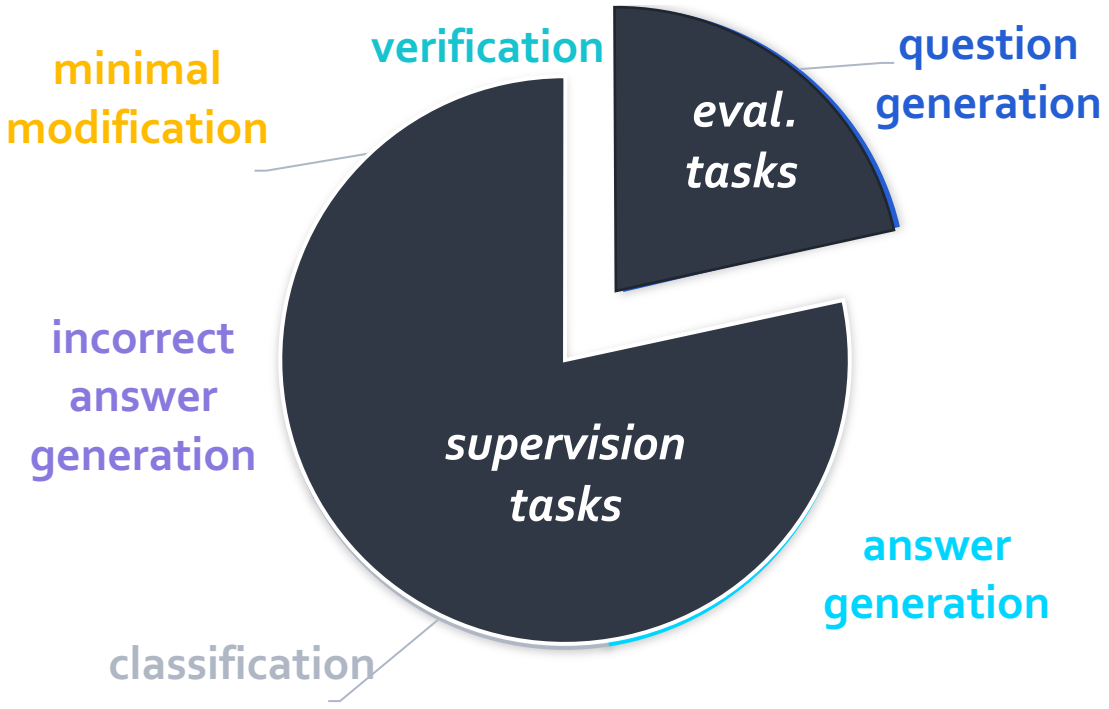
1. Randomly split the tasks
   - 12 **evaluation** tasks
   - 49 **supervision** tasks
2. Leave-one-**category**-out



Evaluating Fine-tuned Models: Setup

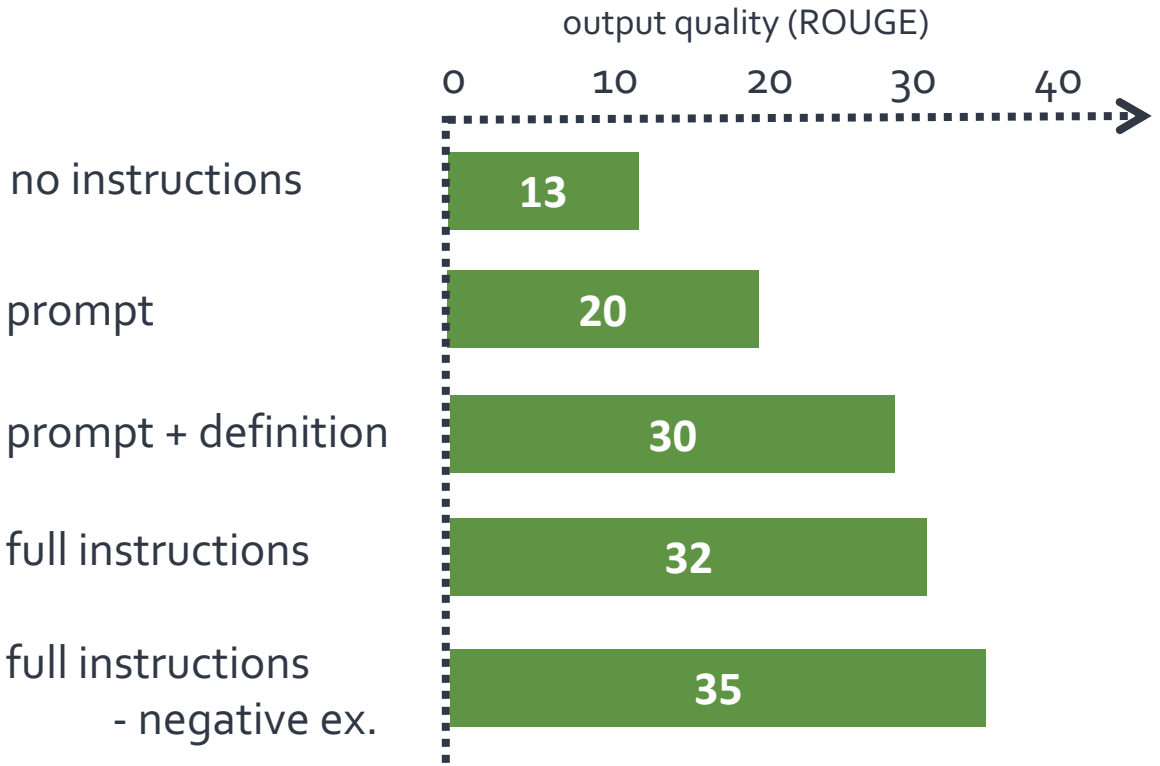
- Splitting the data for fine-tuning experiments:

1. Randomly split the tasks
   - 12 **evaluation** tasks
   - 49 **supervision** tasks
2. Leave-one-**category**-out



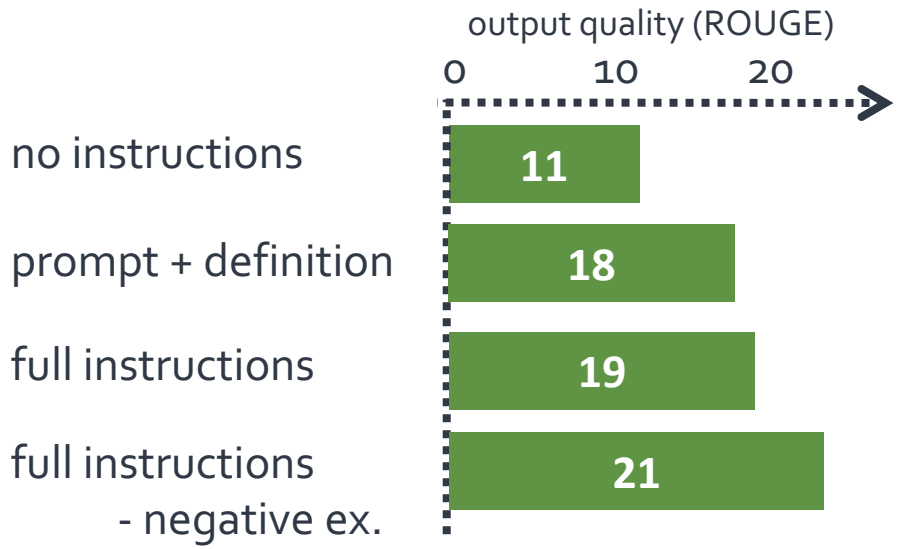
Exp: Generalization to a Random Split

- Can models learn to act w.r.p. instructions?
  - BART (base) [Lewis et al. 2019]
- Small models, too, generalize to **unseen** tasks! 🥳
- All instruction elements (except negative examples) help!

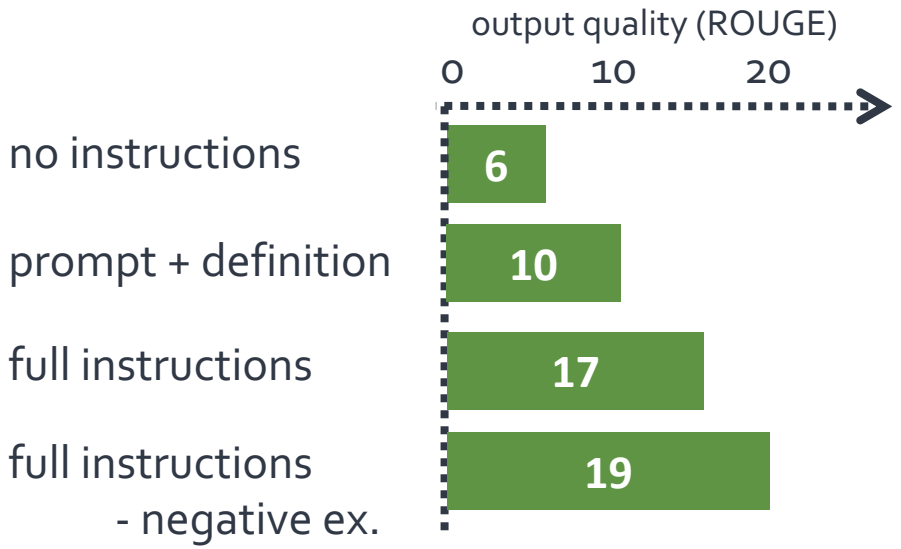


Exp: Generalization to Unseen Categories

- Evaluate on task of a particular category and train on the rest.



eval. on unseen "answer generation" tasks



eval on unseen "question generation" tasks

- Instructions improve generalization to tasks of **unseen** categories! ✨

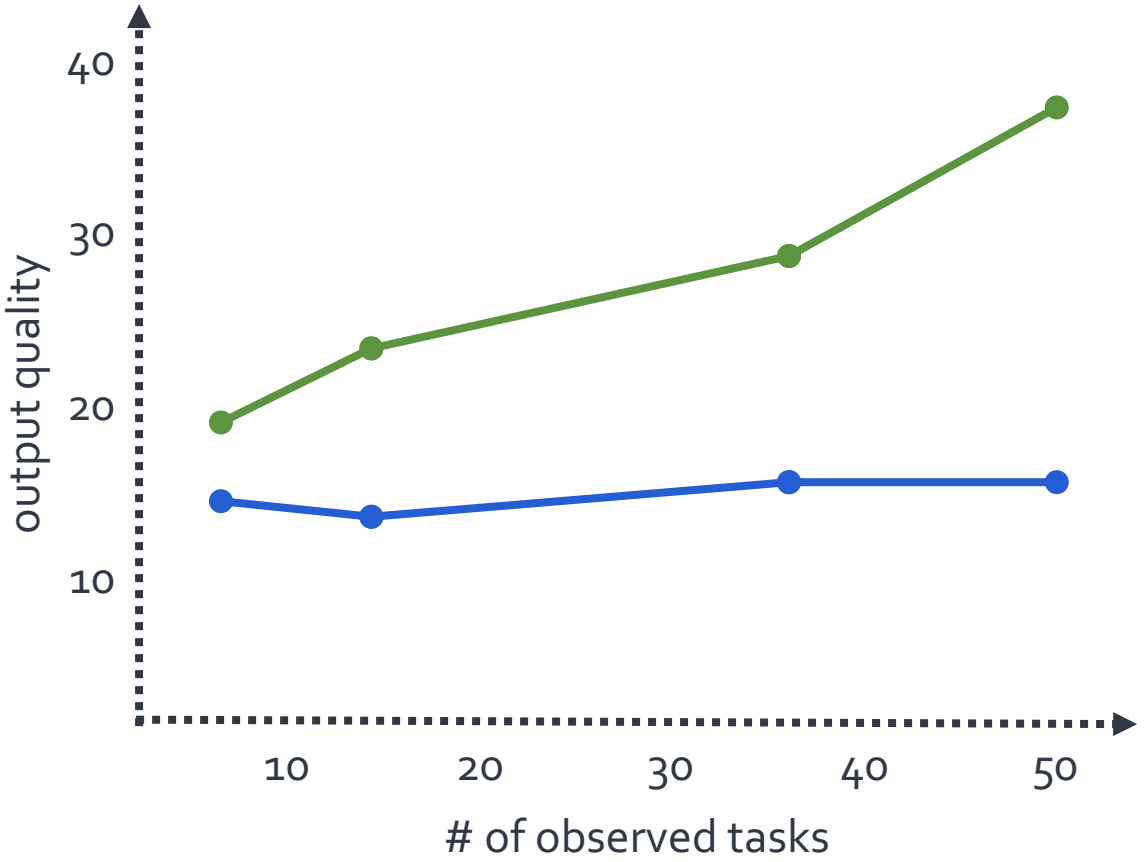
Exp: Generalization vs Size of Observed Tasks

- How the number of observed tasks affects cross-task generalization?

- Generalization to unseen tasks **improves** with **more** observed tasks! 🔥

Full Instructions

No Instructions



Lessons

- Motivating Hypothesis:
  - Can machines generalize to unseen tasks, via natural language instructions?
- *Natural-Instructions*: a dataset of many tasks and their crowdsourcing instructions/annotations.
- Empirical evidence that:
  - Instructions help w/ generalization to unseen tasks!
  - There is notable room to make progress!

http://instructions.apps.allenai.org

That’s it!