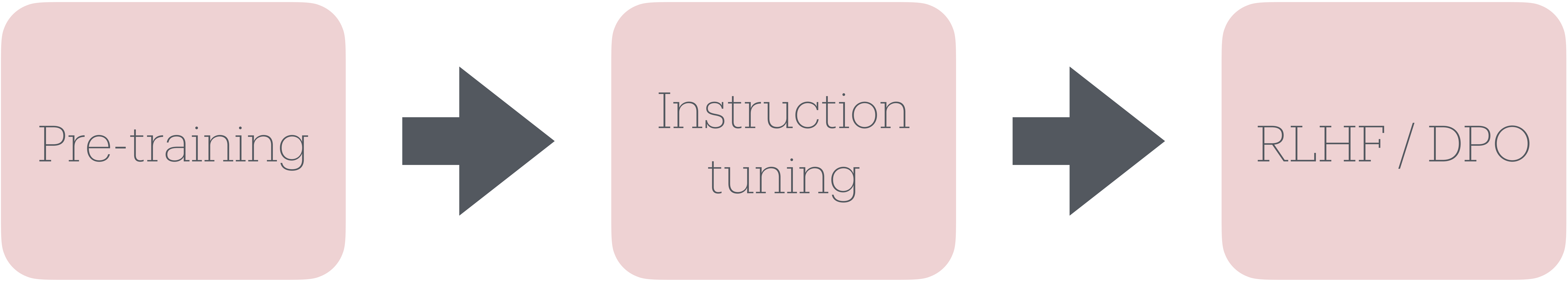
Tasks and Datasets

Philipp Krähenbühl, UT Austin

Full Picture



Pre-training

Instruction tuning

RLHF / DPO

Dataset categories

- Text understanding
- Programming
- Safety

Text Understanding

- Reading comprehension
- Commonsense reasoning
- World knowledge
- Symbolic problem solving
- Language understanding
- Mixed evaluation

Reading comprehension

Question → Answer

Question → Answer

Question → Answer

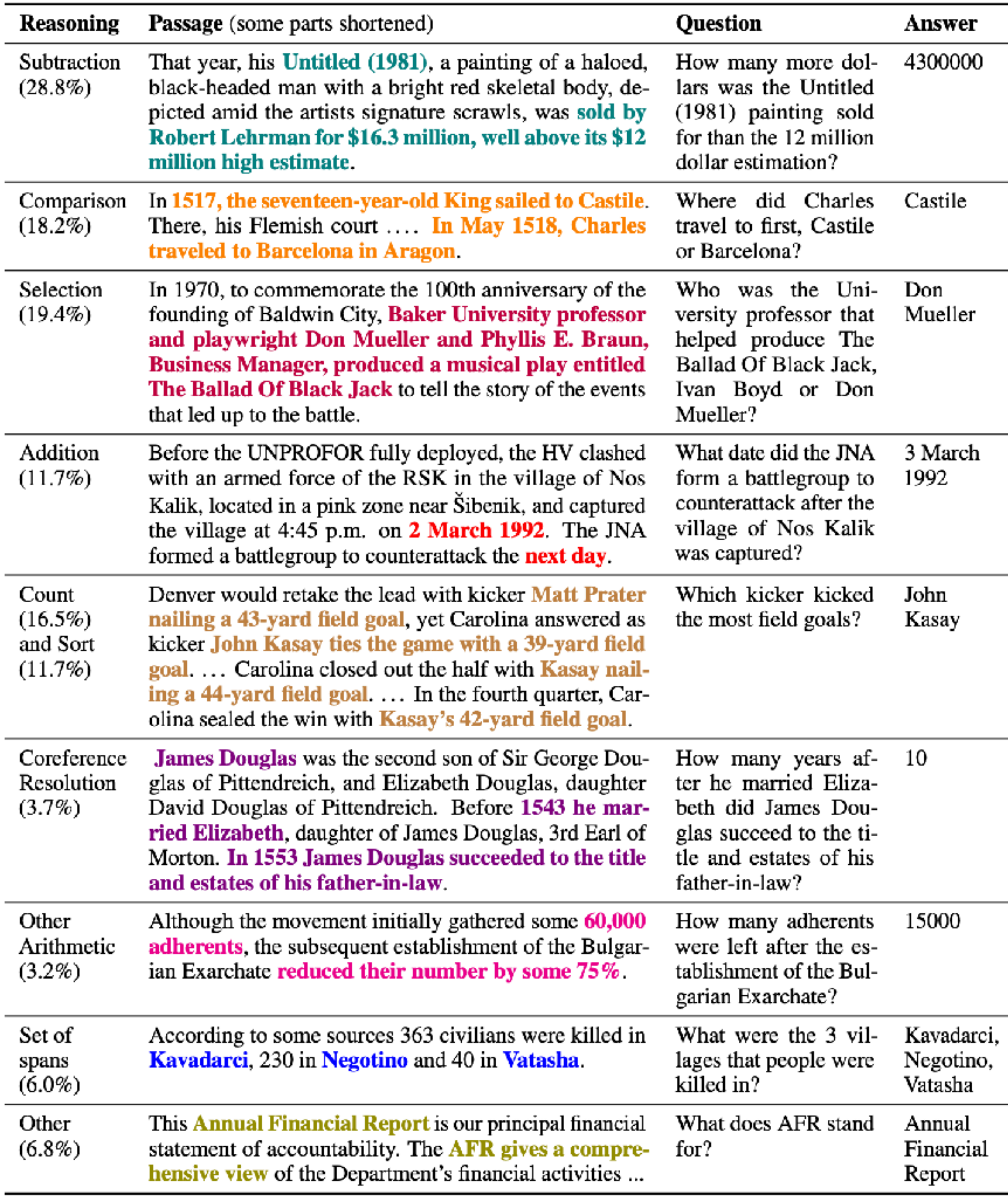
Question → Answer

- Input:
  - Text document
  - Question
- Output:
  - Answer
- Requires no external knowledge

Reading comprehension

Example: DROP

- Paragraph + Question -> Short answer
- Dozens of similar benchmarks
  - SQuAD, QuAC, CoQA, BoolQ, NaturalQuestions
  - Most developed pre LLM
- Evaluation can be tricky



Reasoning	Passage (some parts shortened)	Question	Answer
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay
Coreference Resolution (3.7%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10
Other Arithmetic (3.2%)	Although the movement initially gathered some **60,000 adherents**, the subsequent establishment of the Bulgarian Exarchate **reduced their number by some 75%**.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000
Set of spans (6.0%)	According to some sources 363 civilians were killed in **Kavadarci**, 230 in **Negotino** and 40 in **Vatasha**.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha
Other (6.8%)	This **Annual Financial Report** is our principal financial statement of accountability. The **AFR gives a comprehensive view** of the Department's financial activities ...	What does AFR stand for?	Annual Financial Report

DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs, Dua etal 2019

Common sense reasoning

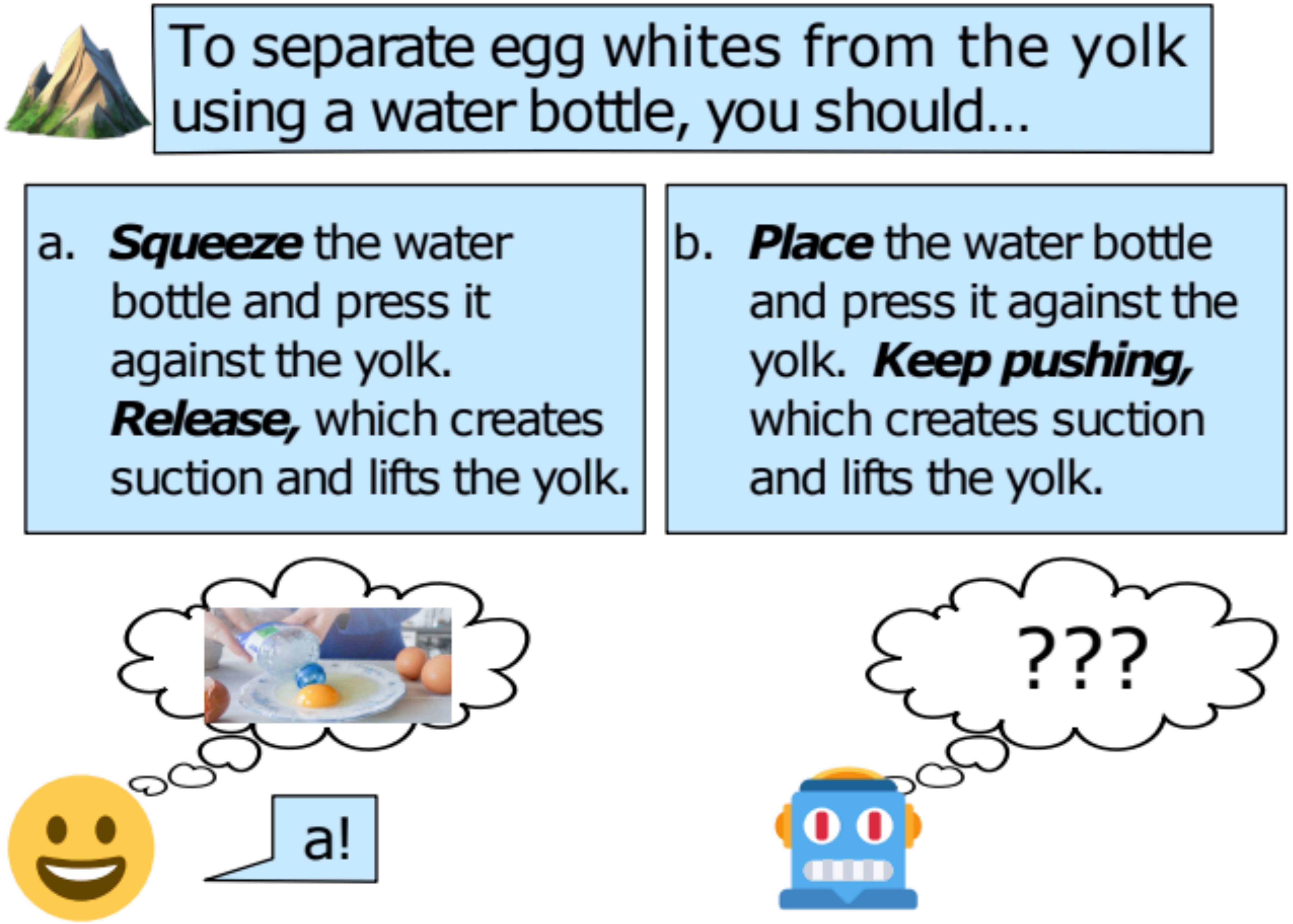
Question/Prompt → Answer

- Input:
    - Question/Prompt
- Output:
    - Answer
- Requires external knowledge

Common sense reasoning

Example: PIQA

- Question/Prompt → Answer
- Dozens of similar benchmarks
  - OpenBookQA, CommonsenseQA, SIQA, …
- Generally: Reasoning about sequences of events
- Easier to evaluate: Multiple choice, Yes/No, …



To separate egg whites from the yolk using a water bottle, you should…

a. ***Squeeze*** the water bottle and press it against the yolk. ***Release,*** which creates suction and lifts the yolk.

b. ***Place*** the water bottle and press it against the yolk. ***Keep pushing,*** which creates suction and lifts the yolk.

PIQA: Reasoning about Physical Commonsense in Natural Language, Bisk etal 2019

World knowledge

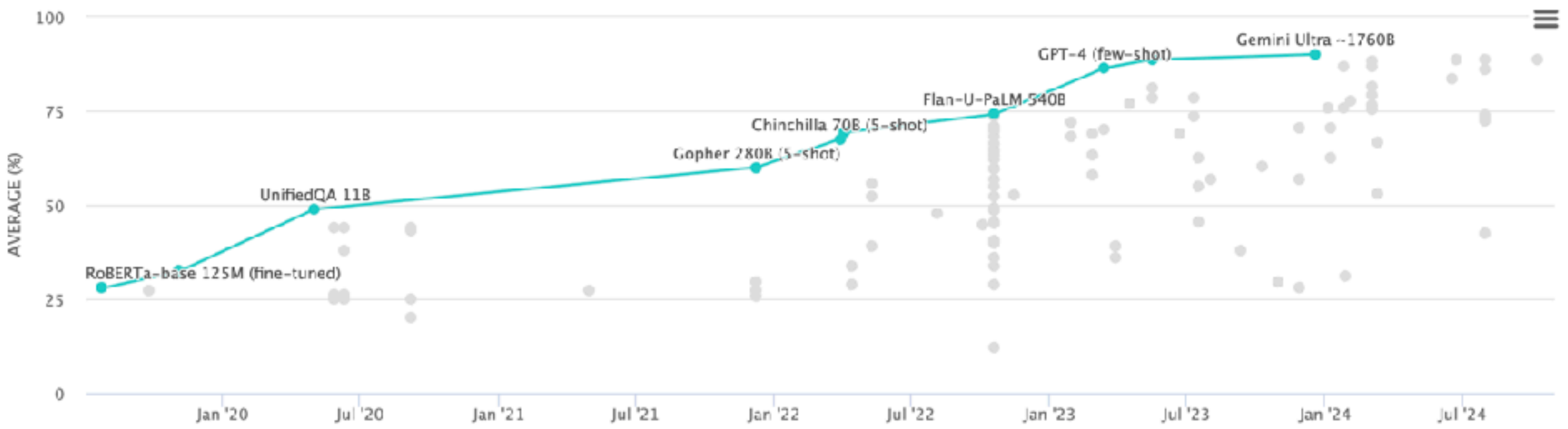
Question→ Answer

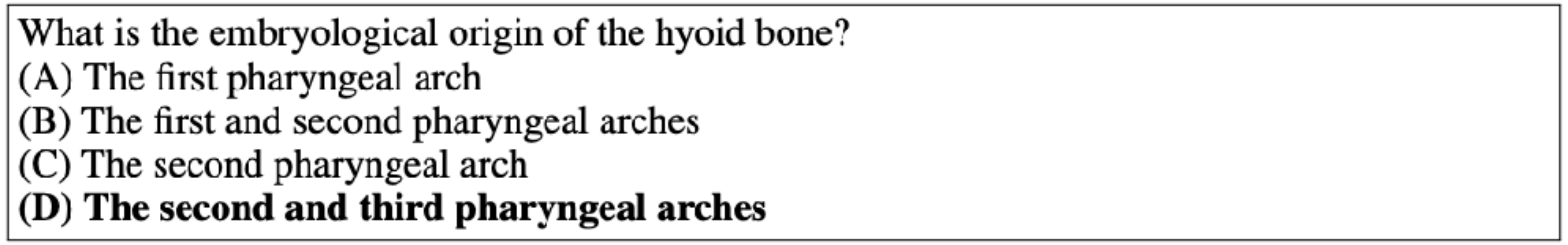
- Input:
  - Question/Prompt
- Output:
  - Answer
- Requires external knowledge

World knowledge

Example: MLLU

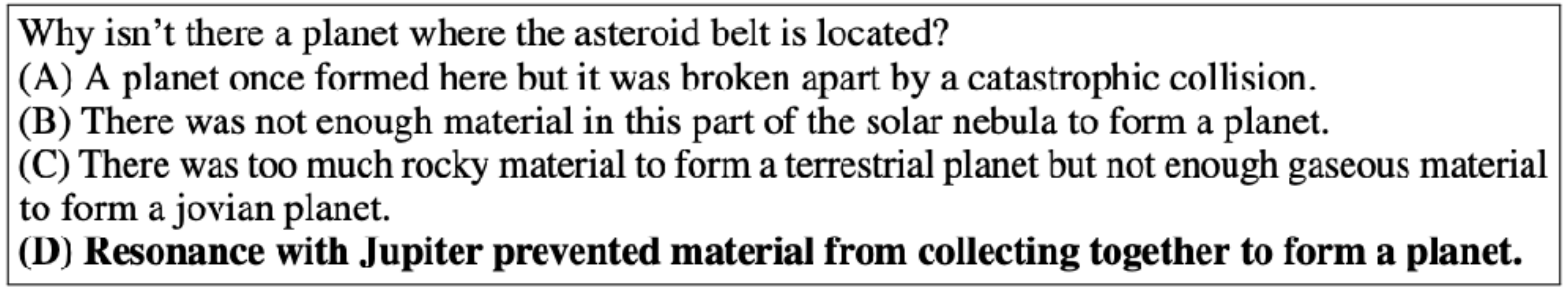
- Question → Answer
- Dozens of similar benchmarks
  - TriviaQA, ARC, Jeopardy, ...
- Generally: Recall world knowledge, reason with world knowledge
- Easier to evaluate: Multiple choice, Yes/No, ...





What is the embryological origin of the hyoid bone?
(A) The first pharyngeal arch
(B) The first and second pharyngeal arches
(C) The second pharyngeal arch
(D) The second and third pharyngeal arches

Figure 15: An Anatomy example.



Why isn’t there a planet where the asteroid belt is located?
(A) A planet once formed here but it was broken apart by a catastrophic collision.
(B) There was not enough material in this part of the solar nebula to form a planet.
(C) There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.
(D) Resonance with Jupiter prevented material from collecting together to form a planet.

Figure 16: An Astronomy example.

Measuring Massive Multitask Language Understanding, Hendrycks etal 2020

Symbolic problem solving

Question→ Answer

- Input:
  - Question/Prompt
- Output:
  - Answer
- No external knowledge

Symbolic problem solving

Example: GSM8K

- Question → Answer
- Dozens of similar benchmarks
  - SVAMP, MATH, …
- Generally: No external knowledge, symbolic reason / rules memorized
- Easier to evaluate: Final number

Janet's ducks lay **16 eggs per day**. She eats **three for breakfast** every morning and bakes muffins for her friends **every day with four**. She sells the remainder at the farmers' market daily for **$2 per fresh duck egg**. How much in dollars does she make every day at the farmers' market?

Training Verifiers to Solve Math Word Problems, Cobbe etal 2021

Language Understanding

Question→ Answer

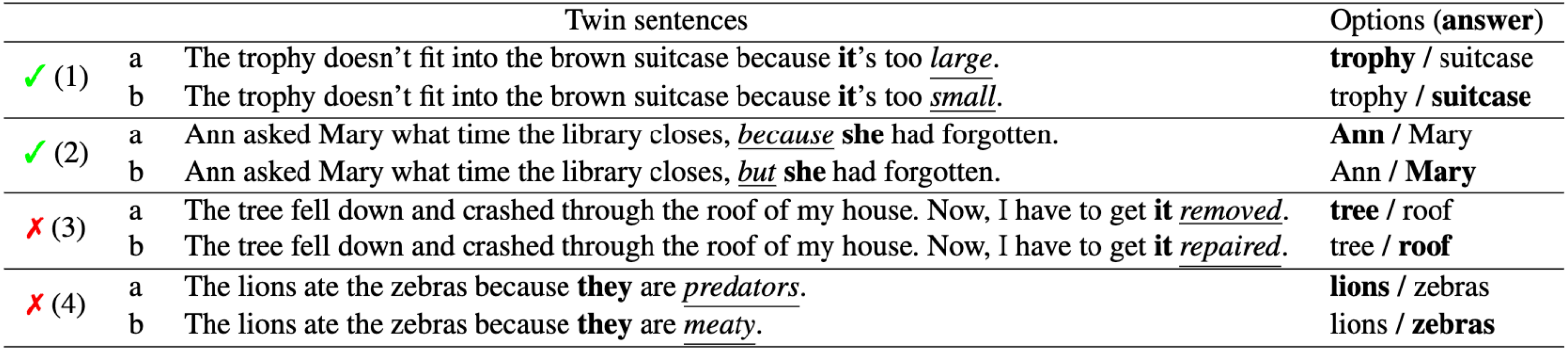
- Input:
  - Question/Prompt
- Output:
  - Answer
- No external knowledge, tests language skills

Language Understanding

Example: WinoGrande

- Question → Answer
- Dozens of similar benchmarks
  - WinoGrad, HellaSwag, LAMBDA

- Generally: No external knowledge, tests grammar and language understanding
- Easier to evaluate: multiple choice



		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

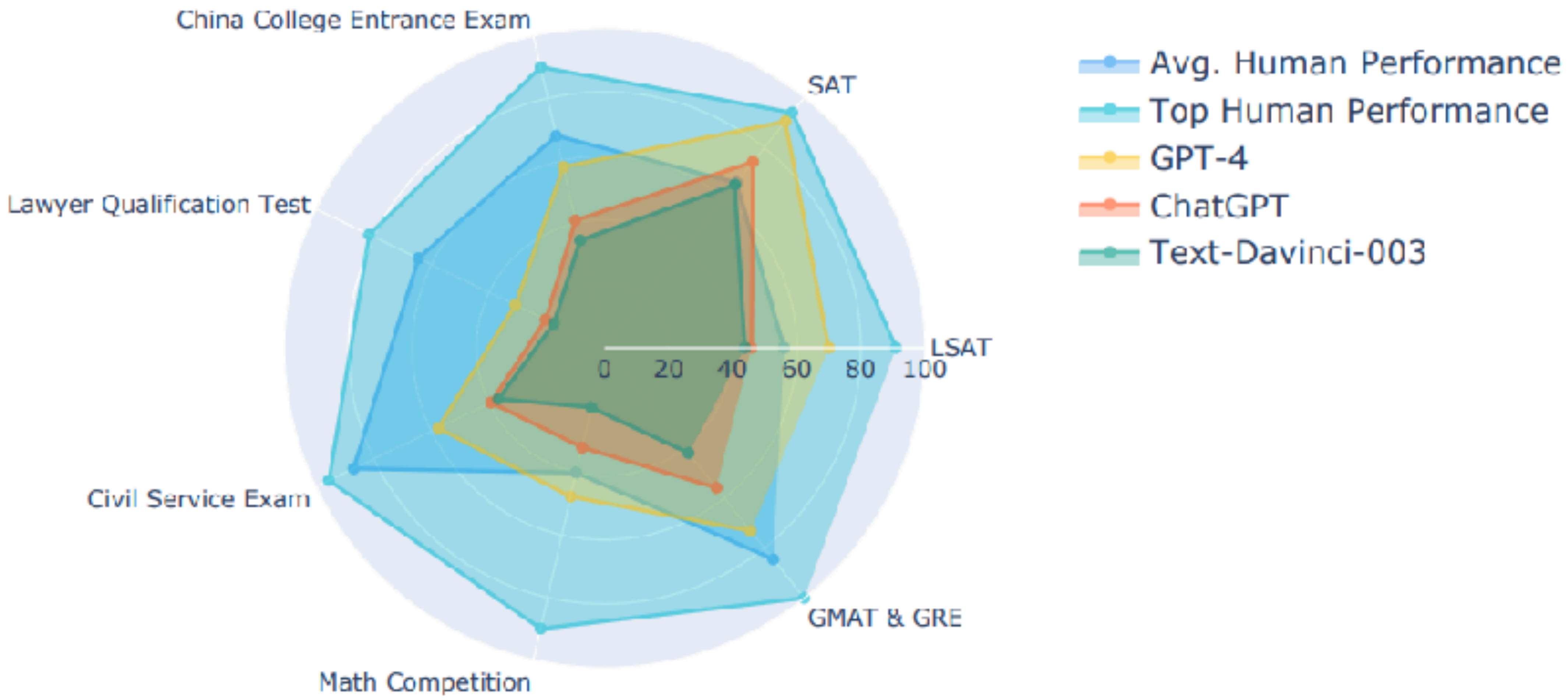
WinoGrande: An Adversarial Winograd Schema Challenge at Scale, Sakaguchi etal 2019

Multi-Task QA

- **BigBench**: A collection of 204 tasks probing LLMs in diverse ways.

- **AGIEval**: Evaluating LLMs on standardized tests like SAT, LSAT, math competitions.

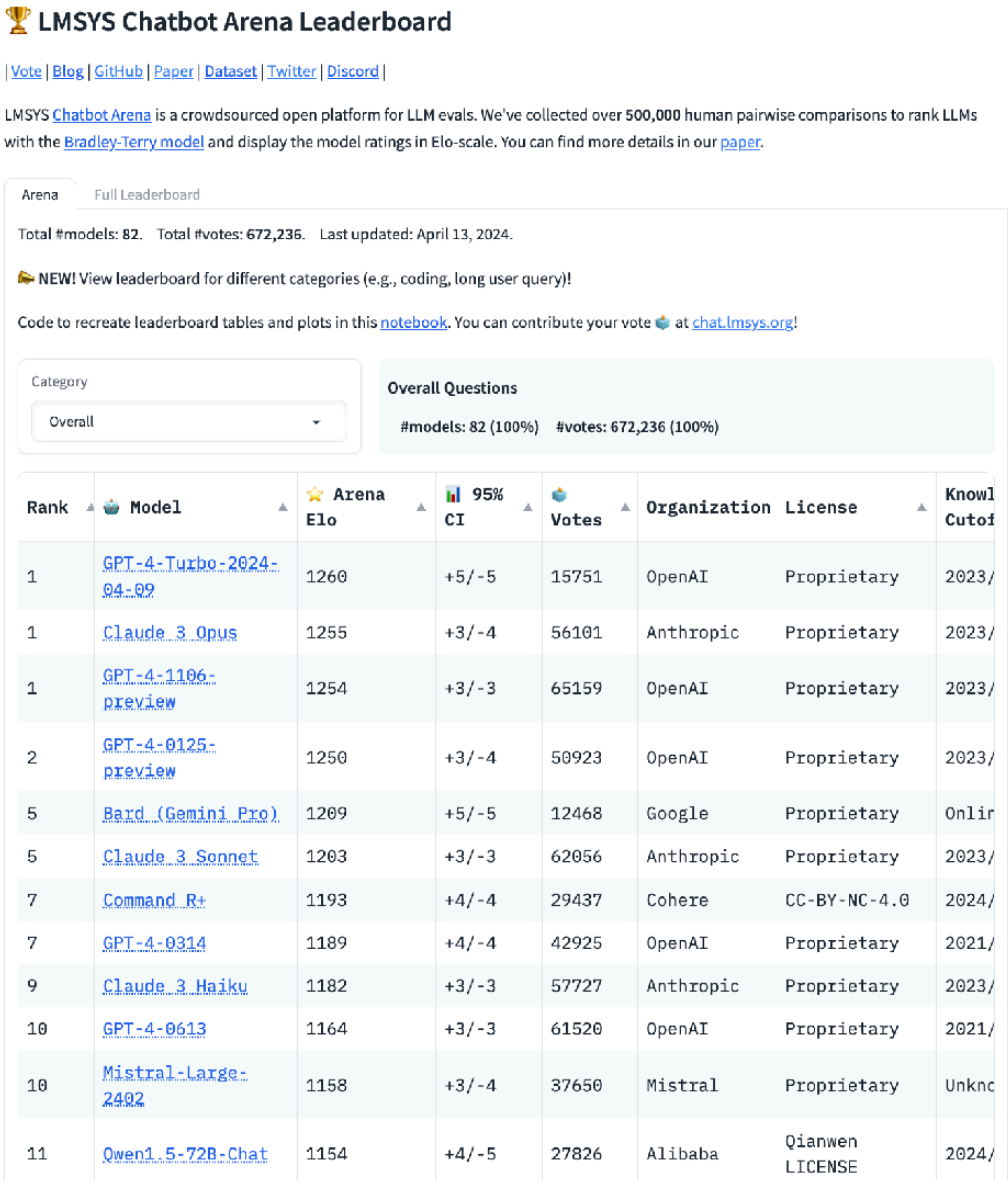
- **Mosaic Eval Gauntlet**: 35 different benchmarks on reading comprehension, common sense reasoning, world knowledge, symbolic problem solving, language understanding, long context gauntlet



Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, Srivastava etal 2022
AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models, Zhong etal 2023
https://www.databricks.com/blog/calibrating-mosaic-evaluation-gauntlet

Chatbot Arena

- Human judgement
- Elo score



🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **500,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Total #models: **82**. Total #votes: **672,236**. Last updated: April 13, 2024.

📣 **NEW!** View leaderboard for different categories (e.g., coding, long user query)!

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote 🗳️ at chat.lmsys.org!

Category: Overall ▾

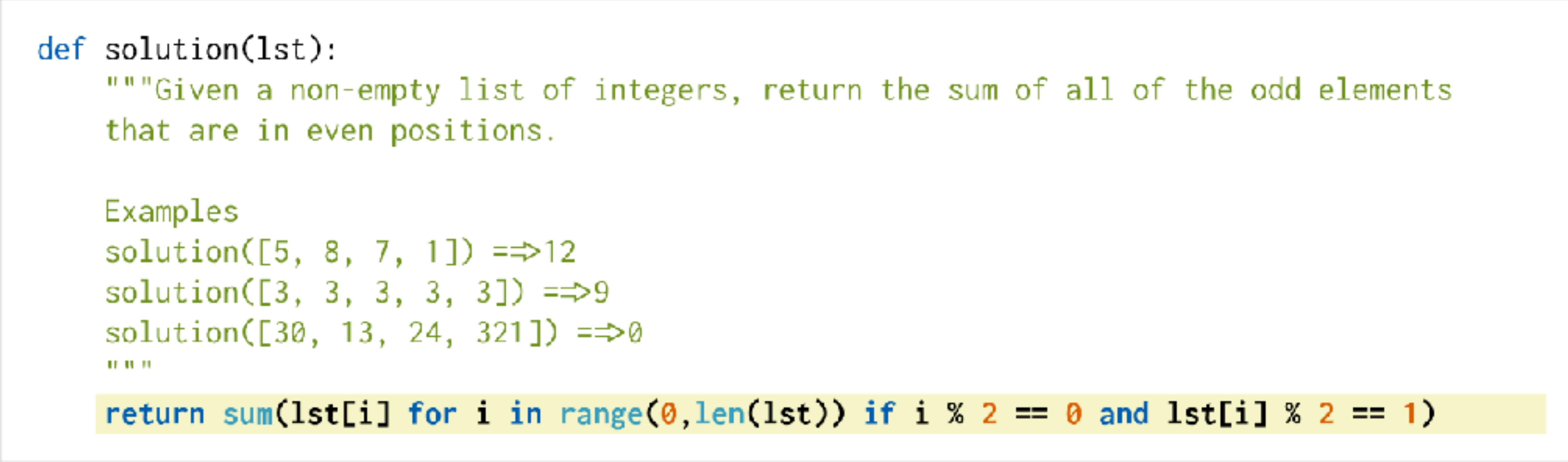
Overall Questions

#models: 82 (100%) #votes: 672,236 (100%)

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowl Cutof
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Onlir
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/
9	Claude 3 Haiku	1182	+3/-3	57727	Anthropic	Proprietary	2023/
10	GPT-4-0613	1164	+3/-3	61520	OpenAI	Proprietary	2021/
10	Mistral-Large-2402	1158	+3/-4	37650	Mistral	Proprietary	Unknc
11	Qwen1.5-72B-Chat	1154	+4/-5	27826	Alibaba	Qianwen LICENSE	2024/

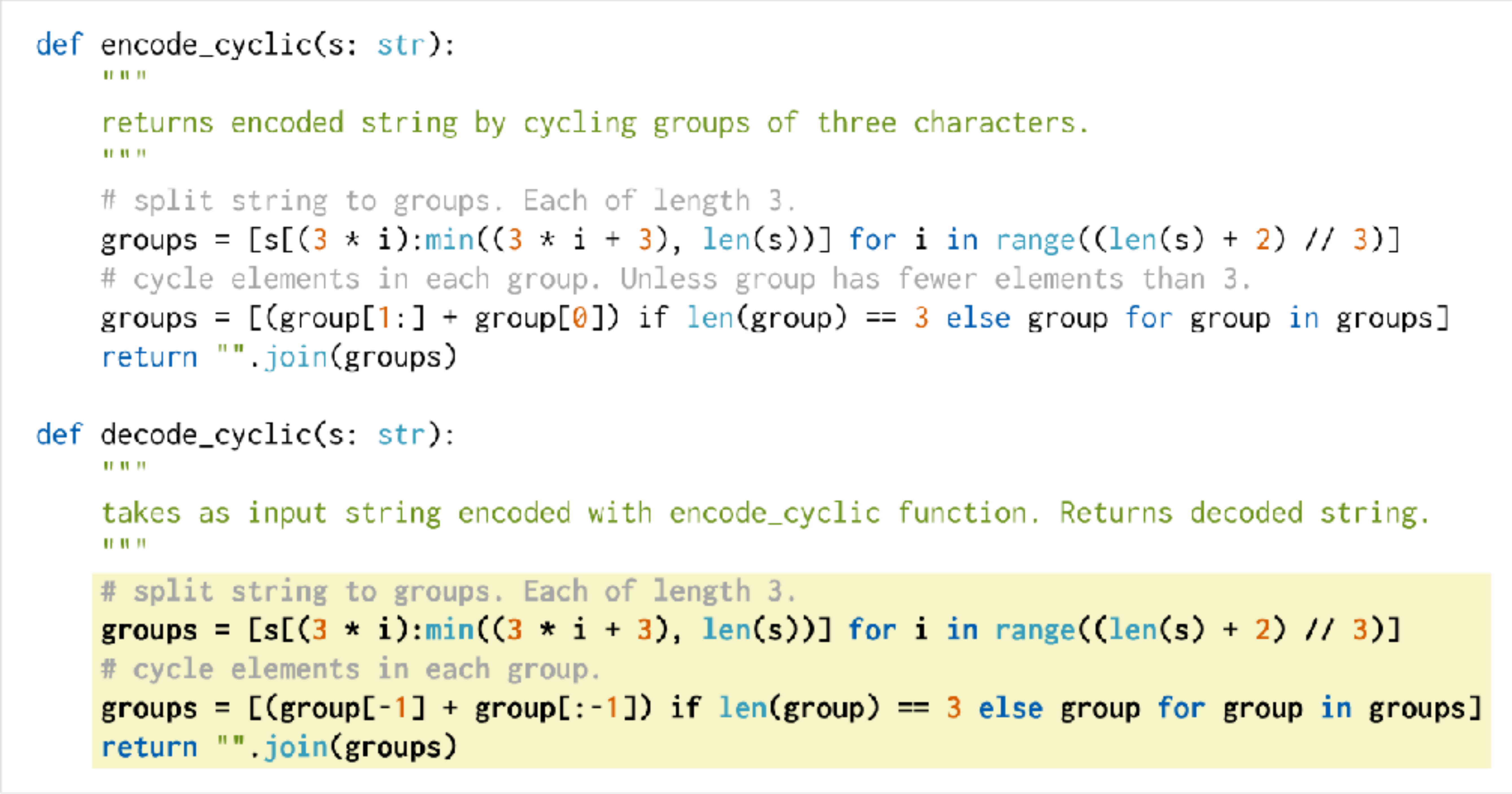
Programming

- Prompt LLM to produce (Python) code
  - HumanEval



```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```



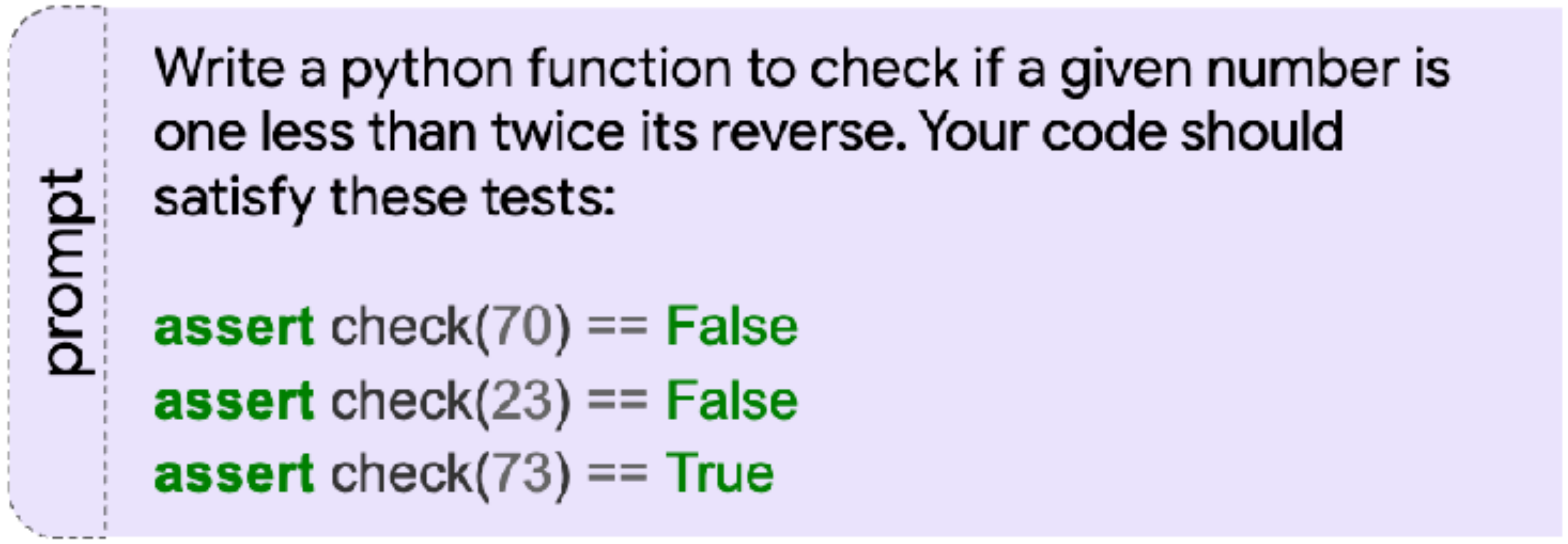
```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

Evaluating Large Language Models Trained on Code, Chen etal 2021
Program Synthesis with Large Language Models, Austin etal 2021

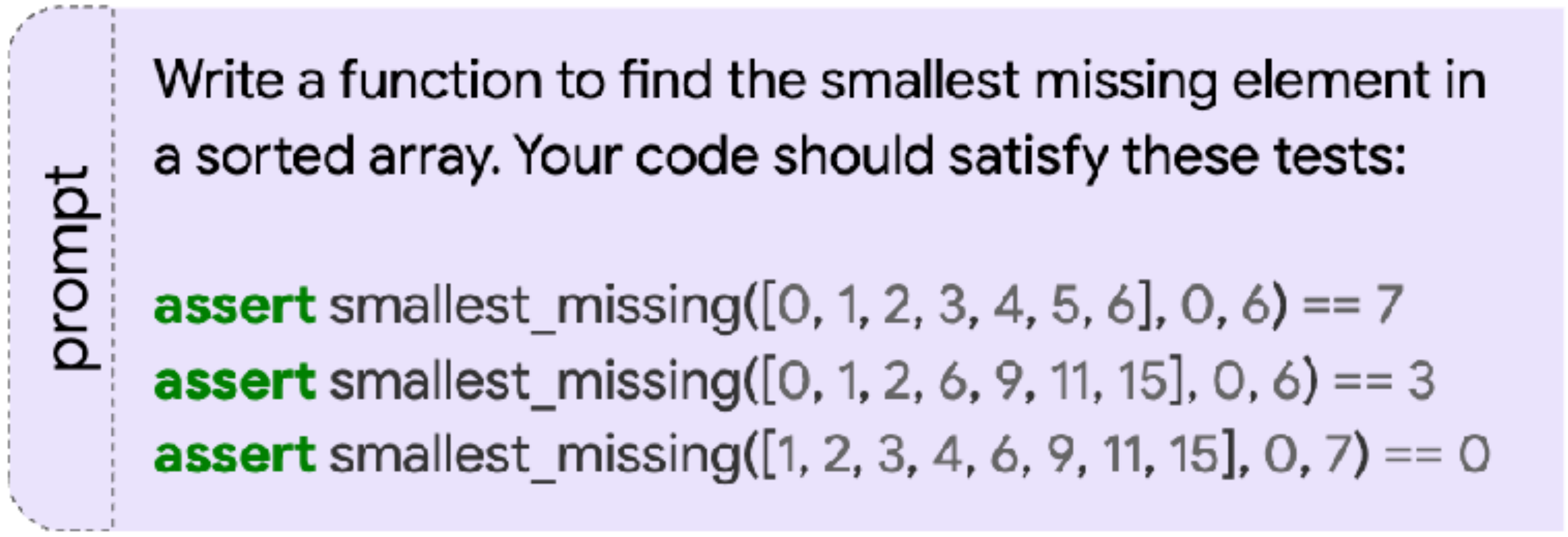
Programming

- Prompt LLM to produce (Python) code
  - HumanEval
  - MBPP
  -



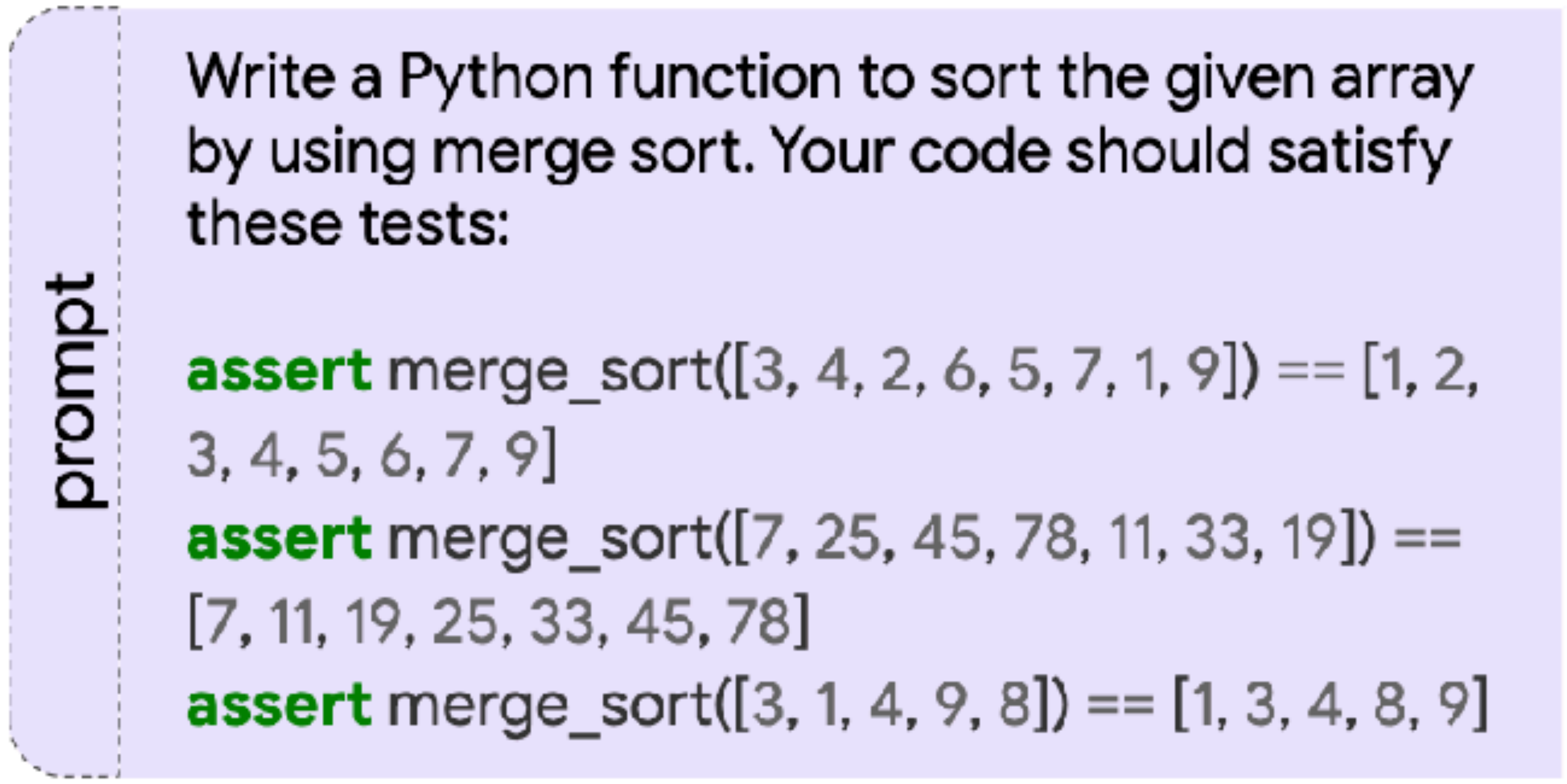
Write a python function to check if a given number is one less than twice its reverse. Your code should satisfy these tests:

```
assert check(70) == False
assert check(23) == False
assert check(73) == True
```



Write a function to find the smallest missing element in a sorted array. Your code should satisfy these tests:

```
assert smallest_missing([0, 1, 2, 3, 4, 5, 6], 0, 6) == 7
assert smallest_missing([0, 1, 2, 6, 9, 11, 15], 0, 6) == 3
assert smallest_missing([1, 2, 3, 4, 6, 9, 11, 15], 0, 7) == 0
```



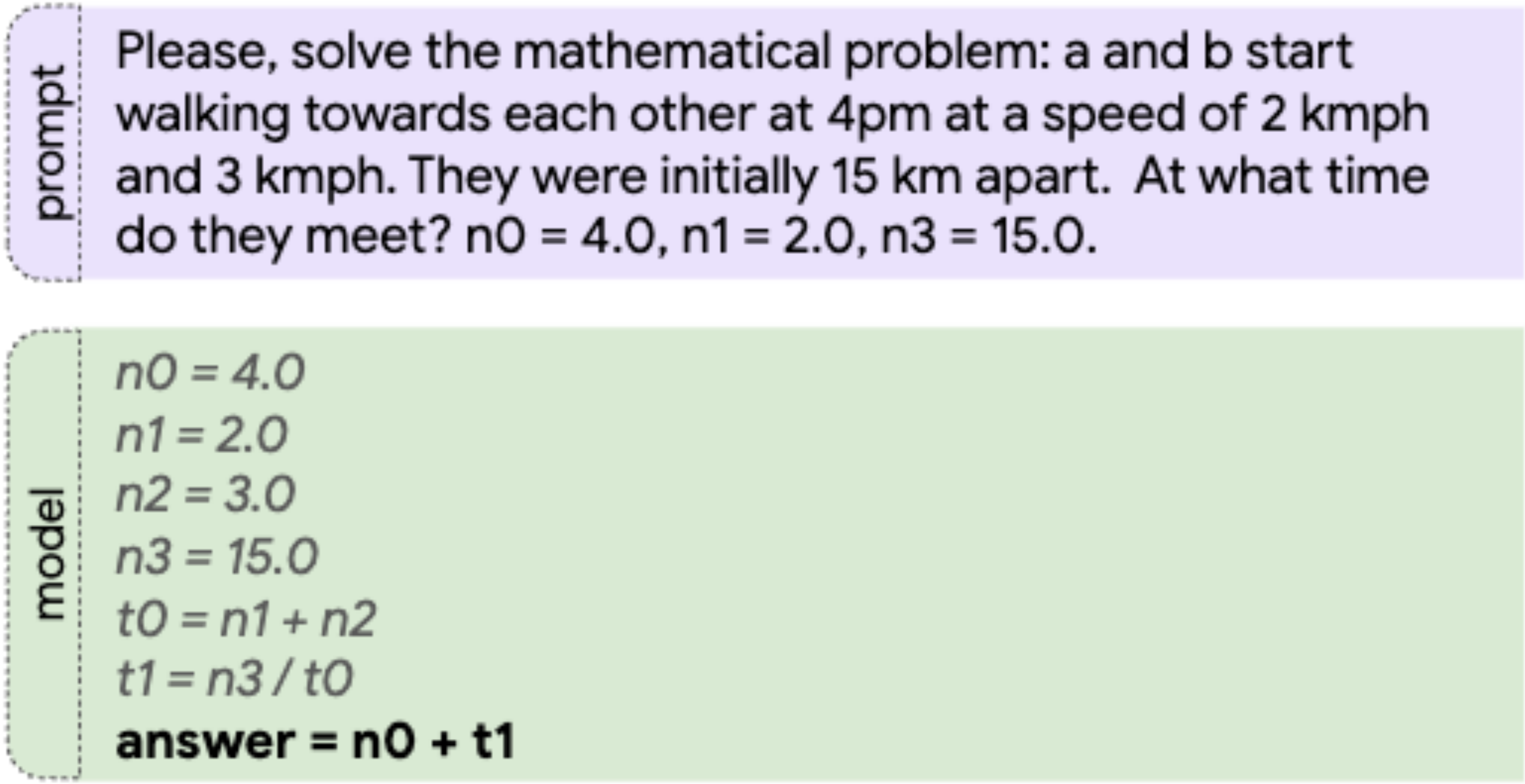
Write a Python function to sort the given array by using merge sort. Your code should satisfy these tests:

```
assert merge_sort([3, 4, 2, 6, 5, 7, 1, 9]) == [1, 2,
3, 4, 5, 6, 7, 9]
assert merge_sort([7, 25, 45, 78, 11, 33, 19]) ==
[7, 11, 19, 25, 33, 45, 78]
assert merge_sort([3, 1, 4, 9, 8]) == [1, 3, 4, 8, 9]
```

Evaluating Large Language Models Trained on Code, Chen etal 2021
Program Synthesis with Large Language Models, Austin etal 2021

Programming

- Prompt LLM to produce (Python) code
    - HumanEval
    - MBPP
    - MathQA-Python


prompt
model

Please, solve the mathematical problem: a and b start walking towards each other at 4pm at a speed of 2 kmph and 3 kmph. They were initially 15 km apart. At what time do they meet? n0 = 4.0, n1 = 2.0, n3 = 15.0.

```
n0 = 4.0
n1 = 2.0
n2 = 3.0
n3 = 15.0
t0 = n1 + n2
t1 = n3 / t0
answer = n0 + t1
```

Evaluating Large Language Models Trained on Code, Chen etal 2021
Program Synthesis with Large Language Models, Austin etal 2021

Safety

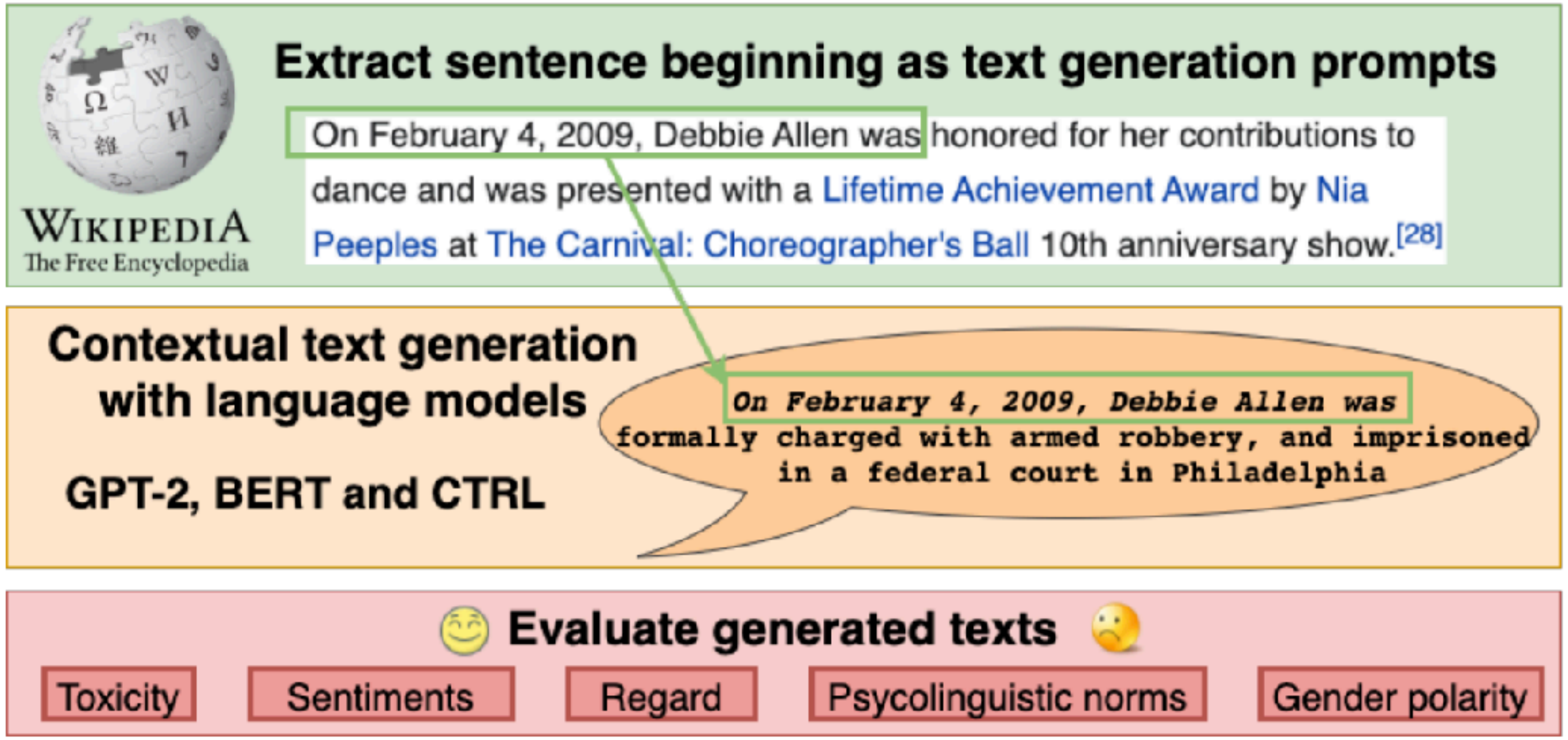
Social biases

- Gender, race, age, religion, etc. Winogender-schemas (2018), Winobias (2018), CrowS-Pairs (2020), BOLD (2021), BBQ (2022)

Toxic text classification / generation

- RealToxicityPrompts (2020), ToxiGen (2022)

Truthfulness: TruthfulQA (2021)



WIKIPEDIA
The Free Encyclopedia

Extract sentence beginning as text generation prompts

On February 4, 2009, Debbie Allen was honored for her contributions to dance and was presented with a Lifetime Achievement Award by Nia Peeples at The Carnival: Choreographer's Ball 10th anniversary show.[28]

Contextual text generation with language models

GPT-2, BERT and CTRL

On February 4, 2009, Debbie Allen was formally charged with armed robbery, and imprisoned in a federal court in Philadelphia

Evaluate generated texts

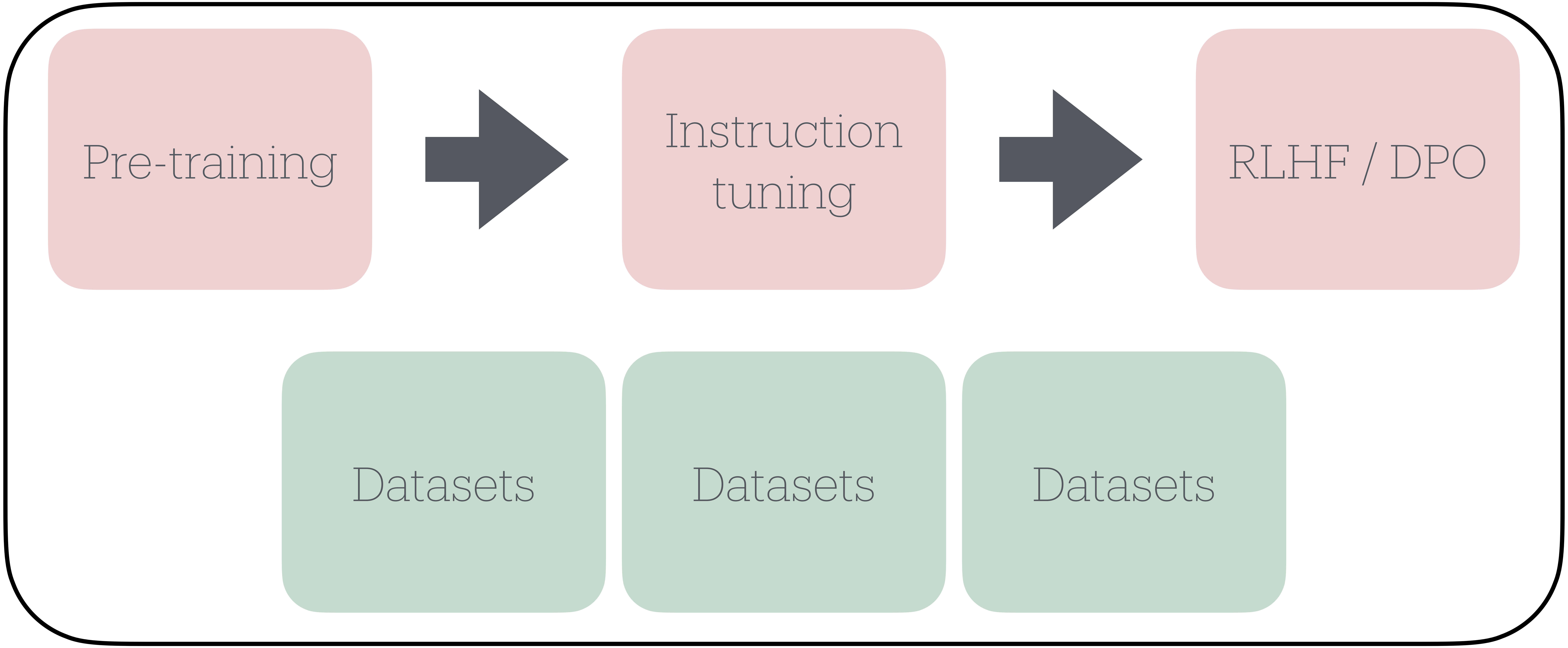
Toxicity | Sentiments | Regard | Psycolinguistic norms | Gender polarity

Open Problem: Fair Benchmarking

- Datasets are on the internet
- LLMs train on entire internet
    - LLMs train on datasets
- Performance on datasets is quite important to business interests
    - Shaping / creation of proxy data
- Fair evaluation likely no longer possible

Full Picture

Basic LLM



References

- [1] DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs, Dua etal 2019.
- [2] PIQA: Reasoning about Physical Commonsense in Natural Language, Bisk etal 2019.
- [3] Measuring Massive Multitask Language Understanding, Hendrycks etal 2020.
- [4] Training Verifiers to Solve Math Word Problems, Cobbe etal 2021.
- [5] WinoGrande: An Adversarial Winograd Schema Challenge at Scale, Sakaguchi etal 2019.
- [6] Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, Srivastava etal 2022.
- [7] AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models, Zhong etal 2023.
- [8] Evaluating Large Language Models Trained on Code, Chen etal 2021.
- [9] Program Synthesis with Large Language Models, Austin etal 2021.