Can We Assess Language Models on Unsolved Questions?

Ludwig's Lab Meeting

April 13, 2026

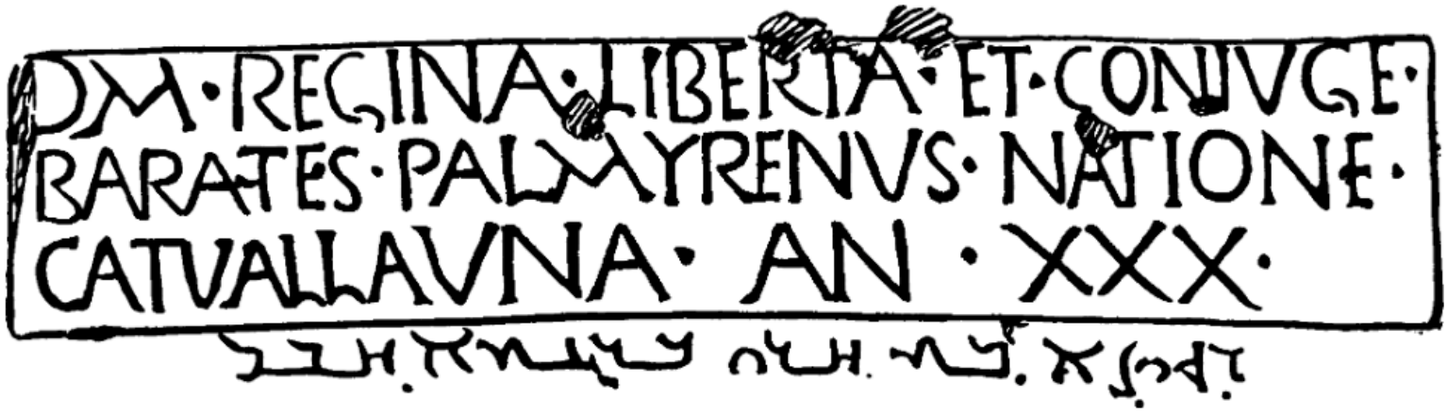
Fan Nie

niefan@stanford.edu

Some benchmarks are difficult...

Classics

Question:



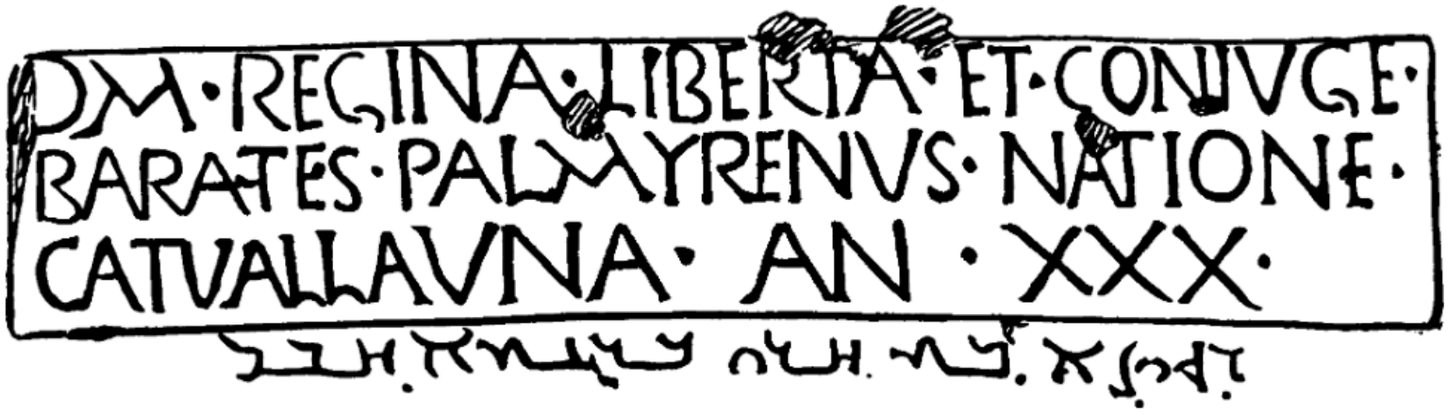
Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYNᵓ BT ḤRY BR ʿTᵓ ḤBL

Henry T
Merton College, Oxford

Some benchmarks are difficult...

📖 Classics

Question:



Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYNᵓ BT ḤRY BR ʿTᵓ ḤBL

Henry T
Merton College, Oxford

But... do people actually care about it?

Some benchmarks are realistic...

Chatbot Arena

Model A

Correct grammar: I are happy.

I'm happy to help you with your English grammar. Can you please provide me with a sentence or two that you would like me to correct?

Model B

Correct grammar: I are happy.

The correct grammar would be "I am happy."

👈 A is better | 👉 B is better | 🤝 Tie | 👎 Both are bad

But...

The Leaderboard Illusion

Shivalika Singh[*1], Yiyang Nan[1], Alex Wang[2], Daniel D'souza[1], Sayash Kapoor[3], Ahmet Üstün[1], Sanmi Koyejo[4], Yuntian Deng[5], Shayne Longpre[6], Noah A. Smith[7,8], Beyza Ermis[1], Marzieh Fadaee[♦1], and Sara Hooker[♦1]

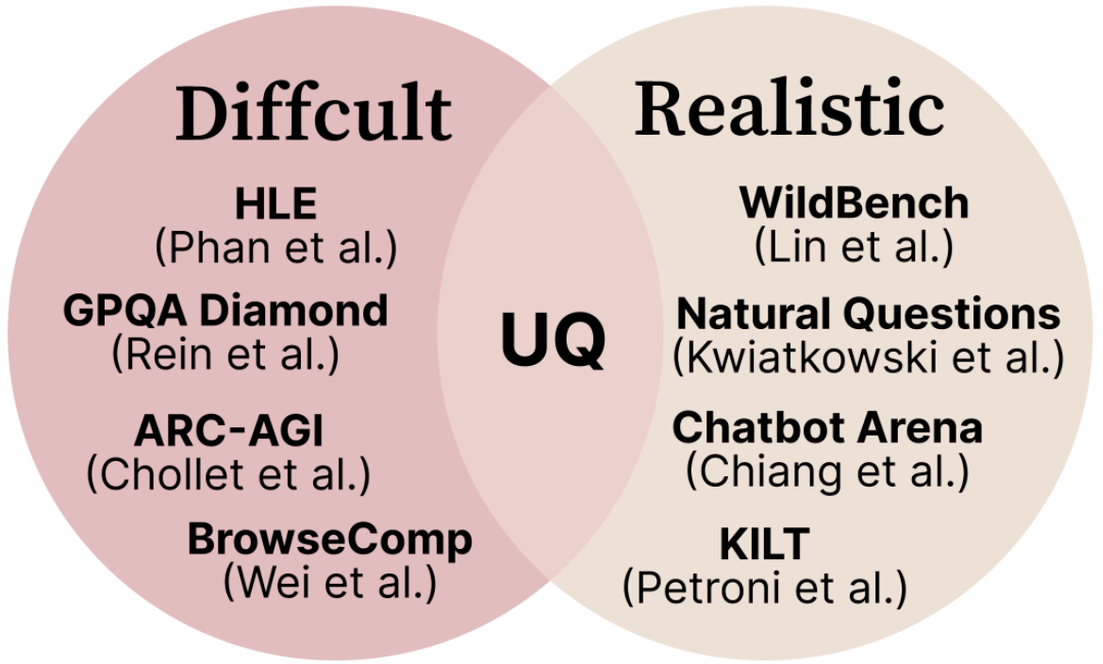
[1]Cohere Labs, [2]Cohere, [3]Princeton University, [4]Stanford University, [5]University of Waterloo, [6]Massachusetts Institute of Technology, [7]Allen Institute for Artificial Intelligence, [8]University of Washington

Corresponding authors: {shiv

Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators

Yann Dubois[1], Balázs Galambosi[2], Percy Liang[1] and Tatsunori B. Hashimoto[1]
[1]Stanford University [2]Independent Researcher

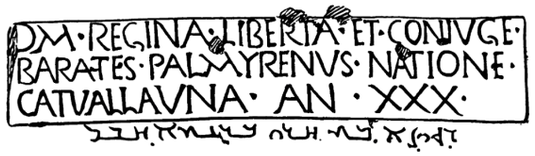
The Difficulty vs Realism Tension



The Difficulty vs Realism Tension

Classics

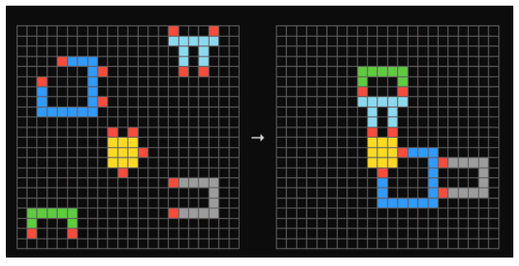
Question:

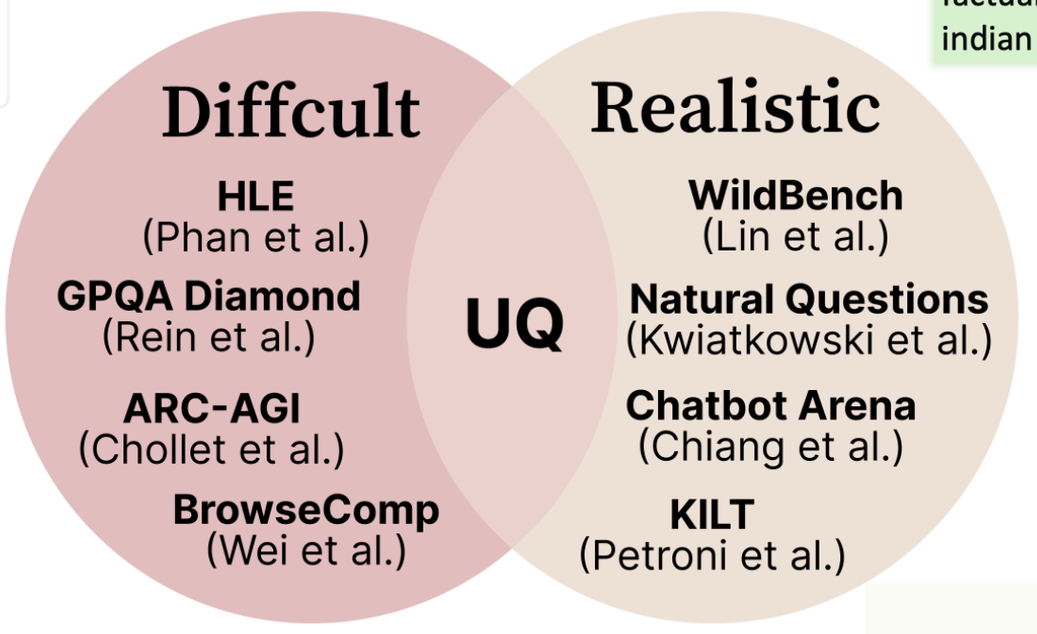


Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYNᵒ BT ḤRY BR ᶜTᵒ ḤBL

Henry T
Merton College, Oxford

hey can you write an essay on the impact of the G20 summit on the global economy, trade, development and the role of young people in shaping the future of the world, it has to have more than 1200 words. Write it beautiful and poetic. Use extensive vocabulary. Use a lot of factual and empirical data. Use some, ancient indian historical references.





Quantum Mechanics

Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p, of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1 - \frac{3p}{4}}$, $A_1 = \sqrt{\frac{p}{4}}X$, $A_2 = \sqrt{\frac{p}{4}}Y$, and $A_3 = \sqrt{\frac{p}{4}}Z$. What could be the correct Kraus Representation of the state $E(\rho)$?

A) $E(\rho) = (1-p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z$
B) $E(\rho) = (1-p)\rho + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$
C) $E(\rho) = (1-p)\rho + \frac{p}{4}X\rho X + \frac{p}{4}Y\rho Y + \frac{p}{4}Z\rho Z$
D) $E(\rho) = (1-p)\rho^2 + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$

Arena

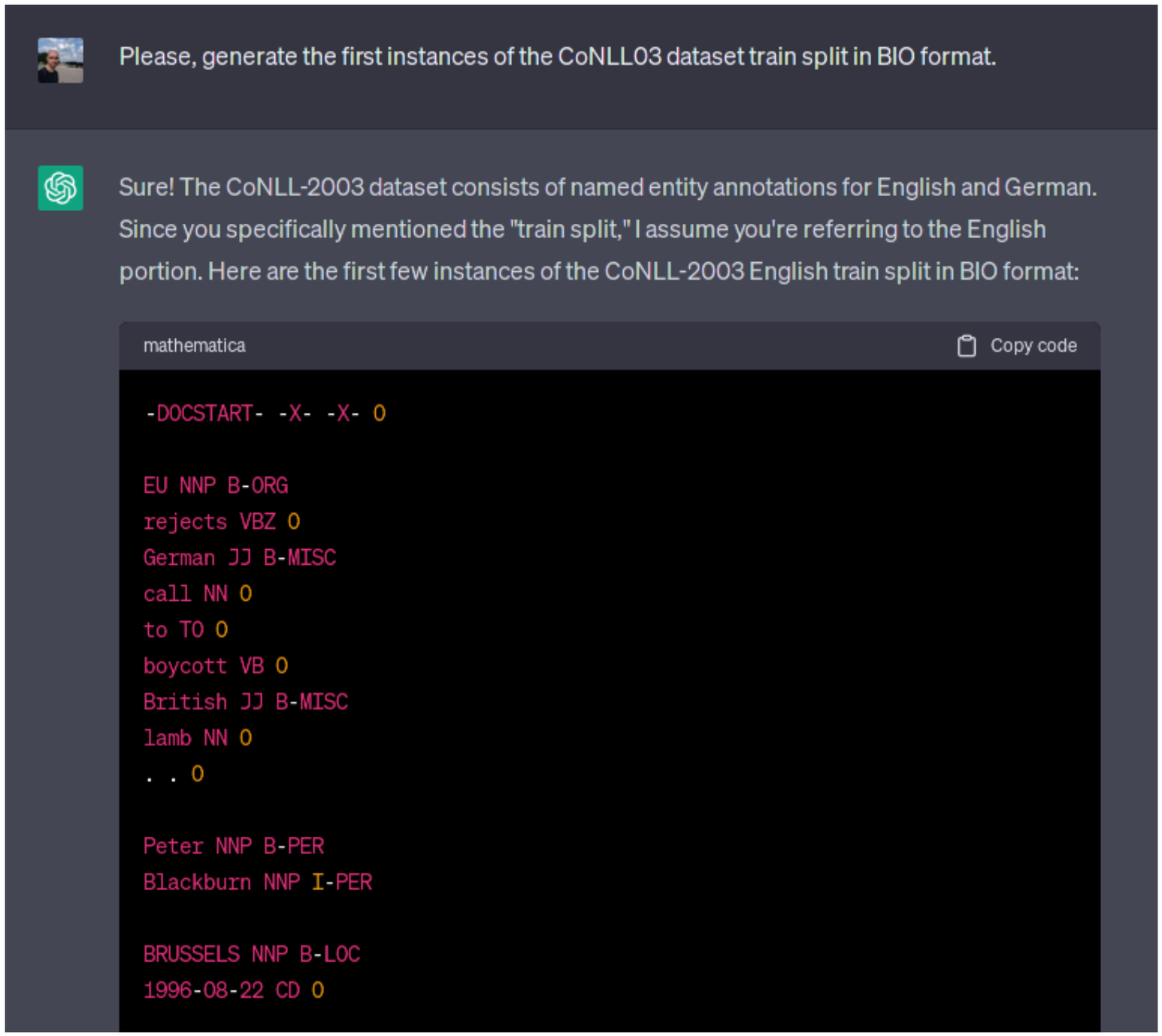
Experience the *frontier*

Polish this email for me:

Hi Sophia, Thank you for the email! Sorry I didn't receive your previous email. I'm still interested in this position, and could you help find another team? I attached my updated resume.

Add files

Benchmarks can also be contaminated...



Benchmarks can also be contaminated...

Example from MATH-500

Problem: For how many positive integers $n > 1$ is it true that 2^{24} is a perfect n^{th} power?
Answer: $\boxed{7}$

Partial Prompt

Prompt: For how many positive integers $n > 1$ is

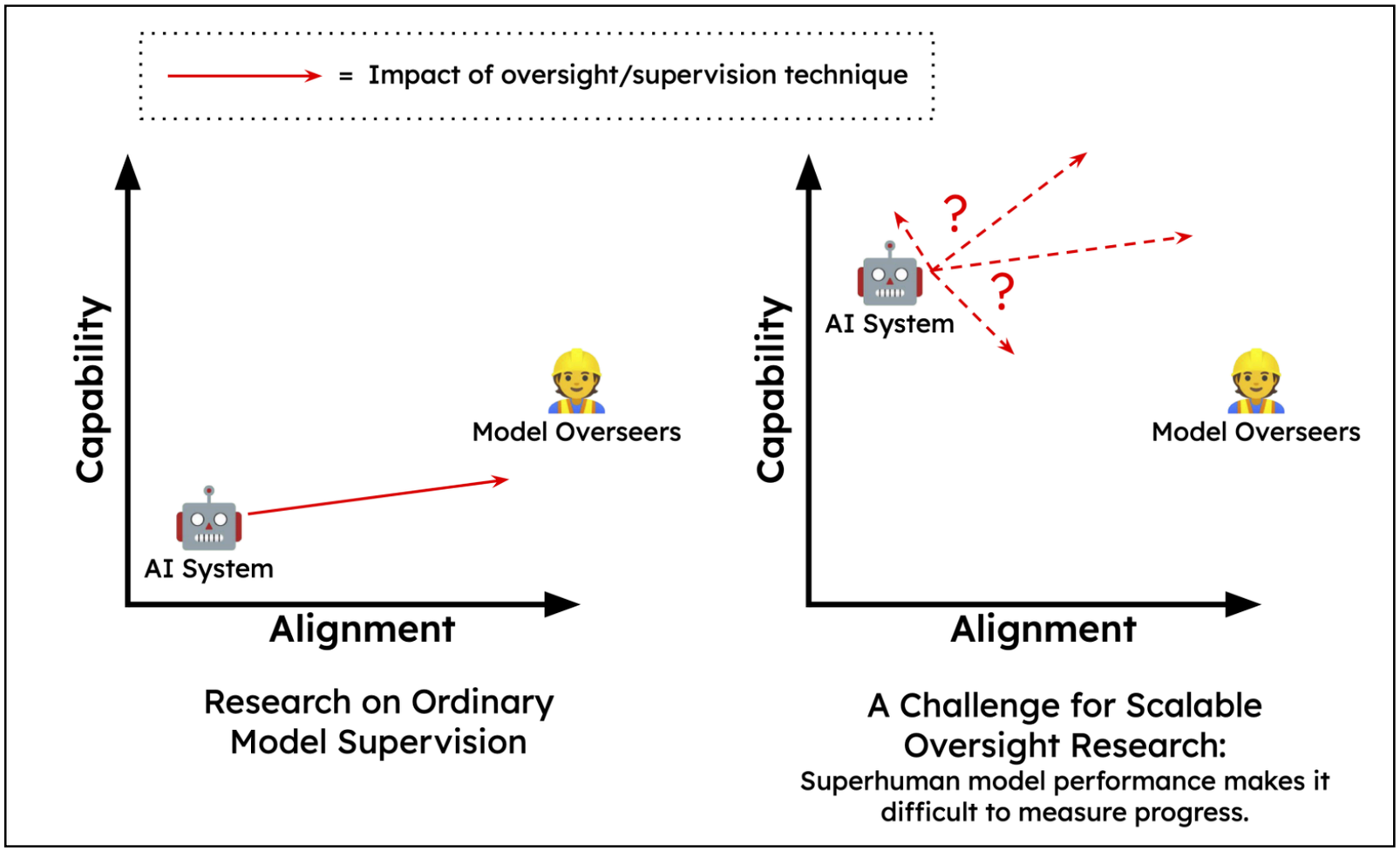
Qwen2.5-7B Output

it true that 2^{24} is a perfect n^{th} power?
To determine for how many positive integers $n > 1$ the number 2^{24} is a perfect n-th power...
Let's count these values. There are 7 such values...
The final answer is $\boxed{7}$.

Llama3.1-8B Output

the set of positive divisors of n (including 1 and n) less than 50 ?
(A) 15 **(B)** 16 **(C)** 17 ...
The number of positive divisors of n is given by...
The number of factors of each of these numbers is...
The answer is $\boxed{\textbf{(B)}\ 16}$

Also, the models are getting too smart...
Human supervision struggles to keep up with model capability!



What's something that's realistic, difficult, and uncontaminated at the same time????

What's something that's realistic, difficult, and uncontaminated at the same time????

Unsolved Questions!

Just evaluate on unsolved questions!

1. **They are hard! Naturally hard**
2. **They are realistic! they exist because people ask about them, and havent gotten an asnwer**
3. **They better not be contaminated (or they are solved)**

Sample Question from Mathematics

- **Title:** Does every ring of integers sit inside a ring of integers that has a power basis?
- **Keywords:** abstract-algebra, number-theory, ring-theory, algebraic-number-theory
- **Site:** math
- **Link:** https://math.stackexchange.com/questions/1754860

Given a finite extension of the rationals, K, we know that $K = \mathbb{Q}[\alpha]$ by the primitive element theorem, so every $x \in K$ has the form

$$x = a_0 + a_1\alpha + \cdots + a_n\alpha^n,$$

with $a_i \in \mathbb{Q}$.

However, the ring of integers, $\mathcal{O}_K$, of K need not have a basis over $\mathbb{Z}$ which consists of 1 and powers of a single element (a power basis). In fact, there exist number fields which require an arbitrarily large number of elements to form such a basis.

Question: Can every ring of integers $\mathcal{O}_K$ that does not have a power basis be extended to a ring of integers $\mathcal{O}_L$ which does have a power basis, for some finite L/K?

Example question from UQ

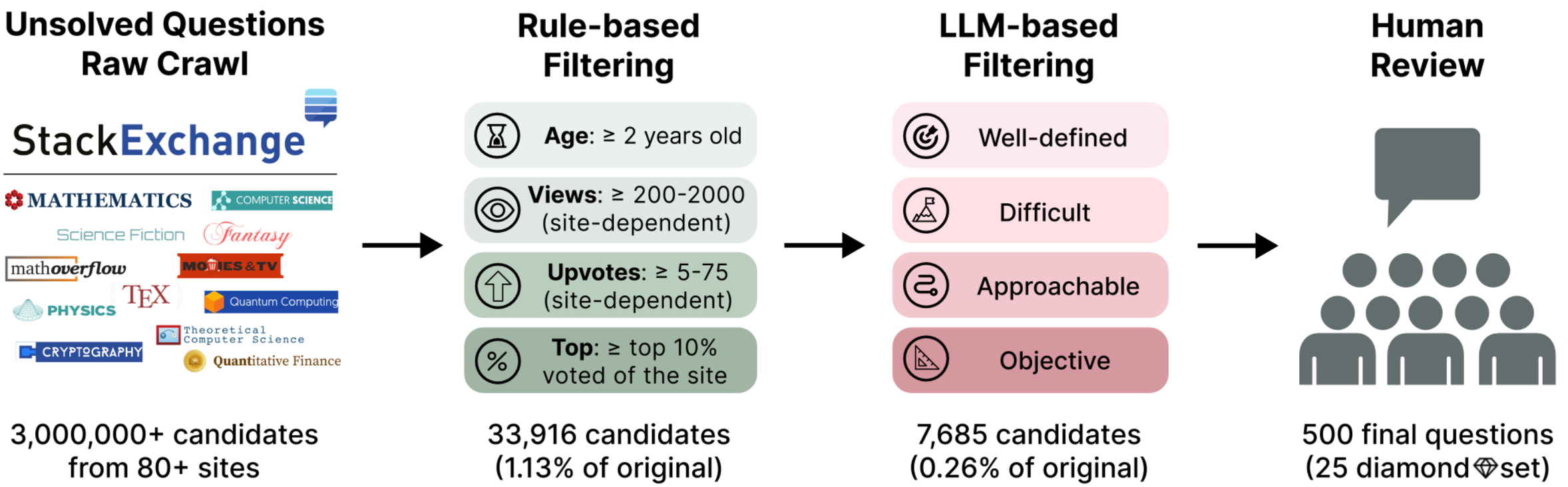
but how do you evaluate on unsolved questions??

Questions:

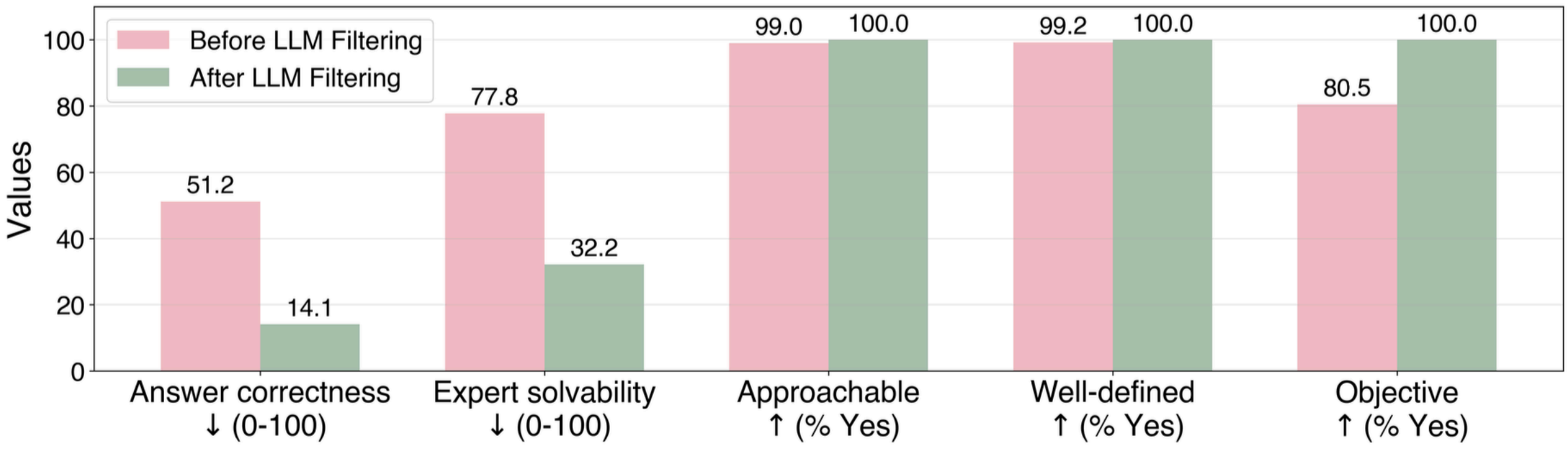
- (Dataset) **Collection:** What makes a good unsolved question?
- (Validator) **Validation:** we don't have answers ..
- (Platform) **Verification**: can we ever know if an answer is correct?...

How to collect and validate the difficulty and quality of questions?

UQ-Dataset: Collection Pipeline

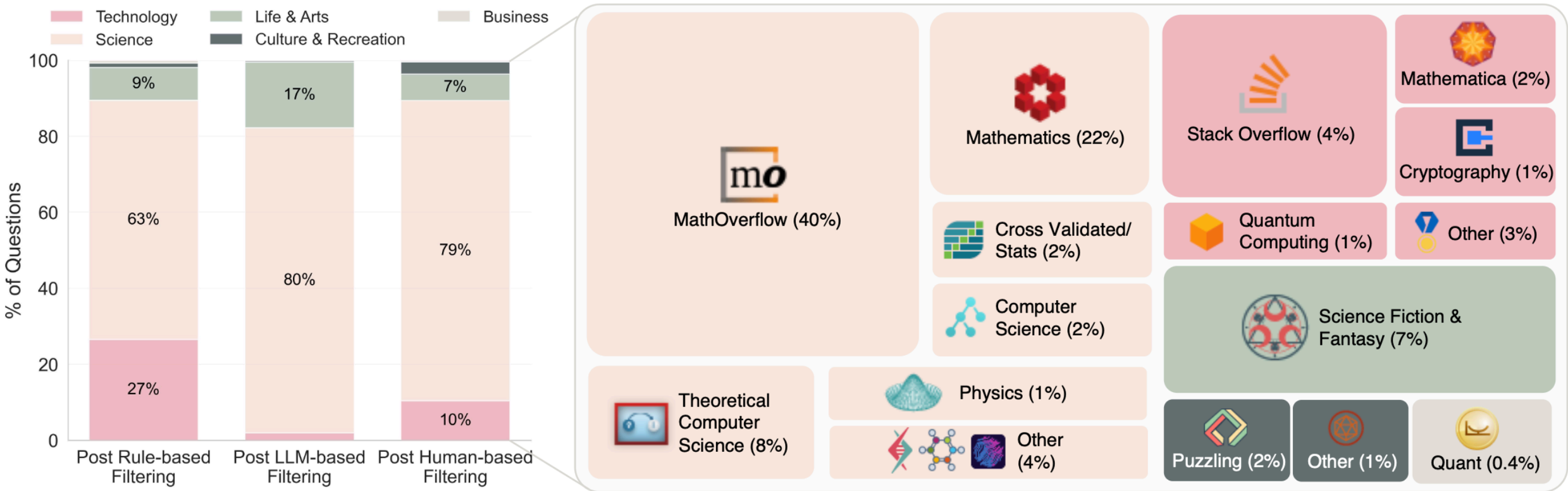


UQ-Dataset: LLM-based questions filters



- Approachable, Well-defined, Objective increases to 100% (we discard any question that fail these metrics)
- Attempted answer correctness and expert solvability largely drop

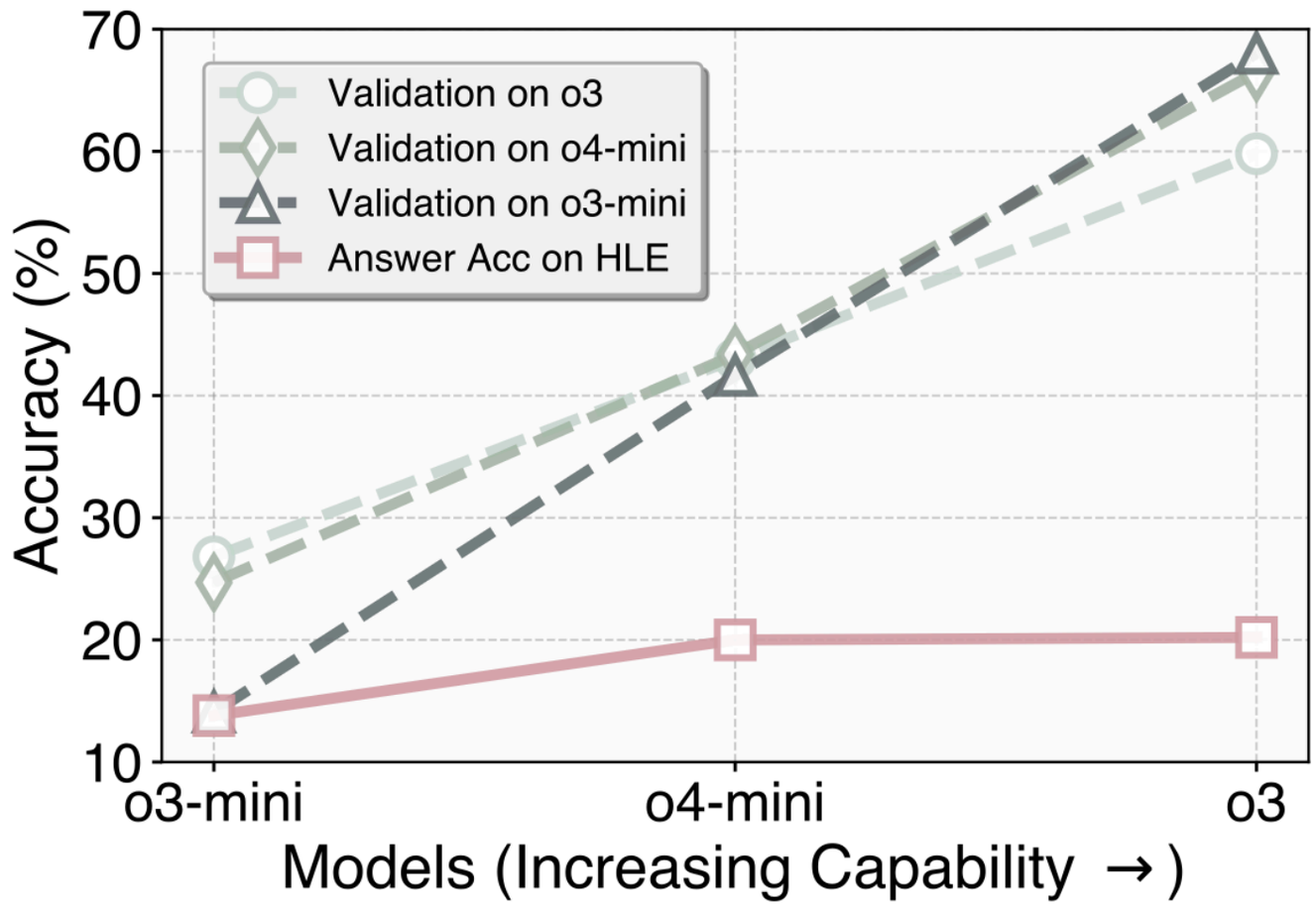
UQ-Dataset: Question Composition



- Left: high-level composition across each of the three-stage filtering
- Right: question composition of UQ-Dataset by Stack Exchange sites

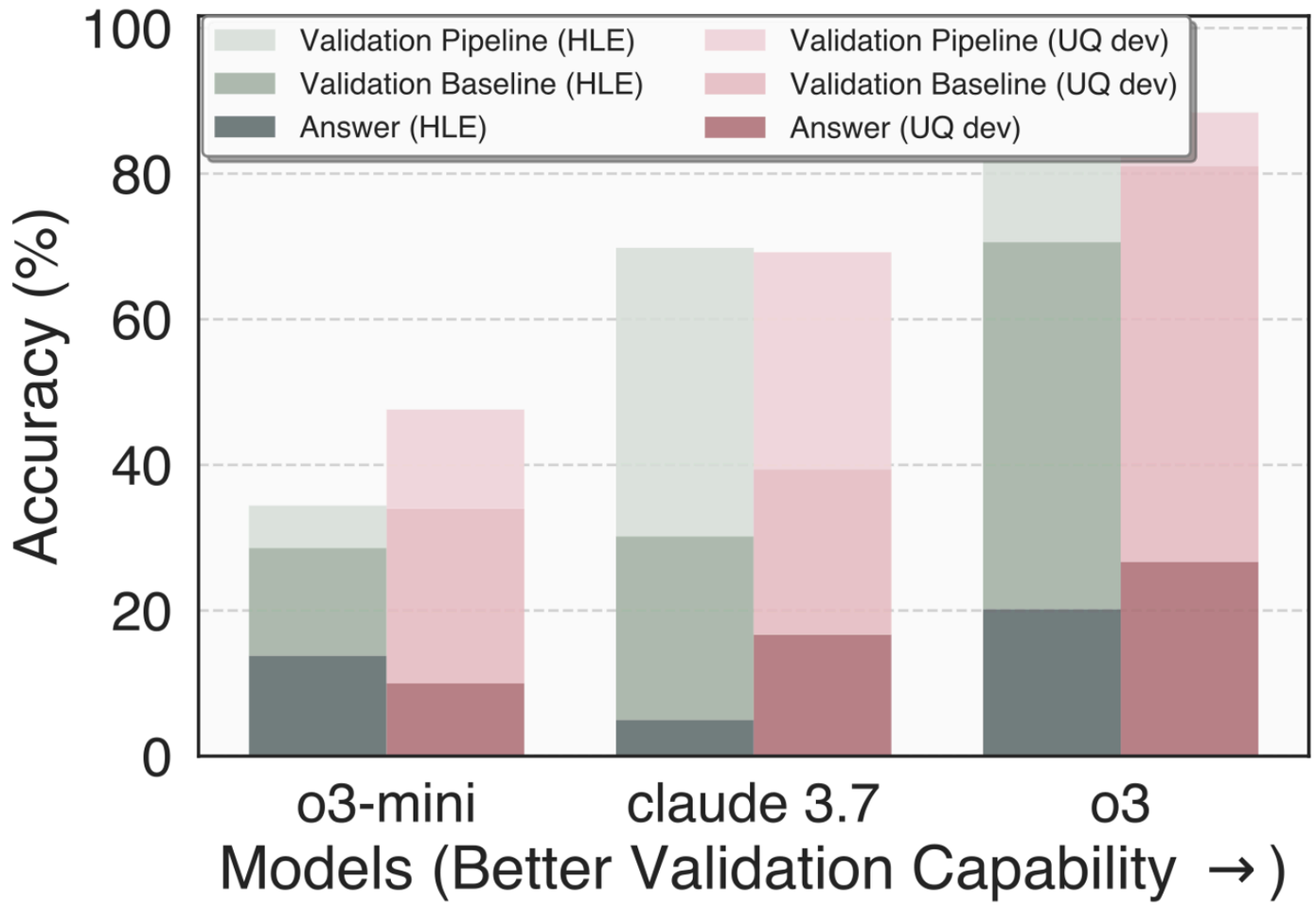
How to assess candidate solutions produced by different models?

Verifying candidate answers to hard questions may be easier than generating them?



Generator-Validator Gap Widens with Model Capability

Verifying candidate answers to hard questions may be easier than generating them?

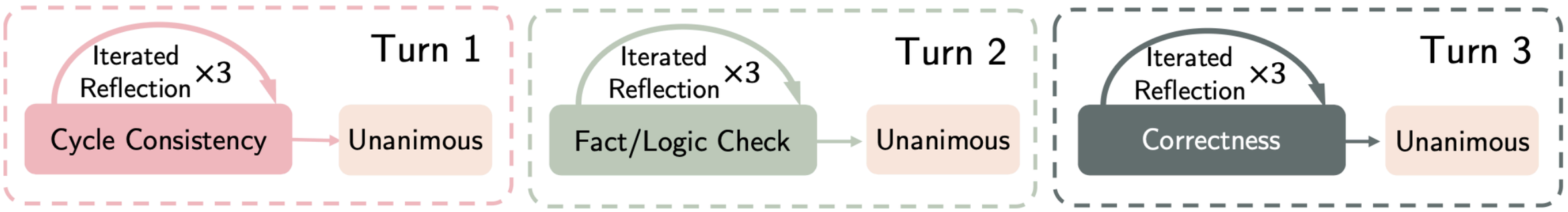


Validator performance shows transfer

UQ-Validator: Design Space

- Low-level Strategies
  - Cycle-consistency, fact/logic check, total correctness, etc
- Mid-level Strategies
  - Repeated sampling, iterated reflection, etc
- High-level Strategies
  - Majority-voting, unanimous-voting, pipeline-verification, etc

Illustration of the default, performant UQ-Validator



Finding 1: Compound Validator Strategies Outperform Simple Prompting Baselines

Model	Strategy	Accuracy (%)	Precision (%)	Recall (%)
Claude Sonnet 3.7	Vanilla Prompt (Baseline)	21.60	13.26	90.77
	Correctness	30.20	14.85	92.31
	Correctness × 5 | Majority	29.40	14.53	90.77
	Correctness × 5 | Unanimous	41.20	15.82	81.52
	Correctness ↻ 5 | Unanimous	54.32	23.08	56.25
	3-Iter Pipeline	73.20	20.00	16.00
o3-mini	Vanilla Prompt (Baseline)	24.00	14.29	96.92
	Correctness	28.60	15.24	98.46
	Correctness × 5 | Majority	29.20	15.18	96.92
	Correctness × 5 | Unanimous	33.00	15.56	93.85
	Correctness ↻ 5 | Unanimous	30.00	15.16	95.38
	3-Iter Pipeline	34.40	15.84	93.85
o3	Vanilla Prompt (Baseline)	58.12	20.73	78.46
	Correctness	70.60	22.00	50.00
	Correctness × 5 | Majority	73.15	25.87	56.92
	Correctness × 5 | Unanimous	83.77	26.47	13.85
	Correctness ↻ 5 | Unanimous	78.60	28.57	43.08
	1-Iter Pipeline	75.40	24.00	42.00
	3-Iter Pipeline	**81.65**	**30.99**	**34.38**
	5-Iter Pipeline	81.50	26.23	25.40
Multi-model ensemble	Correctness (5 Models) | Majority	45.00	17.99	90.77
	Correctness (5 Models) | Unanimous	78.60	25.00	32.31
	3-Iter Pipeline (2 Models) | Unanimous	**85.40**	**40.00**	**24.62**

Finding 1: Compound Validator Strategies Outperform Simple Prompting Baselines

Model	Strategy	Accuracy (%)		Precision (%)	Recall (%)
Claude Sonnet 3.7	Vanilla Prompt (Baseline)	21.60		13.26	90.77
	Correctness	30.20		14.85	92.31
	Correctness × 5 | Majority	29.40		14.53	90.77
	Correctness × 5 | Unanimous	41.20		15.82	81.52
	Correctness ↻ 5 | Unanimous	54.32		23.08	56.25
	3-Iter Pipeline	73.20	+51.6%	20.00	16.00
o3-mini	Vanilla Prompt (Baseline)	24.00		14.29	96.92
	Correctness	28.60		15.24	98.46
	Correctness × 5 | Majority	29.20		15.18	96.92
	Correctness × 5 | Unanimous	33.00		15.56	93.85
	Correctness ↻ 5 | Unanimous	30.00		15.16	95.38
	3-Iter Pipeline	34.40		15.84	93.85
o3	Vanilla Prompt (Baseline)	58.12		20.73	78.46
	Correctness	70.60		22.00	50.00
	Correctness × 5 | Majority	73.15		25.87	56.92
	Correctness × 5 | Unanimous	83.77		26.47	13.85
	Correctness ↻ 5 | Unanimous	78.60		28.57	43.08
	1-Iter Pipeline	75.40		24.00	42.00
	3-Iter Pipeline	**81.65**		**30.99**	**34.38**
	5-Iter Pipeline	81.50		26.23	25.40
Multi-model ensemble	Correctness (5 Models) | Majority	45.00		17.99	90.77
	Correctness (5 Models) | Unanimous	78.60		25.00	32.31
	3-Iter Pipeline (2 Models) | Unanimous	**85.40**		**40.00**	**24.62**

Finding 2: Attaining High Precision is Difficult

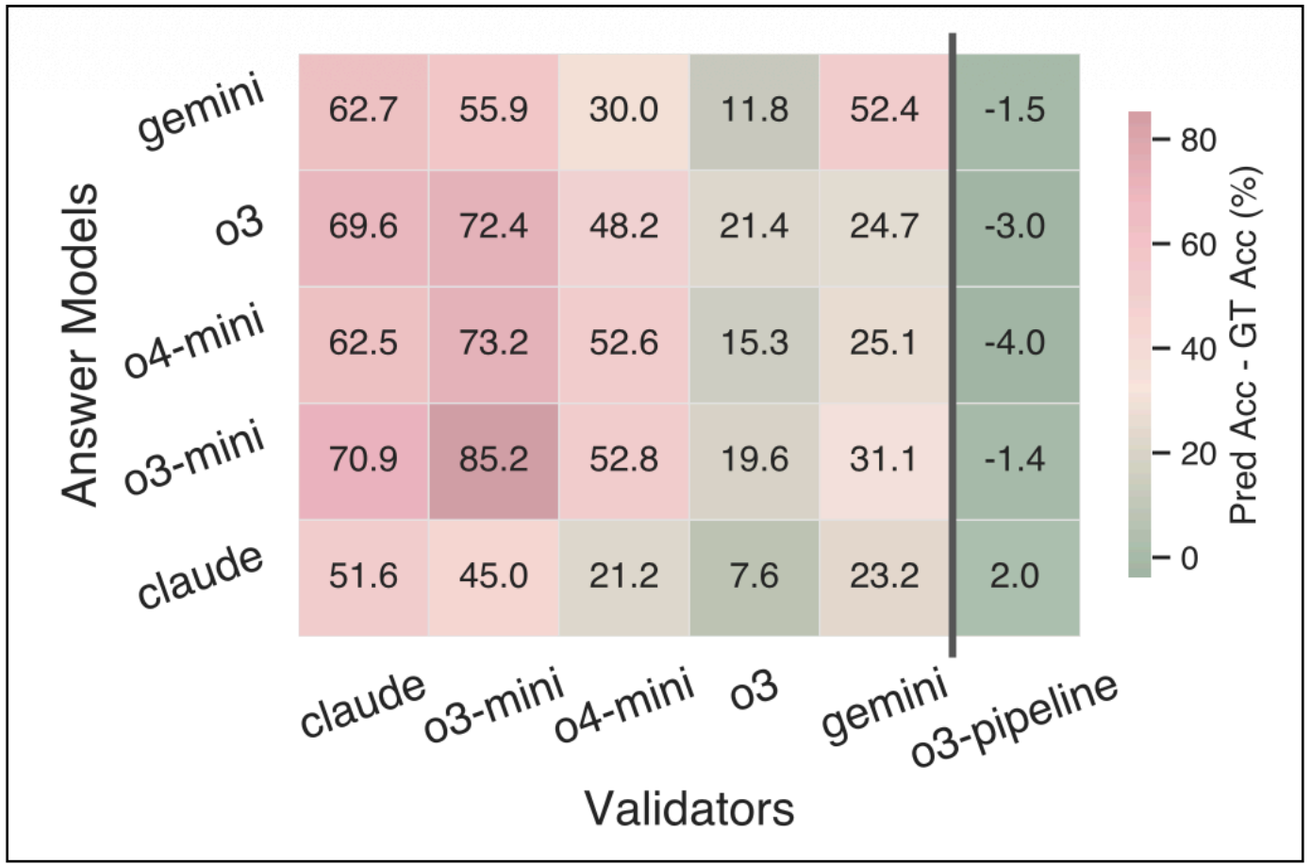
Model	Strategy	Accuracy (%)	Precision (%)	Recall (%)
Claude Sonnet 3.7	Vanilla Prompt (Baseline)	21.60	13.26	90.77
	Correctness	30.20	14.85	92.31
	Correctness × 5 | Majority	29.40	14.53	90.77
	Correctness × 5 | Unanimous	41.20	15.82	81.52
	Correctness ↻ 5 | Unanimous	54.32	23.08	56.25
	3-Iter Pipeline	73.20	20.00	16.00
o3-mini	Vanilla Prompt (Baseline)	24.00	14.29	96.92
	Correctness	28.60	15.24	98.46
	Correctness × 5 | Majority	29.20	15.18	96.92
	Correctness × 5 | Unanimous	33.00	15.56	93.85
	Correctness ↻ 5 | Unanimous	30.00	15.16	95.38
	3-Iter Pipeline	34.40	15.84	93.85
o3	Vanilla Prompt (Baseline)	58.12	20.73	78.46
	Correctness	70.60	22.00	50.00
	Correctness × 5 | Majority	73.15	25.87	56.92
	Correctness × 5 | Unanimous	83.77	26.47	13.85
	Correctness ↻ 5 | Unanimous	78.60	28.57	43.08
	1-Iter Pipeline	75.40	24.00	42.00
	3-Iter Pipeline	**81.65**	**30.99**	**34.38**
	5-Iter Pipeline	81.50	26.23	25.40
Multi-model ensemble	Correctness (5 Models) | Majority	45.00	17.99	90.77
	Correctness (5 Models) | Unanimous	78.60	25.00	32.31
	3-Iter Pipeline (2 Models) | Unanimous	**85.40**	**40.00**	**24.62**

Finding 2: Attaining High Precision is Difficult

Metric	Answer Models			
	o3	**Claude Sonnet 3.7**	**Gemini 2.5 Pro**	**GPT-4o**
% answers passed UQ-Validator	0%	0%	12%	0%
% answers passed human reviewers (i.e., GT accuracy)	0%	0%	4%	0%
Human/UQ-Validator judgment agreement	100%	100%	92%	100%
Human-rated accuracy of UQ-Validator reasoning trace	96%	96%	76%	100%

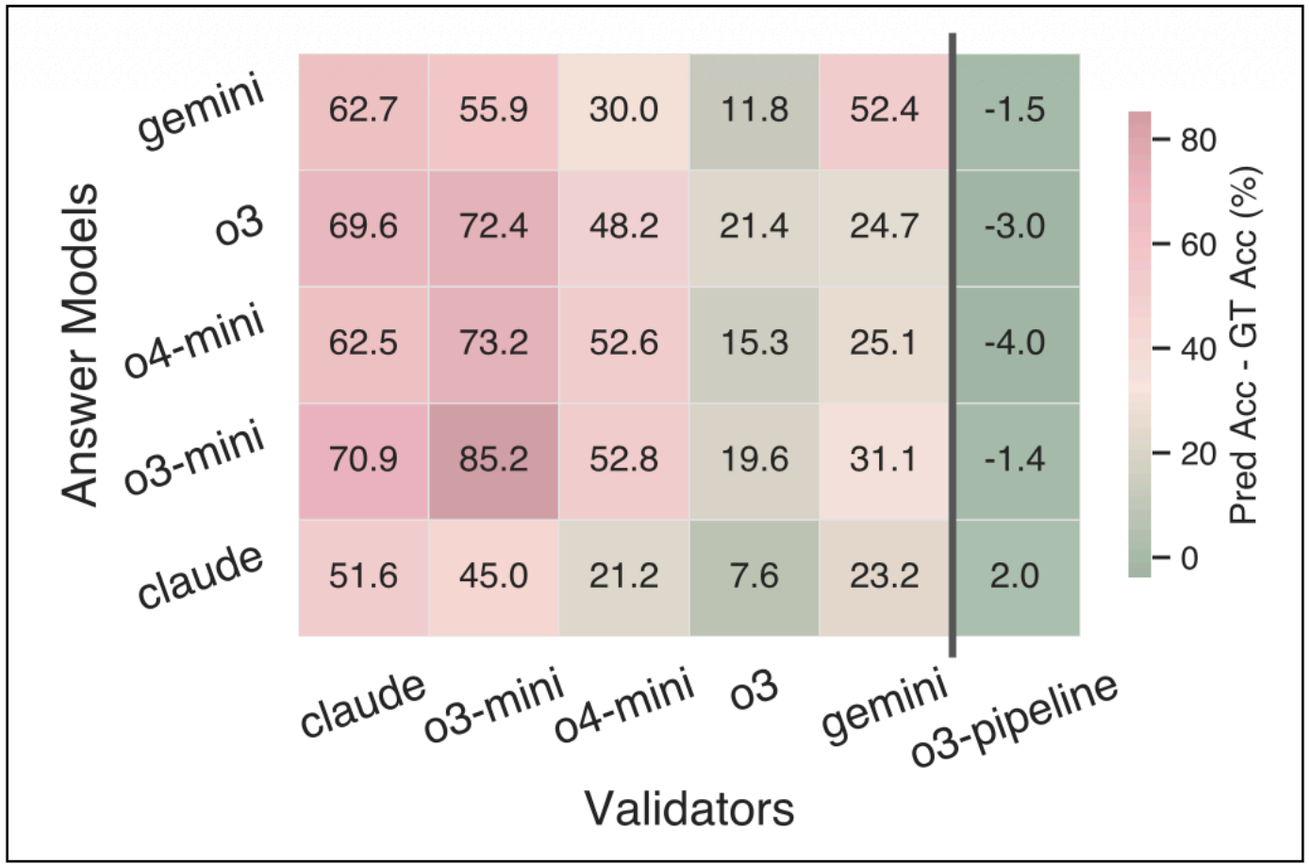
Nevertheless, the best resulting UQ-Validator remains useful for human reviewers!

Finding 3: Simple Validators Show Over-Optimism and Self-Bias



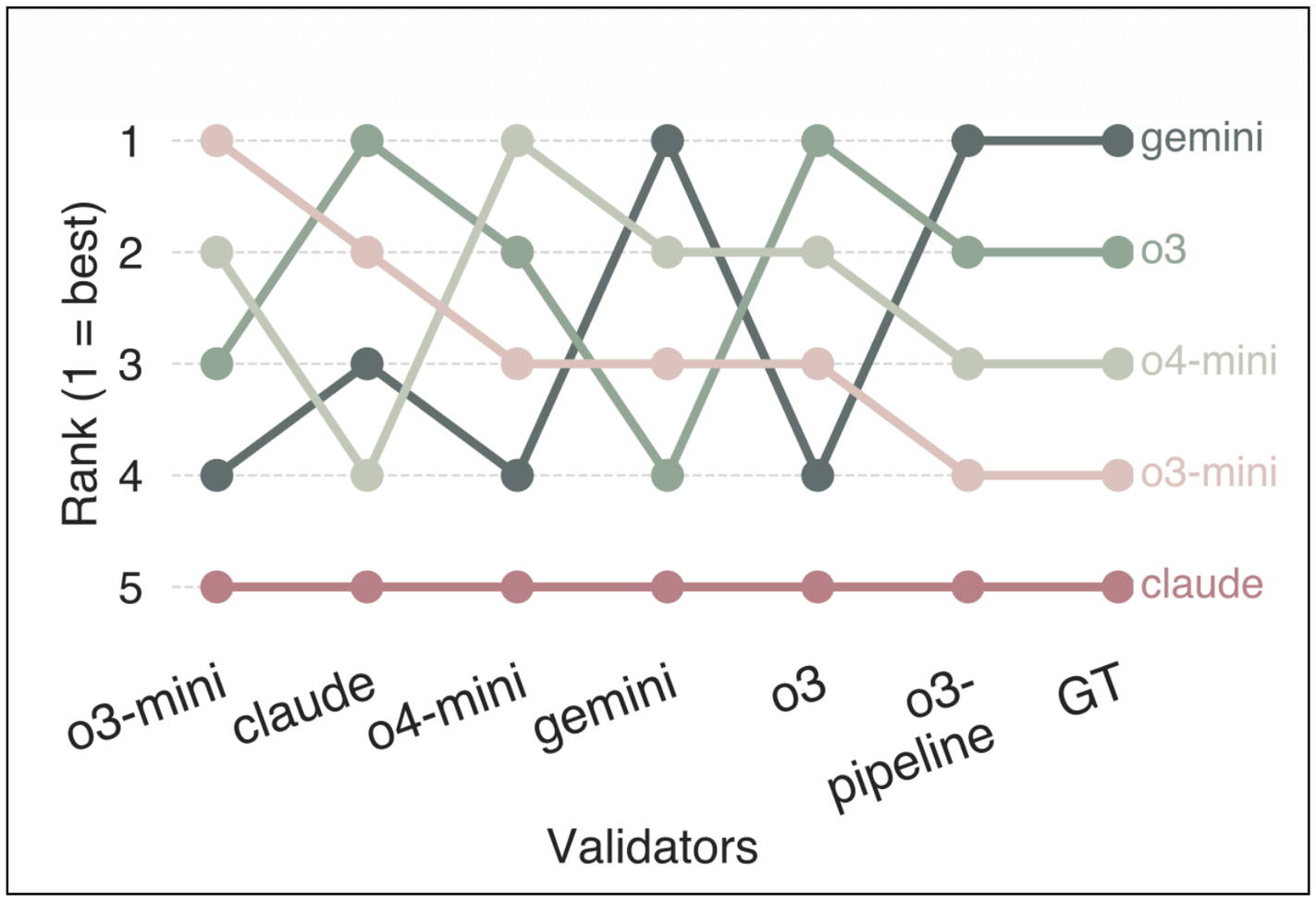
Simple LLM validators overrate self and sibling answers.

Finding 4: Compound Validator Strategies Mitigate Over-Optimism and Self-Bias



3-iter o3 pipeline largely removes over-optimism across all models!

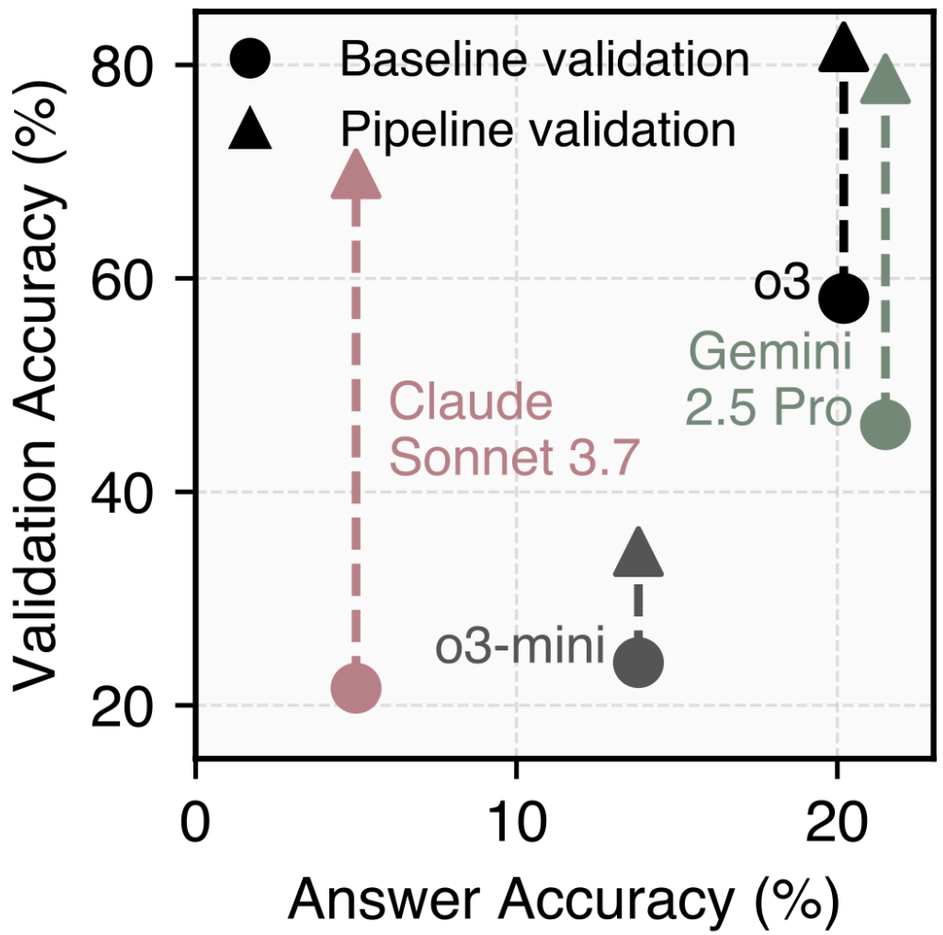
Finding 5: Model Rankings Are Unstable Across Validator Performance



Model ranking is unstable across validator performance

Using 3-iter pipeline aligns better with GT ranking

Finding 6: Better Answer Generators May Not Be Better Answer Validators

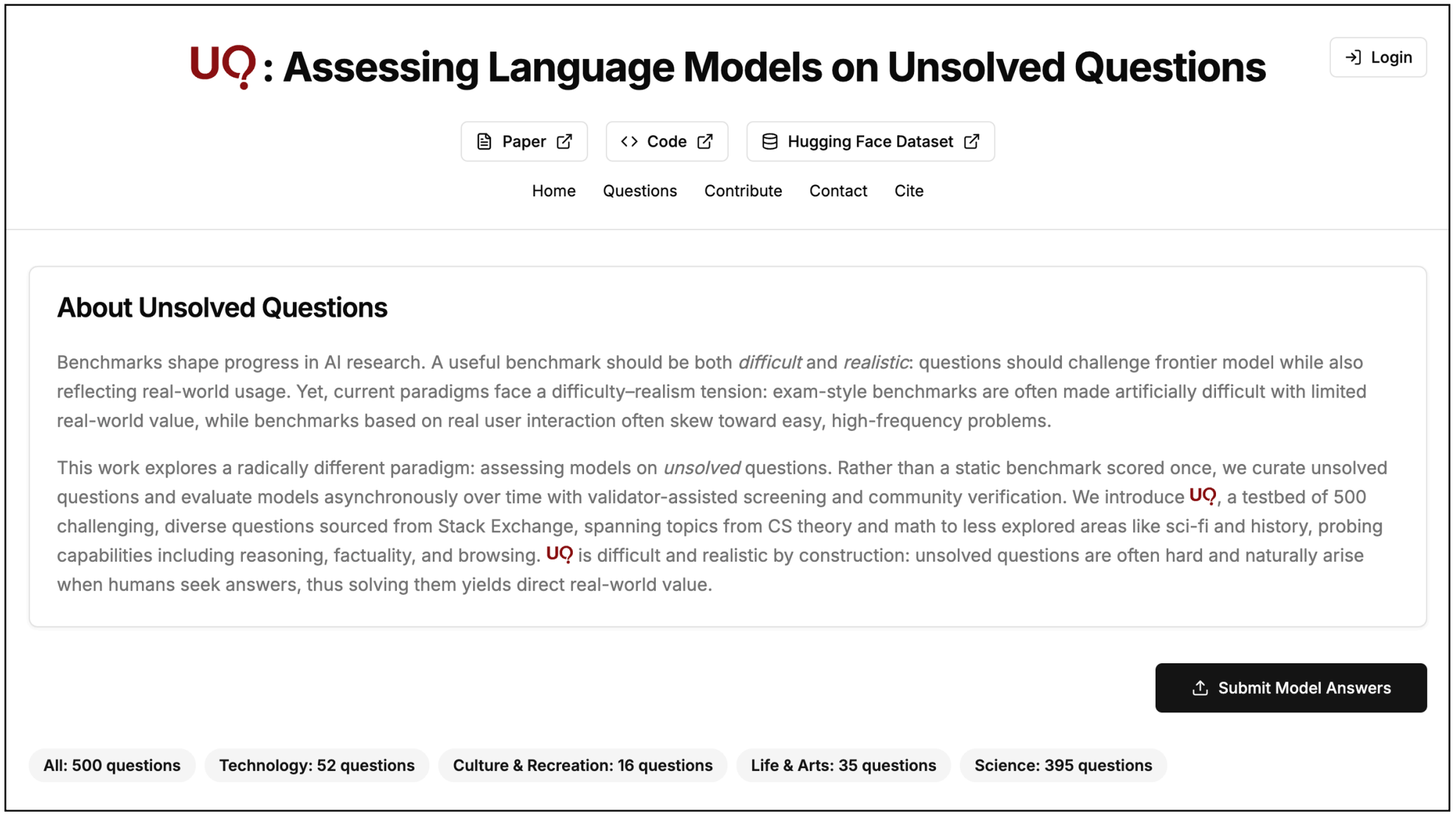


- Better answer performance is broadly indicative of better validation performance, but it's not always the case.

Oracle-free validator alone is not enough for unsolved questions.

Adding New Knowledge Requires Human Consensus!

UQ-Platform: An Open Platform for Community-Based Evaluation



Partial Model Evaluation

Answer Model	UQ-Validator Pass Rate		Human Pass Rate (2025-08-24)*
	# Passed	%	
o3-pro	75 / 500	15.0%	4 / 46
↪ on UQ diamond subset	3 / 25	4.0%	0 / 2
Gemini 2.5 Pro	25 / 500	5.0%	3 / 10
o4-mini (high)	25 / 500	5.0%	2 / 14
o3	44 / 500	8.8%	1 / 25
↪ on UQ diamond subset	1 / 25	2.0%	0 / 1
DeepSeek-R1-0528	11 / 500	2.2%	1 / 5
Claude Opus 4	7 / 500	1.4%	0 / 3
Claude Sonnet 3.7 (16K)	6 / 500	1.2%	0 / 3
GPT-4o	0 / 500	0.0%	0 / 0
Total unique questions	144 / 500	28.8%	10 / 91

Takeaways

- UQ introduces a new evaluation paradigm
  - Moving from create-once & run-once benchmarks to discovering new knowledge through continuous evaluation
- UQ has three components:
  - UQ-Dataset: 500 unsolved questions sourced from Stack Exchange
  - UQ-Validator: compound validation strategies to provide evaluation signals and pre-screen candidate solutions for human review
  - UQ-Platform: an open platform where experts collectively verify questions and solutions, enabling ongoing, asynchronous, and community-driven evaluation

Teams

UQ: Assessing Language Models on Unsolved Questions

Fan Nie[♠] * **Ken Ziyu Liu**[♠] *
Zihao Wang[♠] **Rui Sun**[♠] **Wei Liu**[♠]
Weijia Shi[♡] **Huaxiu Yao**[♣] **Linjun Zhang**[◇] **Andrew Y. Ng**[♠]
James Zou[♠] **Sanmi Koyejo**[♠] **Yejin Choi**[♠] **Percy Liang**[♠] **Niklas Muennighoff**[♠] *
[♠]Stanford University [♡]University of Washington [♣]UNC Chapel Hill [◇]Rutgers University

Ongoing next step

- One-sided automated verification for open problems
- Self-improvement on unsolved questions with weak signals
- ……

Questions