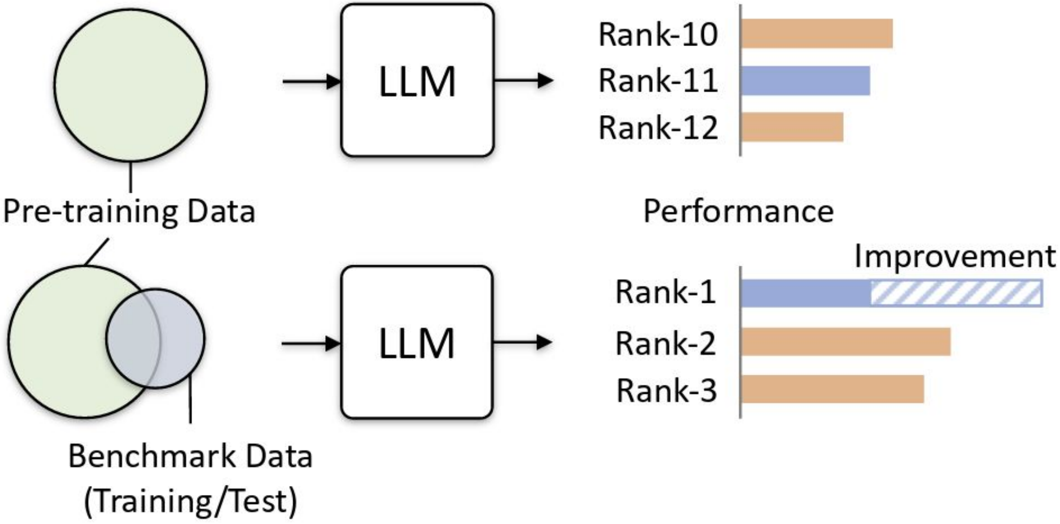
Pitfalls in measuring generalization

Marianna Nezhurina
Juelich Supercomputing Centre, LAION

Data leakage: Training data and test data overlap

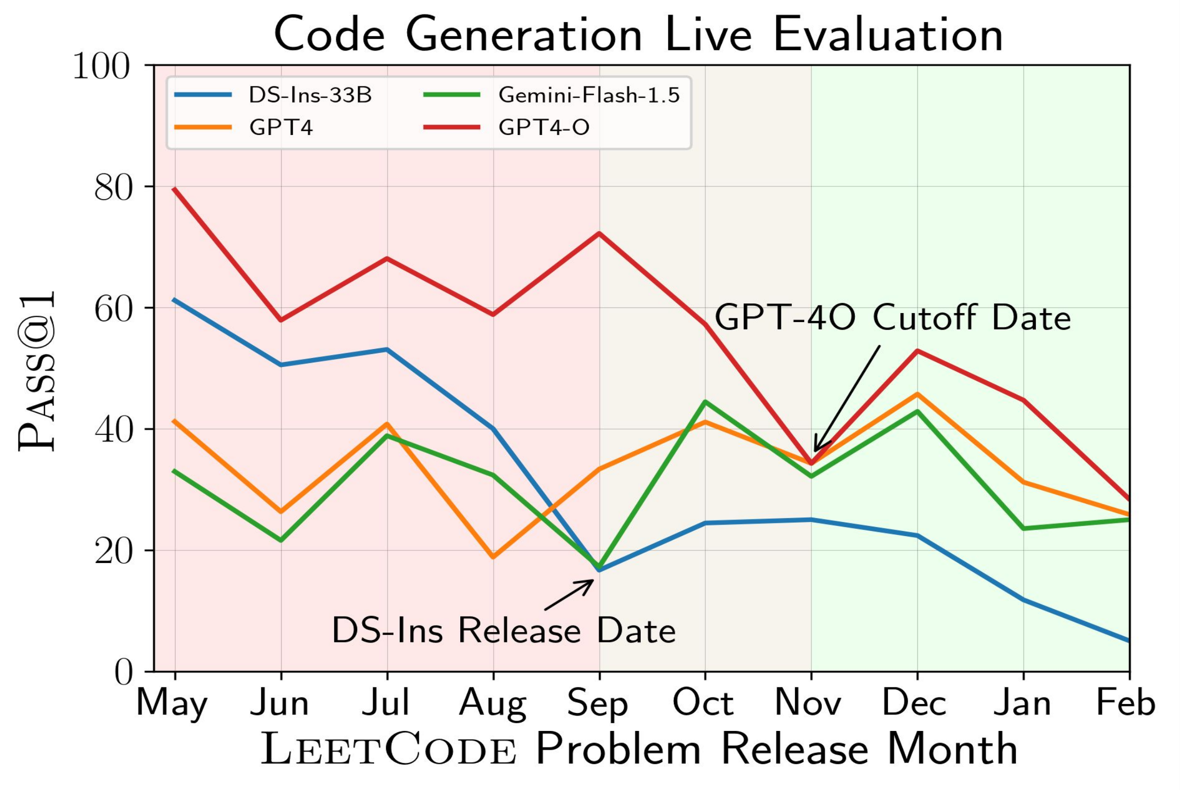
- Pre-training **data is large** and **prepared ahead of tests**
- **High overlap between train and test** data can dramatically boost performance of LLMs on a particular benchmark
- Data **leakage can skew the assessment** of relative model capabilities



Zhou, Kun, et al. "Don't make your llm an evaluation benchmark cheater." arXiv preprint arXiv:2311.01964 (2023).

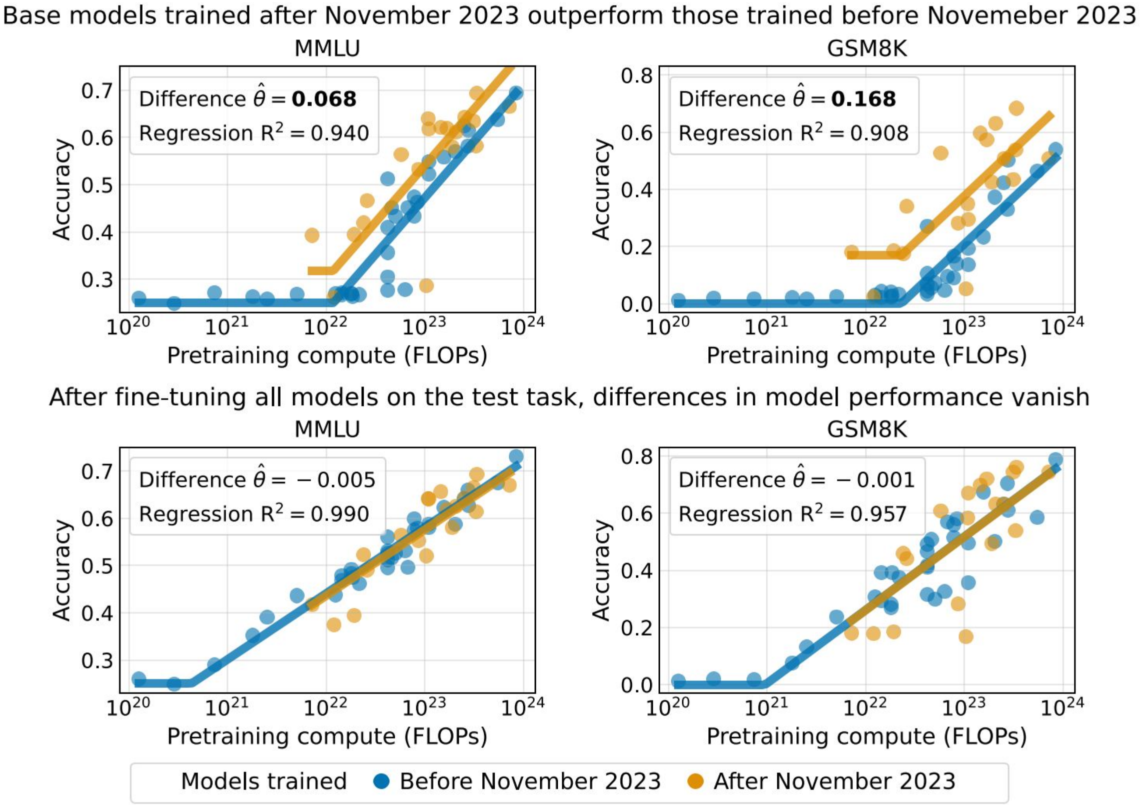
Contamination-free benchmarks

New benchmarks paradigms such as MixEval, LiveCodeBench (on the right), LiveBench try to solve this problem by constantly updating their problem sets with newly available problems.



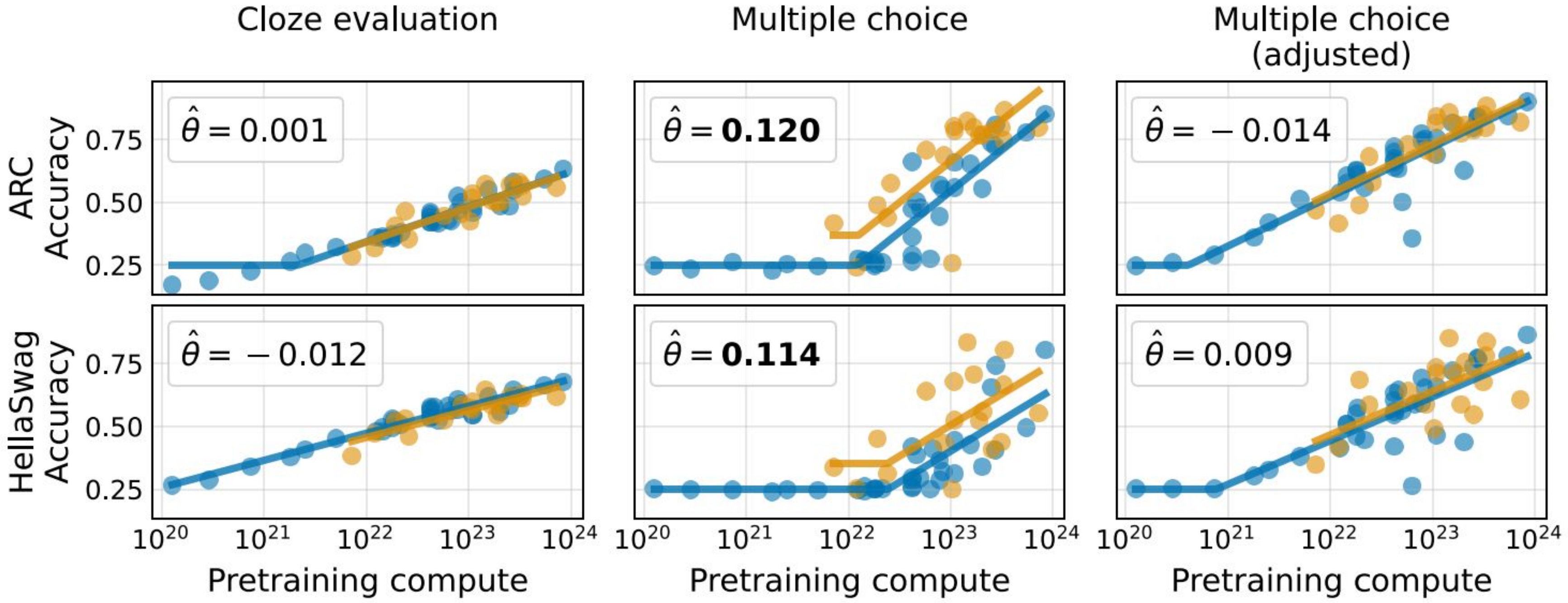
Jain, Naman, et al. "Livecodebench: Holistic and contamination free evaluation of large language models for code." arXiv preprint arXiv:2403.07974 (2024).

Increased training data size makes it harder to expect that a test set will not be in the training distribution.

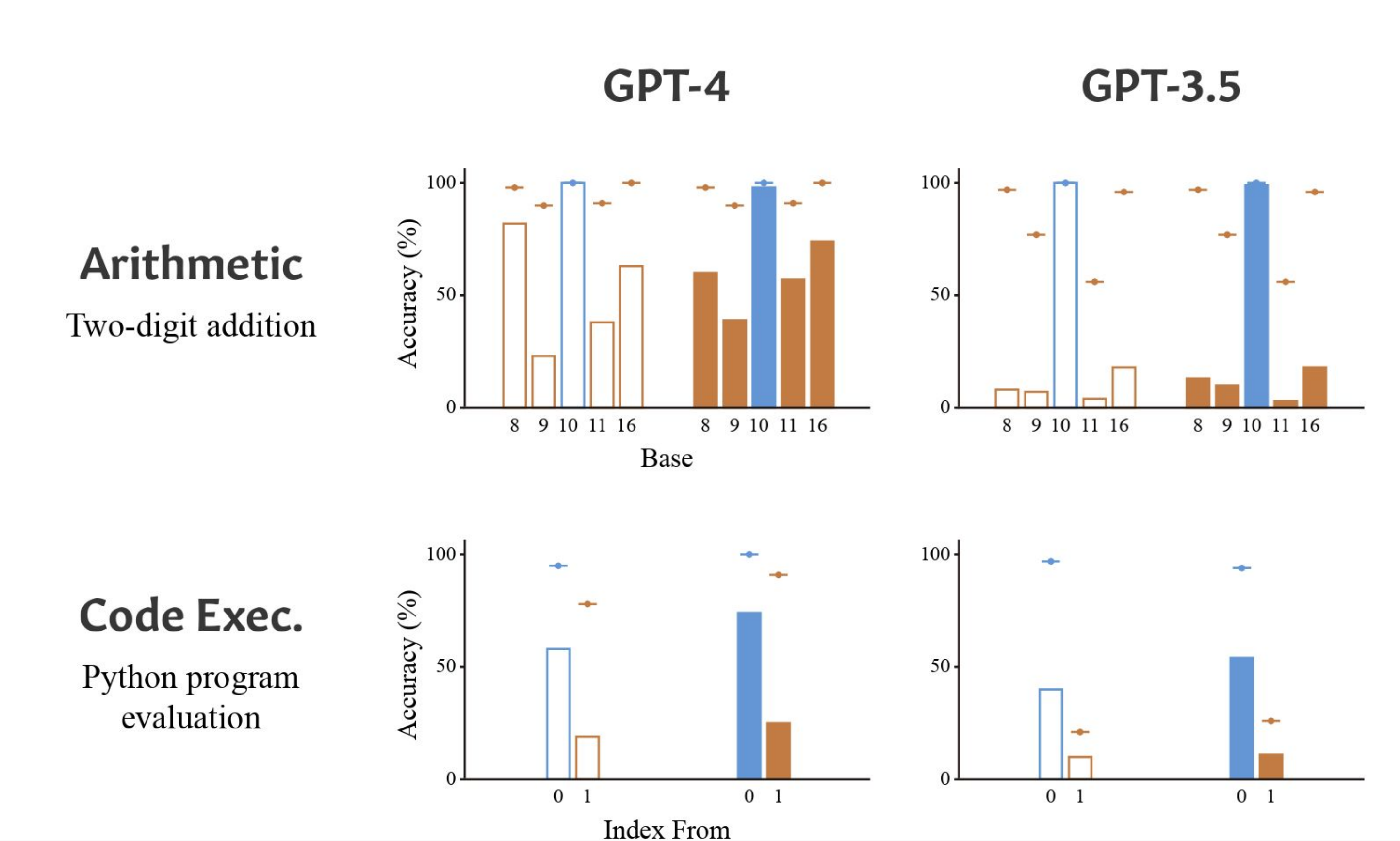


Dominguez-Olmedo, Ricardo, Florian E. Dorner, and Moritz Hardt. "Training on the test task confounds evaluation and emergence." arXiv preprint arXiv:2407.07890 (2024).

Newer models don’t have better scores after reformulation of questions (they are just more familiar with original question format)

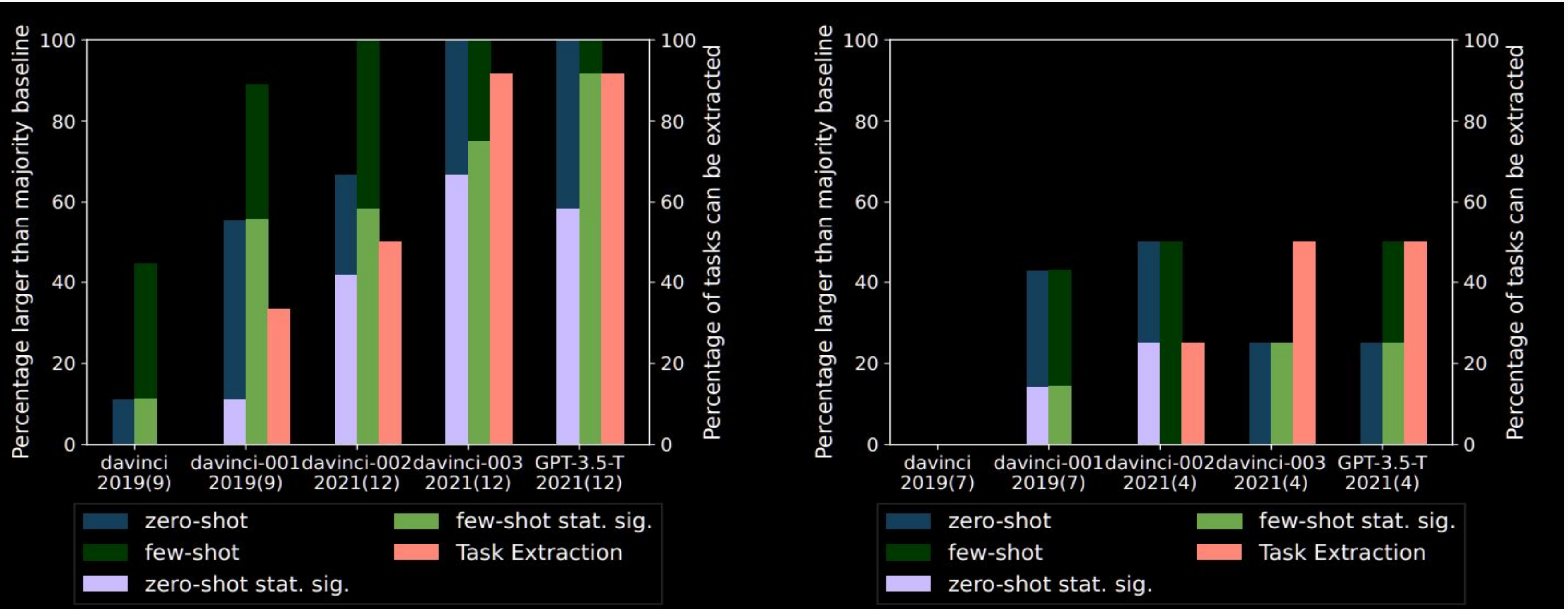


Counterfactual Tasks: The blue and orange bars represent the default and counterfactual conditions respectively, either with or without 0-shot chain-of-thought (0-CoT).



Wu, Zhaofeng, et al. "Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks." arXiv preprint arXiv:2307.02477 (2023).

Evaluation on new vs old benchmarks



(a) GPT-3 series on pre-collection datasets

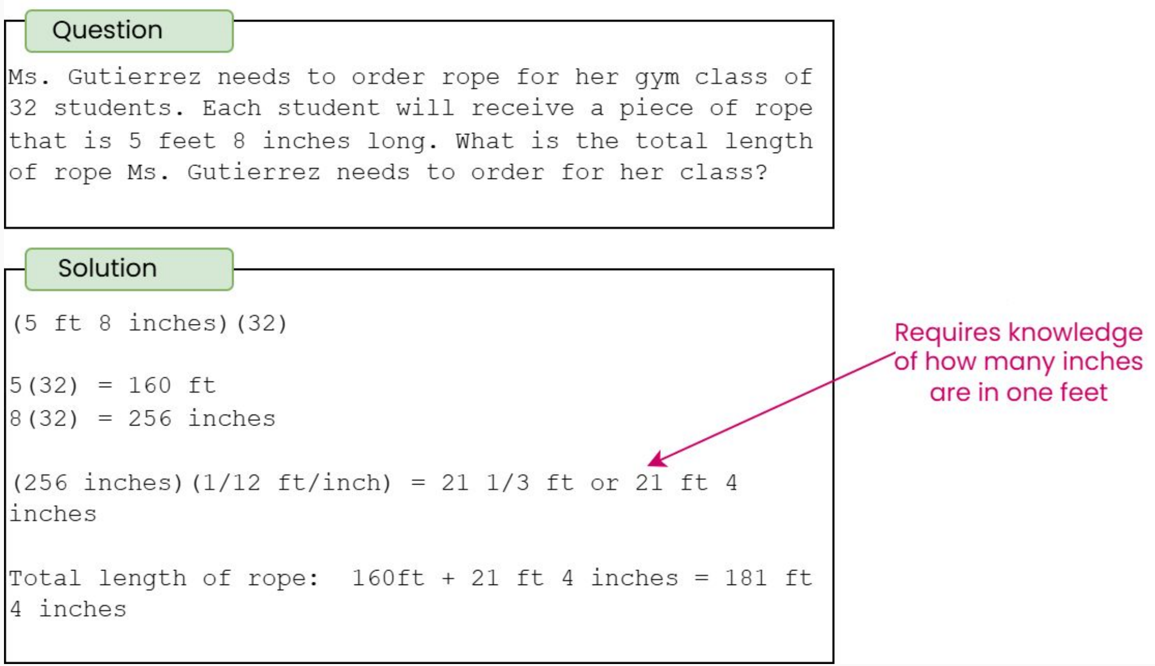
(b) GPT-3 series on post-collection datasets

Li, C., & Flanigan, J. (2024). Task Contamination: Language Models May Not Be Few-Shot Anymore. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 18471-18480.

LLM Benchmarks like MMLU have questions that might require specific knowledge for the answer

Consider following question from the simplest MMLU subset elementary mathematics)

If you don't know how many inches are in one feet you will not be able to solve it!



Question

Ms. Gutierrez needs to order rope for her gym class of 32 students. Each student will receive a piece of rope that is 5 feet 8 inches long. What is the total length of rope Ms. Gutierrez needs to order for her class?

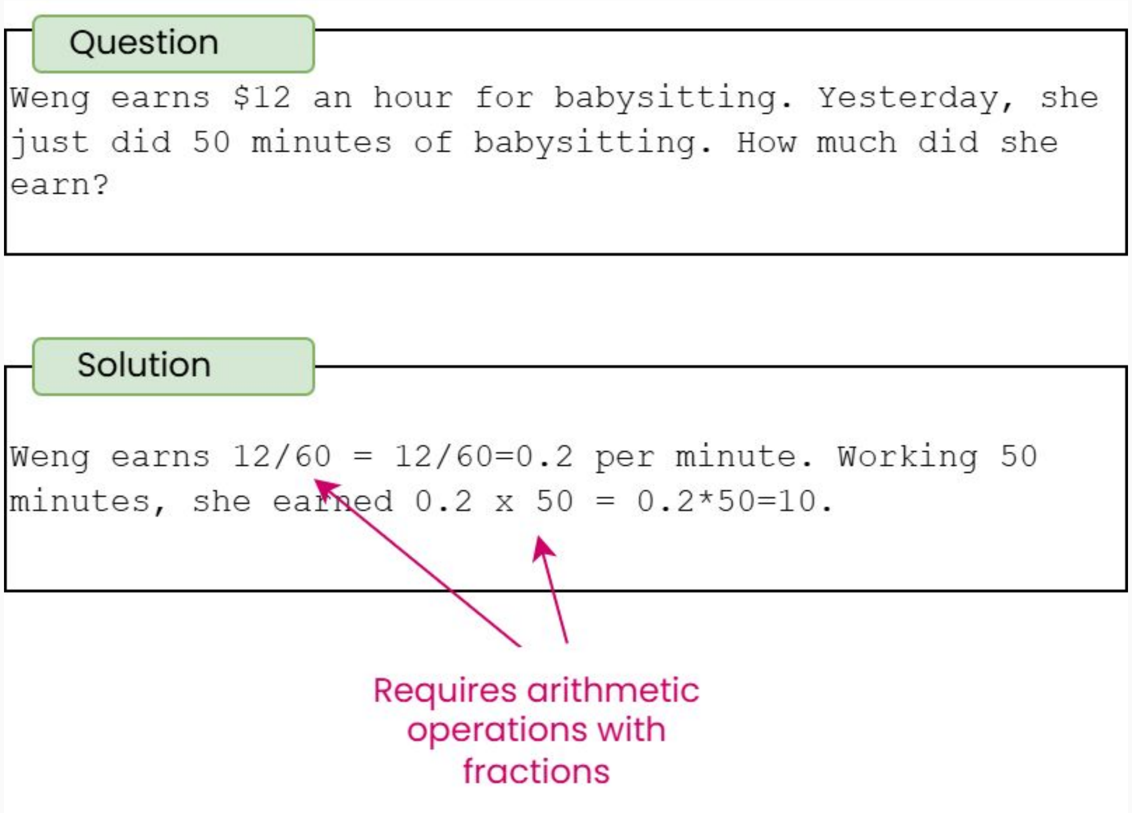
Solution

```
(5 ft 8 inches)(32)

5(32) = 160 ft
8(32) = 256 inches

(256 inches)(1/12 ft/inch) = 21 1/3 ft or 21 ft 4
inches

Total length of rope:  160ft + 21 ft 4 inches = 181 ft
4 inches
```

Requires knowledge of how many inches are in one feet

Typical GSM8k question

In the GSM8k benchmark, question usually require knowledge of basic arithmetics with both integers and fractions. These operations can be tricky even for humans (for kids for example) or people who are distracted.



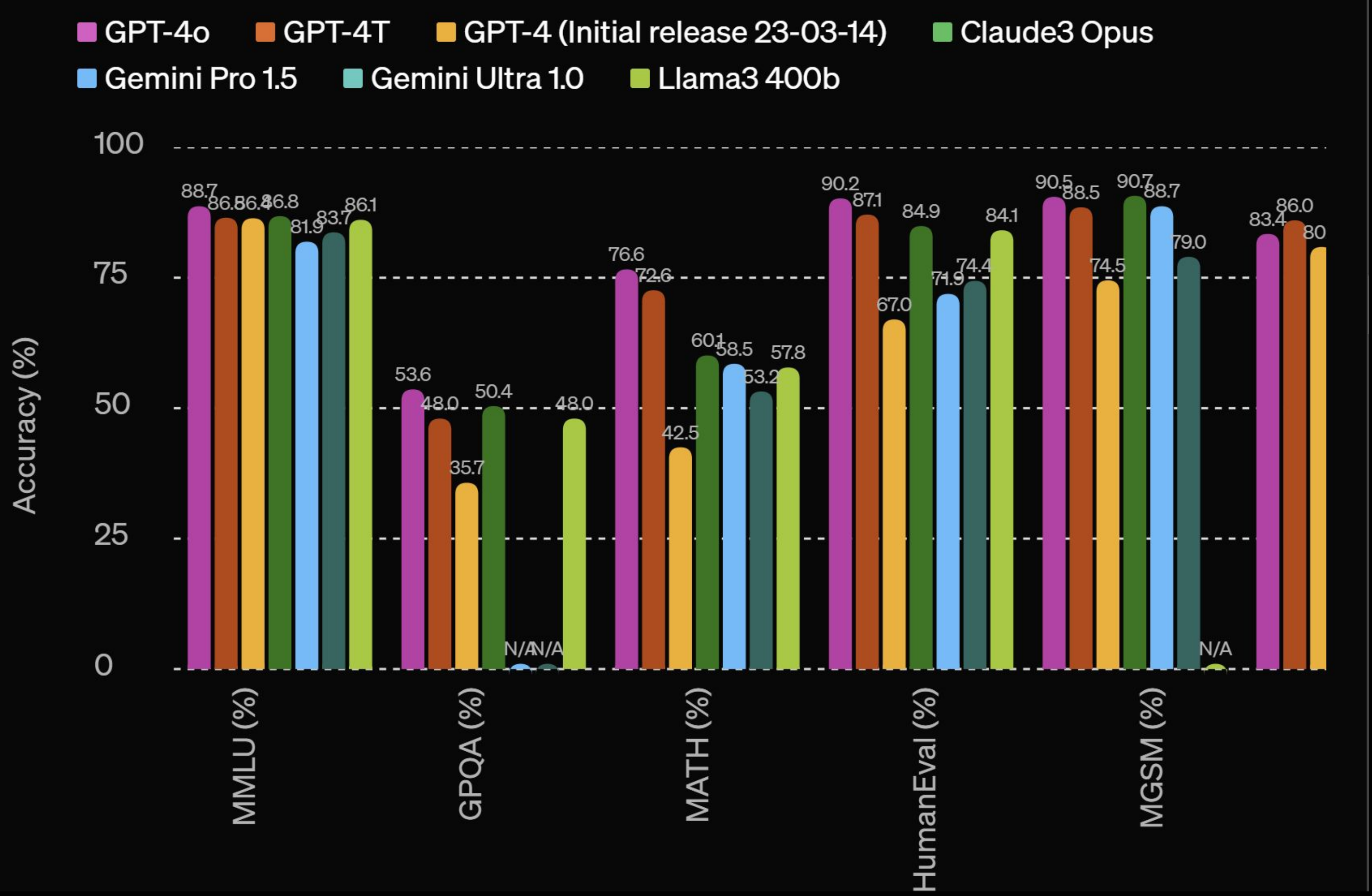
Question

Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Solution

Weng earns 12/60 = 12/60=0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = 0.2*50=10.

Requires arithmetic operations with fractions



https://openai.com/index/hello-gpt-4o/

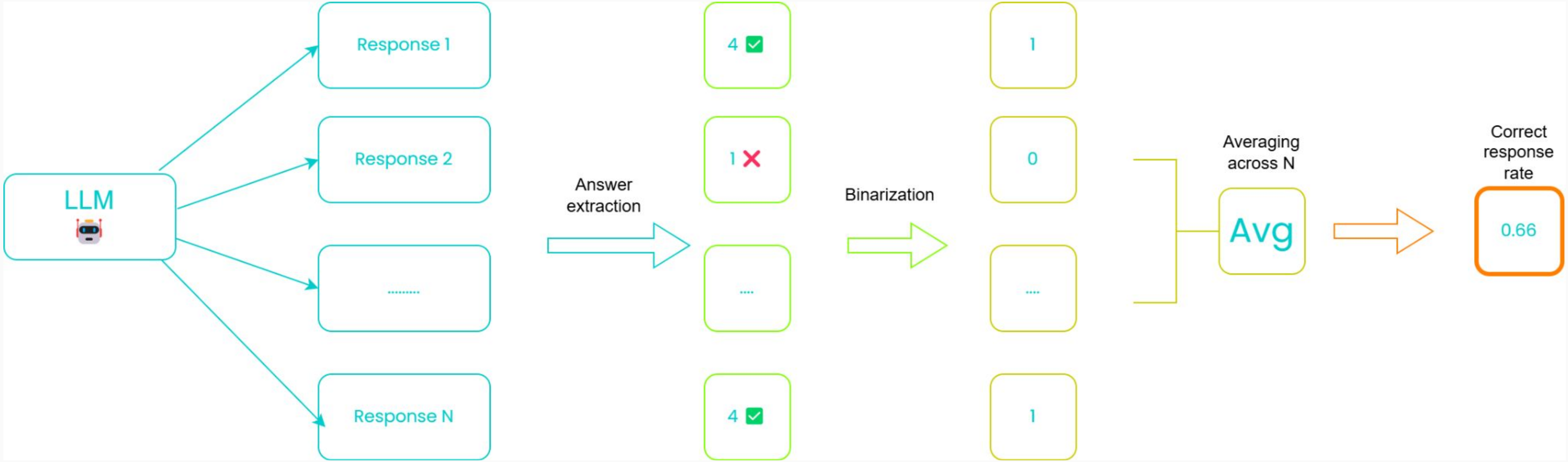
Can an LLM that is able to solve PhD-level tasks solve simple problems?

Welcome to Wonderland!

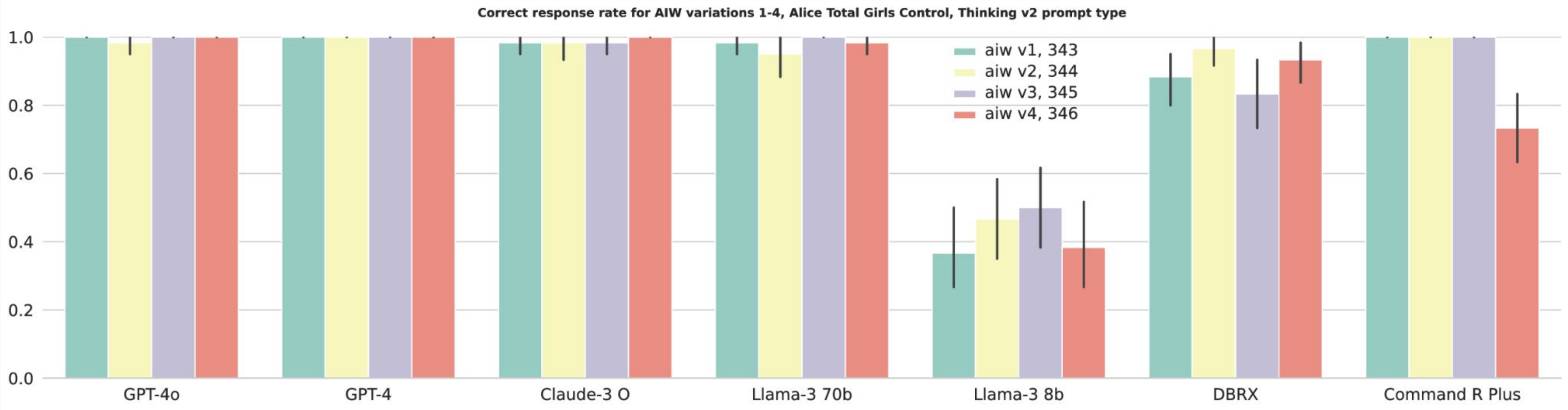
- Very short problem statement.
- No required specific knowledge to solve.
- No advanced arithmetics (the only operation is incrementing by one.
- Example on the right →

Alice has N brothers and she also has M sisters. **How many girls are in the family?**

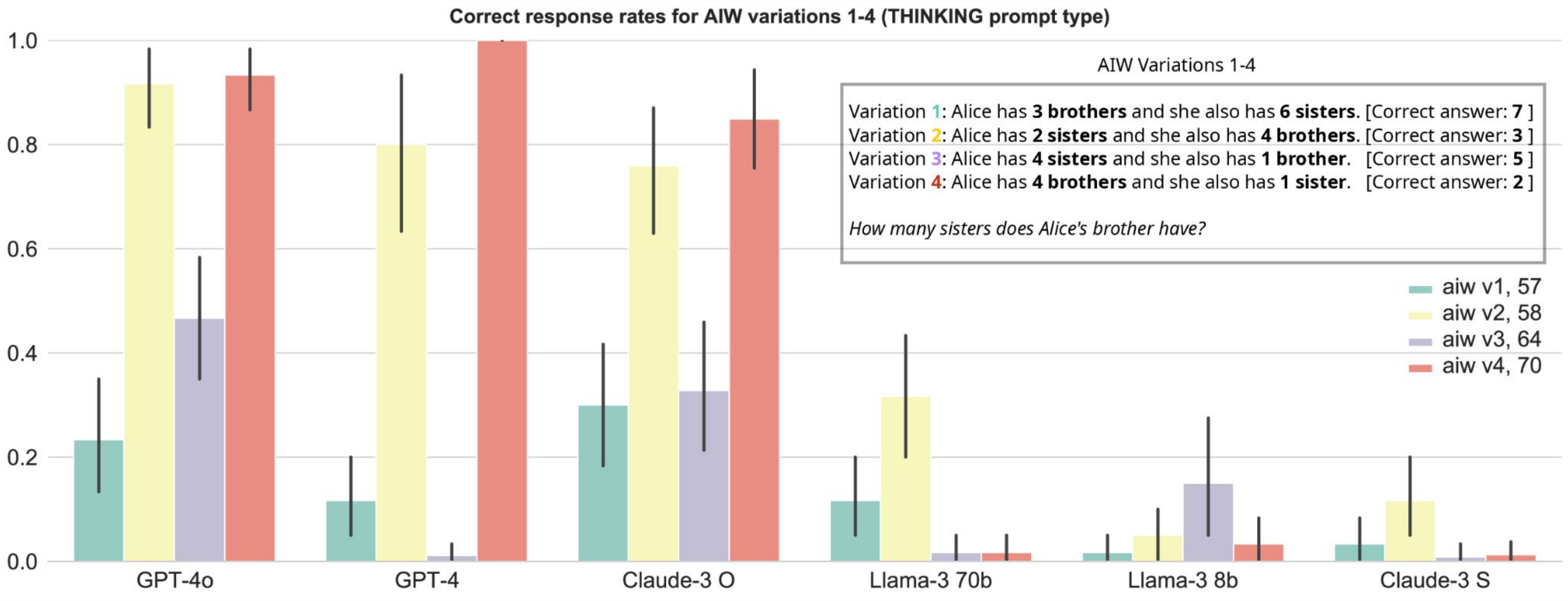
Evaluation procedure: for each problem variation sample N times, binarize answers (1 - correct, 0 - incorrect). Average across N.



Question: Alice has M brothers and she also has N sisters. How many girls are in the family? (Answer: N+1)



Question: Alice has M brothers and she also has N sisters. How many sisters does Alice brother have? (Answer: N+1)



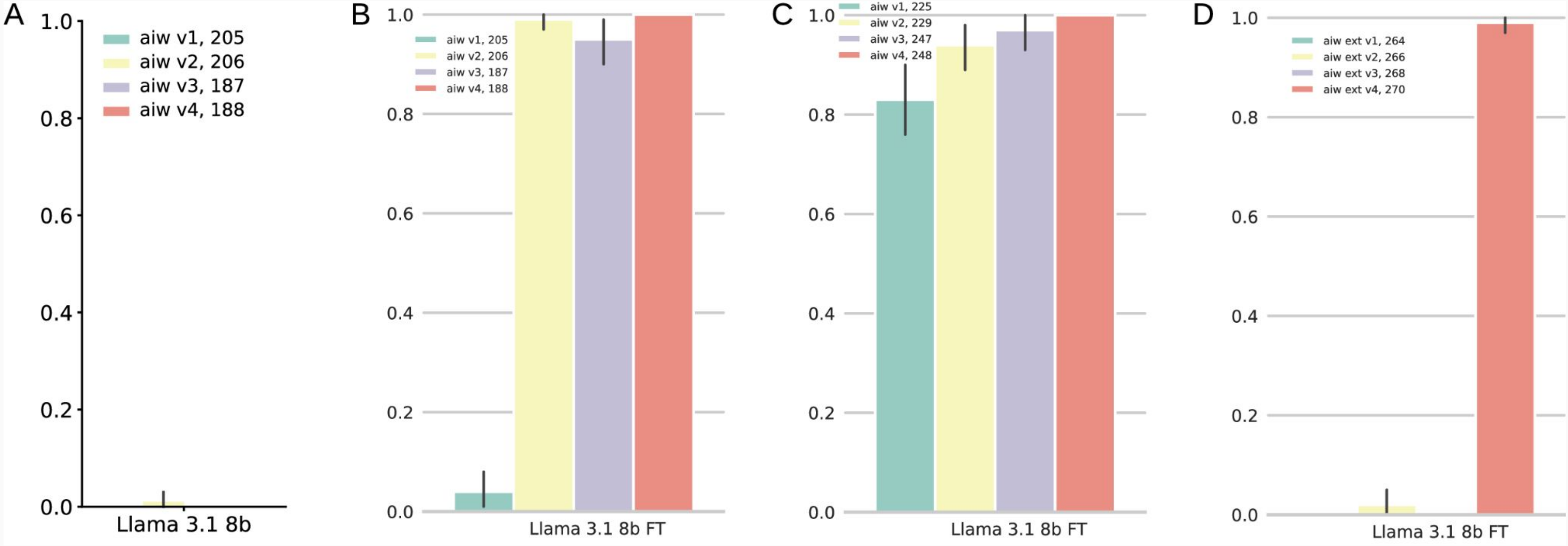
Correct response rates for AIW variations 1-4 (THINKING prompt type)

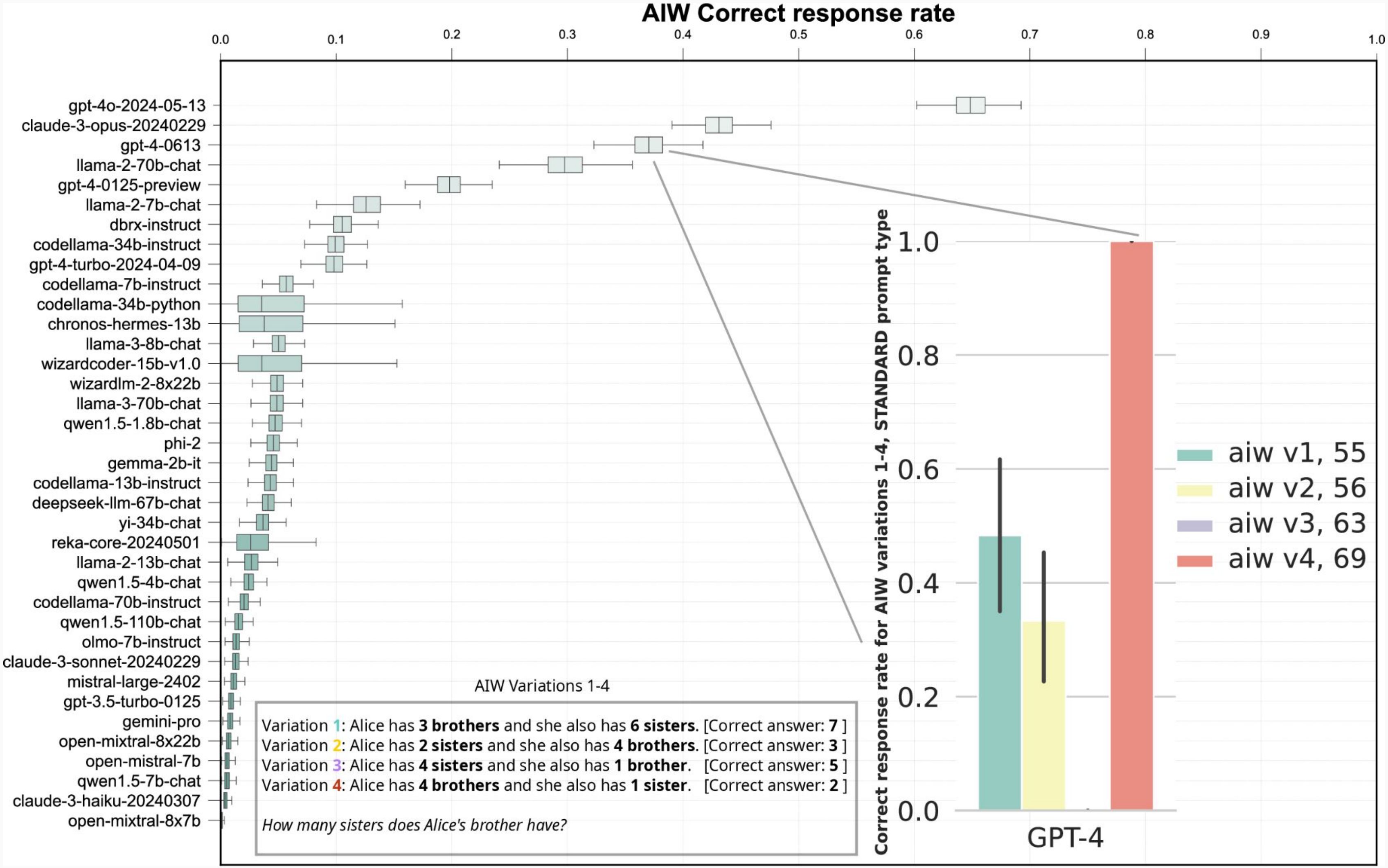
AIW Variations 1-4

Variation **1**: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation **2**: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation **3**: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation **4**: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]

How many sisters does Alice's brother have?

Hypothesis: some problems are in the training data



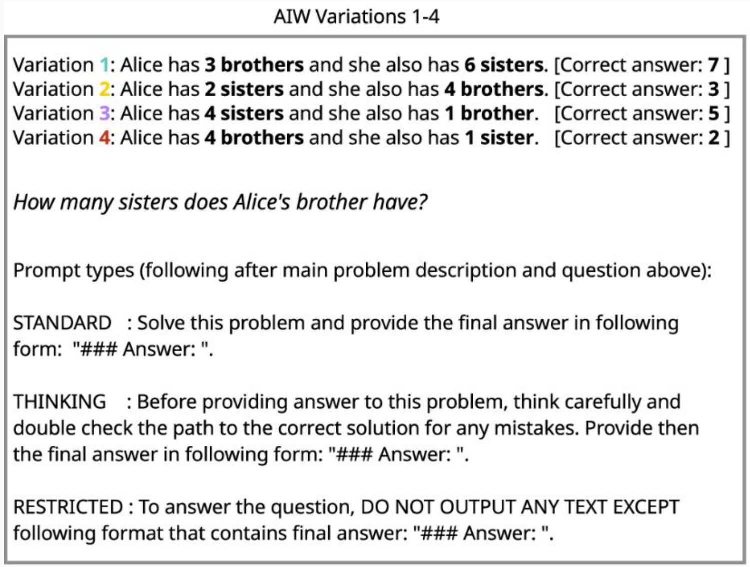


AIW Variations 1-4

Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]

How many sisters does Alice's brother have?

Overview of AIW



AIW Variations 1-4

Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]

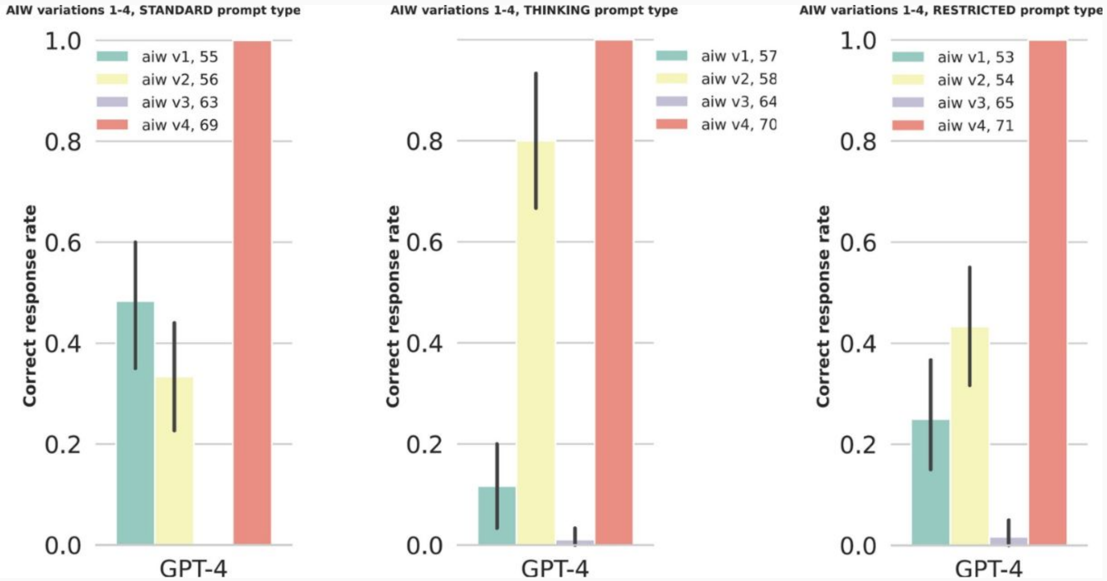
How many sisters does Alice's brother have?

Prompt types (following after main problem description and question above):

STANDARD : Solve this problem and provide the final answer in following form: "### Answer: ".

THINKING : Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".

RESTRICTED : To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".

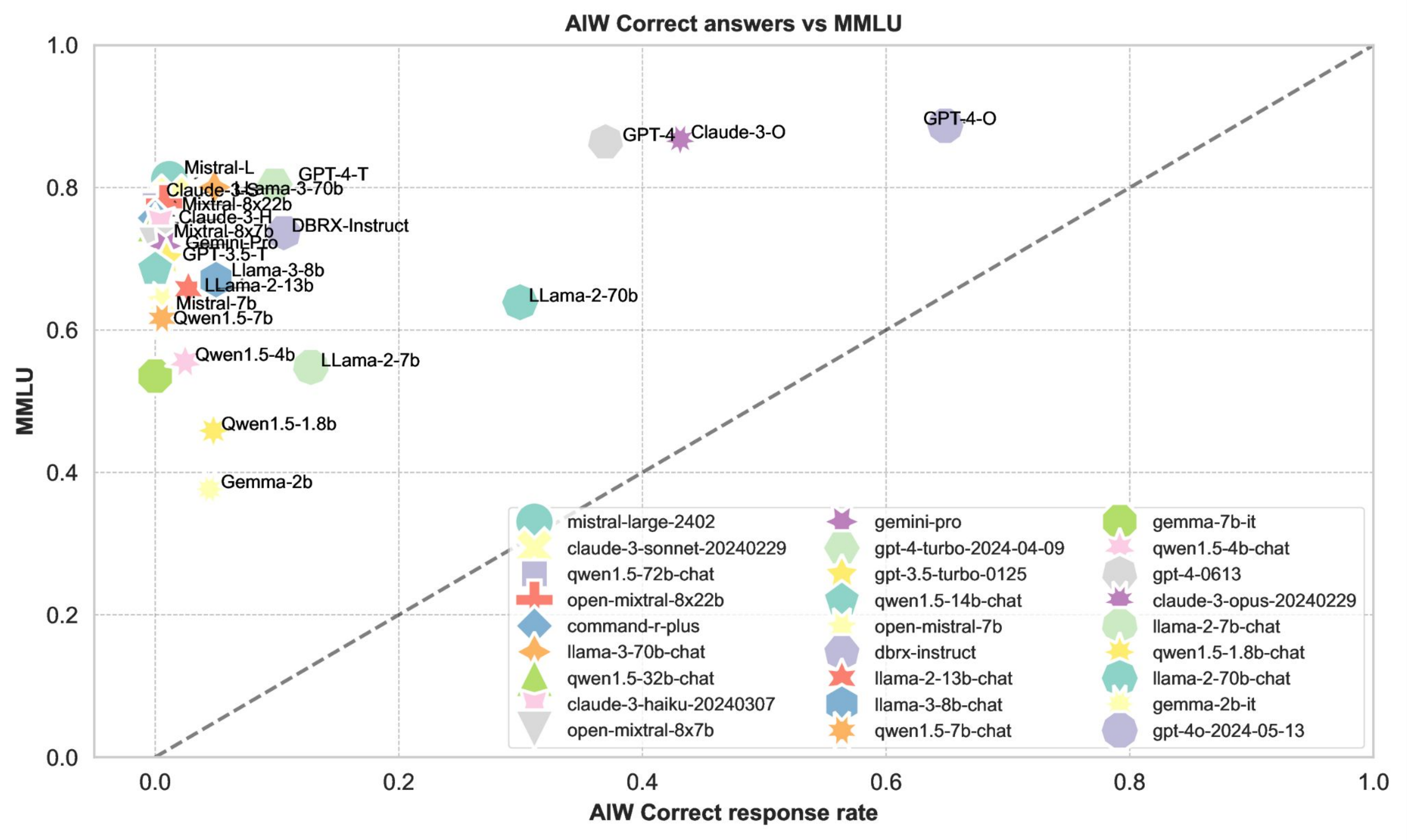


AIW variations 1-4, STANDARD prompt type

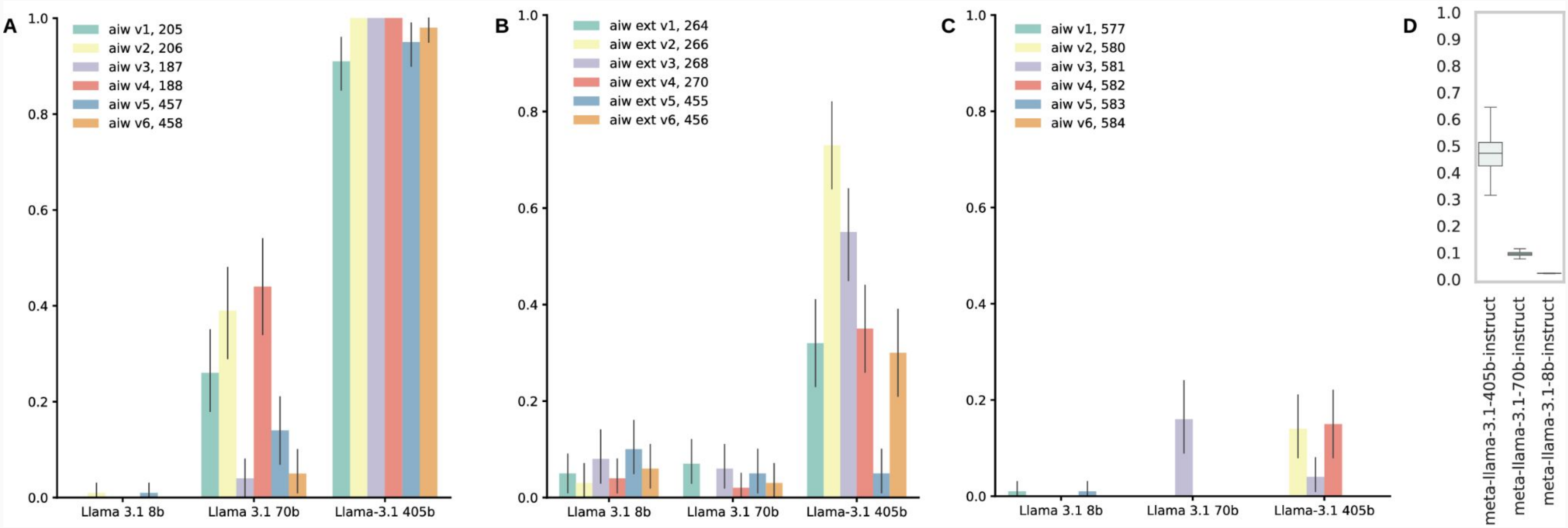
AIW variations 1-4, THINKING prompt type

AIW variations 1-4, RESTRICTED prompt type

Performance on AIW vs a standardized benchmark (MMLU)



Scale is not all you need but it’s still important



Outlook

- **Dynamic benchmarks:** remove confound of data leakage and models being more familiar with question structure.
- **Simple yet hard to cheat on questions:** evaluate basic capabilities that require abstract reasoning and ability to generalize not only *very hard* problems.