Humanity’s last exam

Woyi Wei
Matrik.Nr:13005880

Overview of Benchmarking in AI

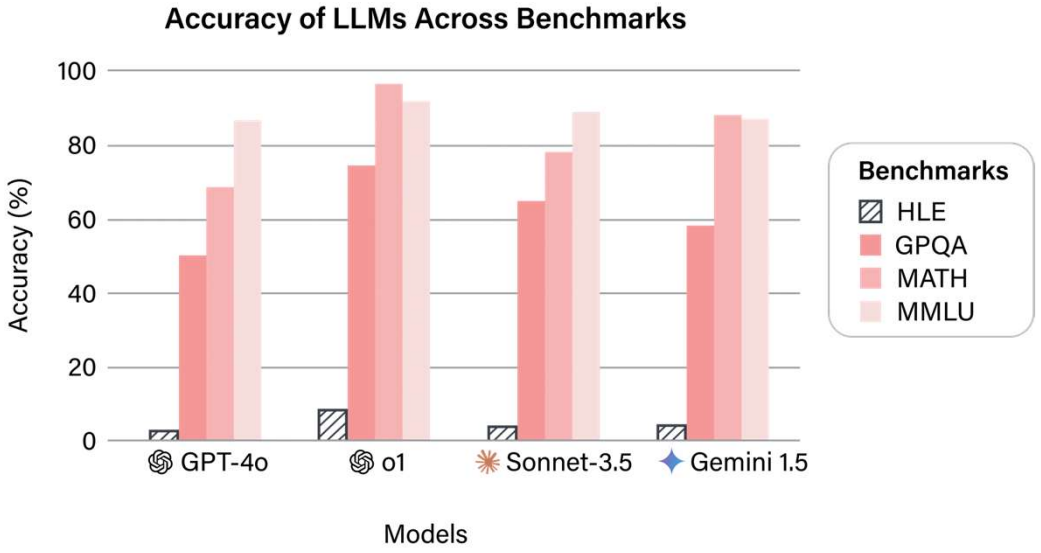
LLMs have demonstrated exceptional performance on standard benchmarks.

current benchmarks are reaching performance saturation, with many models scoring near-perfect.

This plateau limits our ability to assess the true capability of LLMs.

What is Benchmark Saturation?

- Benchmark saturation occurs when performance reaches near-perfect levels, leaving little room for further model improvement.
- In most cases, models achieve accuracy over 90% on common benchmarks like MMLU.



Humanity's Last Exam (HLE)

HLE was introduced to address the saturation problem by providing a more complex, multi-modal benchmark.

HLE consists of a community-contributed question bank that has been vetted.

Trivia

Question:

In Greek mythology, who was Jason's maternal great-grandfather?

Introduction

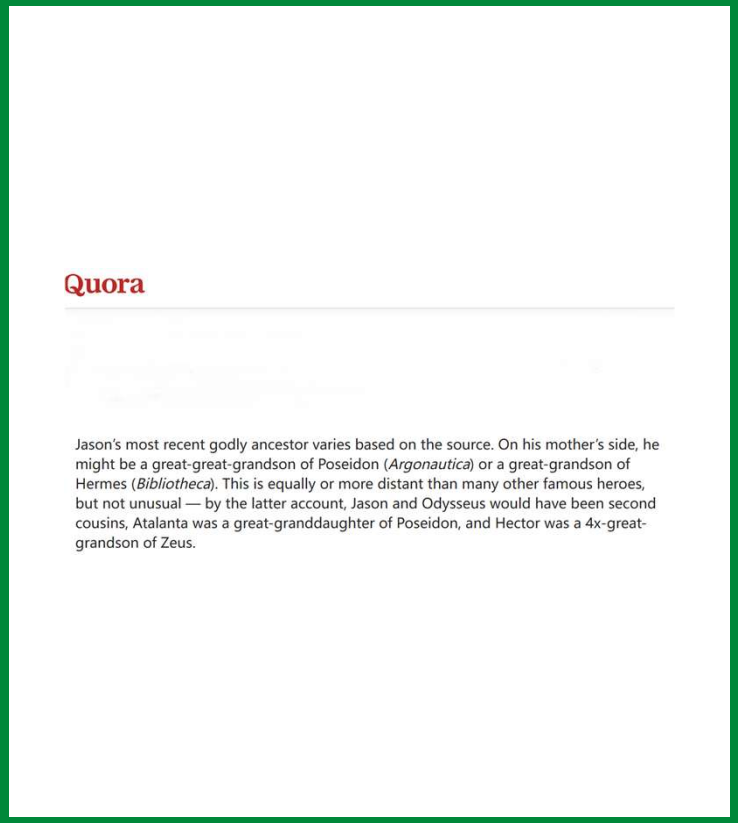
In Greek mythology, who was Jason's maternal great-grandfather?

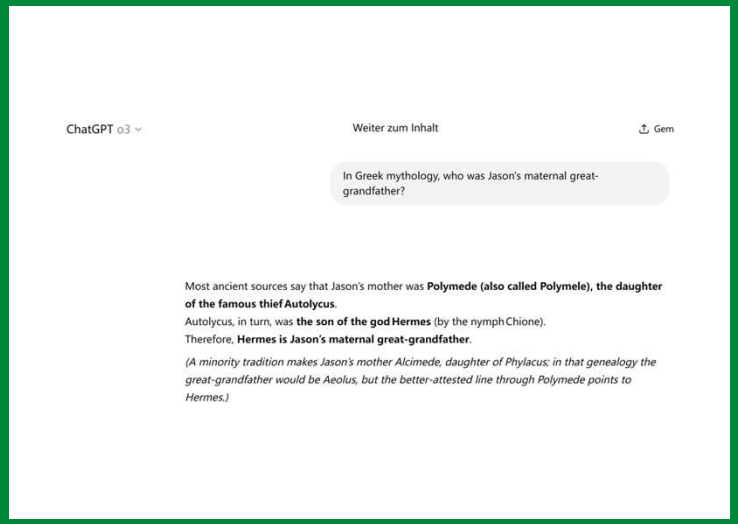
In Greek mythology, **Jason's maternal great-grandfather** was **Sisyphus**, the cunning king of Corinth who was famously punished in the underworld by being forced to eternally roll a boulder uphill.

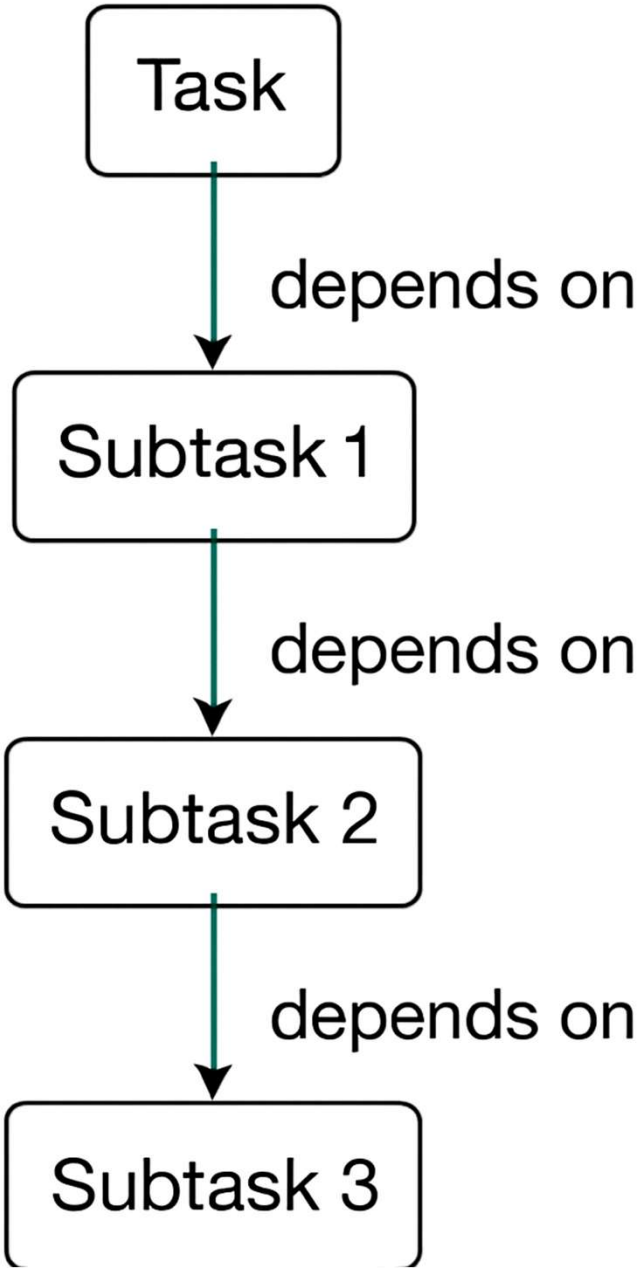
Here's the lineage breakdown:

- **Sisyphus** (great-grandfather)
  - └── **Athamas** (grandfather, son of Sisyphus and Merope)
  - └── **Ino** (mother, daughter of Athamas and Nephele)
  - └── **Jason** (son of Aeson, but raised by Ino after Aeson's overthrow by Pelias)

However, some versions suggest that **Ino** was Jason's aunt (Aeson's sister), not his mother, making **Athamas** his grandfather or great-uncle. But the most consistent connection places **Sisyphus** as Jason's maternal great-grandfather through Athamas and Ino.







LLMs often don't do well on tasks that require composition into smaller subtasks.

In this case there is a chain of relations that depend on the previous result.

Dataset Composition

A mix of question types: 76% exact-match and 24% multiple-choice (MCQ).

Ecology

Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Edward V
Massachusetts Institute of Technology

We call a class of graphs $\mathcal{G}$ well-behaved if, for each $G \in \mathcal{G}$ the Markov chain $M(G)$ converges to a unique stationary distribution, and the unique stationary distribution is the uniform distribution.

Which of the following graph classes is well-behaved?

Answer Choices:
A. The class of all non-bipartite regular graphs
B. The class of all connected cubic graphs
C. The class of all connected graphs
D. The class of all connected non-bipartite graphs
E. The class of all connected bipartite graphs.

Marc R
Queen Mary University of London

Evaluation Protocol

Zero-shot Chain-of-Thought (CoT) prompts were used to assess model performance.

The structure of prompts includes: Explanation, Answer, and Confidence levels.

O3-mini judge was employed to evaluate equivalence across different models and ensure consistency.

Zero-Shot Chain-of-Thought Guidance

- Encouraging Stepwise Thinking:

```
Explanation:
Step 1: Identify relevant facts...
Step 2: Apply formula...
Step 3: Simplify result...
```

Design of Dual Prompt Templates

- Multiple-Choice Questions

Your response should be in the following format:

Explanation: {your explanation for your answer choice}
Answer: {your chosen answer}
Confidence: {your confidence score between 0% and 100% for your answer}

- Exact-Match Questions

Your response should be in the following format:

Explanation: {your explanation for your final answer}
Exact Answer:{your succinct, final answer}
Confidence: {your confidence score between 0% and 100% for your answer}

Structured Judge Response

Structured Judge Response Example

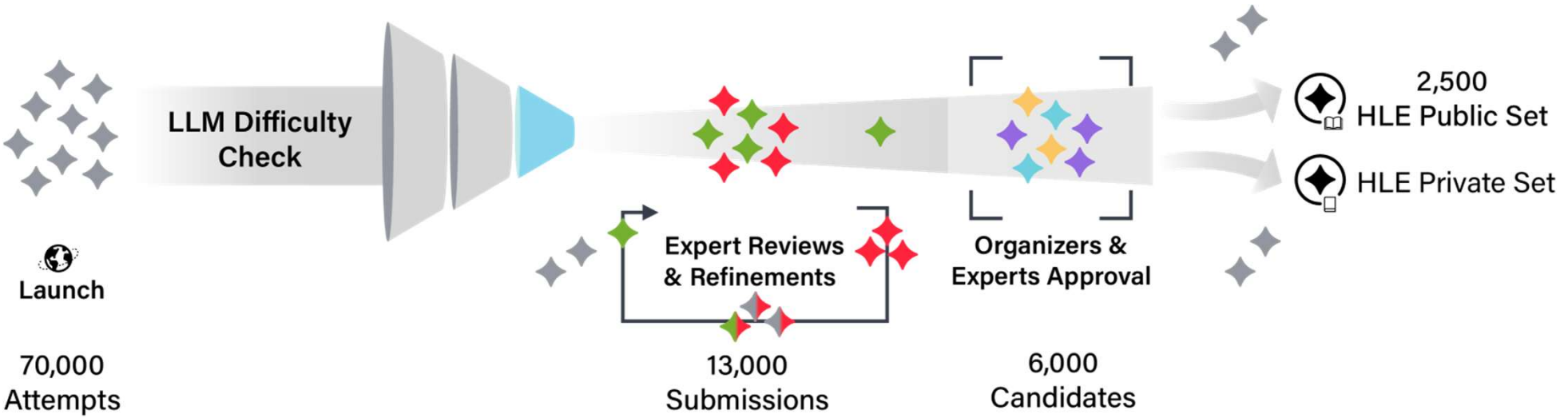
[correct_answer]: $\frac{\cos\left(\frac{\pi}{n}\right)}{2\left(1+\cos\left(\frac{\pi}{n}\right)\right)}$

[extracted_model_answer]: $\frac{\cot\left(\frac{\pi}{n}\right)}{2\cot\left(\frac{\pi}{2n}\right)}$

[reasoning]: Using the half-angle identity $\tan\left(\frac{\pi}{2n}\right) = \frac{\sin\left(\frac{\pi}{n}\right)}{1+\cos\left(\frac{\pi}{n}\right)}$ and recalling that $\cot x = \frac{\cos x}{\sin x}$, we express $\cot\left(\frac{\pi}{n}\right) = \frac{\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}$ and $\cot\left(\frac{\pi}{2n}\right) = \frac{1+\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}$. Forming the quotient $\frac{\cot\left(\frac{\pi}{n}\right)}{2\cot\left(\frac{\pi}{2n}\right)} = \frac{\frac{\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}}{2\cdot\frac{1+\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}} = \frac{\cos\left(\frac{\pi}{n}\right)}{2(1+\cos\left(\frac{\pi}{n}\right))}$, where the common $\sin\left(\frac{\pi}{n}\right)$ cancels out. This detailed inline simplification shows that the model answer is equivalent to the correct answer.

[correct]: yes

Gating & Filtering Process



Difficulty Check

Difficulty gating criteria:

- Exact-match questions require every model to answer incorrectly
- Multiple-choice questions permit at most one model to guess correctly by chance, eliminating “prop questions” prone to lucky hits.

Review and Refinements

- Questions should usually (but do not always need to) be at a graduate / PhD level or above. （For STEM）
- Questions should ask for something precise and have an objectively correct, univocal answer.
- Questions should be original and not derived from textbooks or Google.
- Questions need to be in English.

Score	Scoring Guideline	Description
0	Discard	The question is out of scope, not original, spam, or otherwise not good enough to be included in the HLE set and should be discarded.
1	Major Revisions Needed	Major revisions are needed for this question or the question is too easy and simple.
2	Some Revisions Needed	Difficulty and expertise required to answer the question is borderline. Some revisions are needed for this question.
3	Okay	The question is sufficiently challenging but the knowledge required is not graduate-level nor complex. Minor revisions may be needed for this question.
4	Great	The knowledge required is at the graduate level or the question is sufficiently challenging.
5	Top-Notch	Question is top-notch and perfect.
Unsure	-	Reviewer is unsure if the question fits the HLE guidelines, or unsure if the answer is right.

Expert Approval

- Recruitment of students from top United States universities to fully solve a sample of HLE questions. 、
- Errors flagged routed among organizers, original question authors, and auditors until consensus reached. Audit data used to further refine the dataset.

Expert Approval

Score	Scoring Guideline	Description
0	Discard	The question is out of scope, not original, spam, or otherwise not good enough to be included in the HLE set and should be discarded.
1	Not sure	Major revisions are needed for this question or you're just unsure about the question. Please put your thoughts in the comment box and an organizer will evaluate this.
2	Pending	You believe there are still minor revisions that are needed on this question. Please put your thoughts in the comment box and an organizer will evaluate this.
3	Easy questions models got wrong	These are very basic questions that models got correct or the question was easily found online. Any questions which are artificially difficult (large calculations needing a calculator, requires running/rendering code, etc.) should also belong in this category. The models we evaluate cannot access these tools, hence it creates an artificial difficulty bar. Important: "Found online" means via a simple search online. Research papers/journals/books are fine
4	Borderline	The question is not interesting OR The question is sufficiently challenging, but 1 or more of the models got the answer correct.
5	Okay to include in HLE benchmark	Very good questions (usually has score of 3 in the previous review round). You believe it should be included in the HLE Benchmark.
6	Top question in its category	Great question (usually has a score of 4-5 in the previous review round), at a graduate or research level. Please note that "graduate level" is less strict for Non-STEM questions. For Non-STEM questions and Trivia, they are fine as long as they are challenging and interesting.

Searchability & Post-Release Community Audit

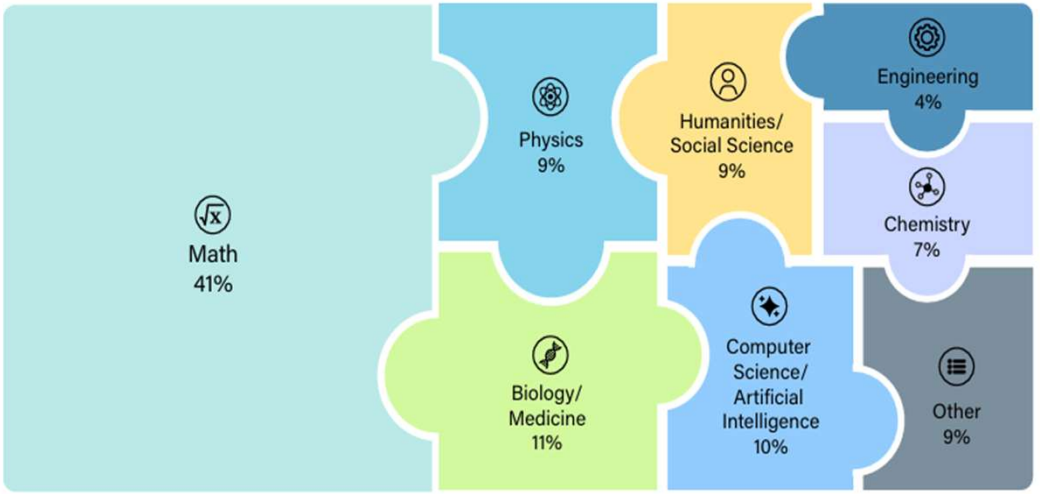
- Dual-model comparison: For questions that passed the first two rounds, examine whether retrieval-enabled models (e.g., GPT-4o search, Perplexity Sonar) answer correctly with search enabled but fail with search disabled. Questions showing “retrieval-enabled model correct + non-retrieval model incorrect” undergo manual review to confirm that a simple online query suffices; otherwise, those questions receive removal or revision.

- Community feedback: Upon public release, a “crowdsourced bug bounty” program opens for reports of label errors or question ambiguities; organizers and original authors jointly confirm and correct reported issues.

Motivating HLE

HLE introduces 2,500 questions spanning over 100 subjects

It challenges models by including multi-modal elements (text and images)



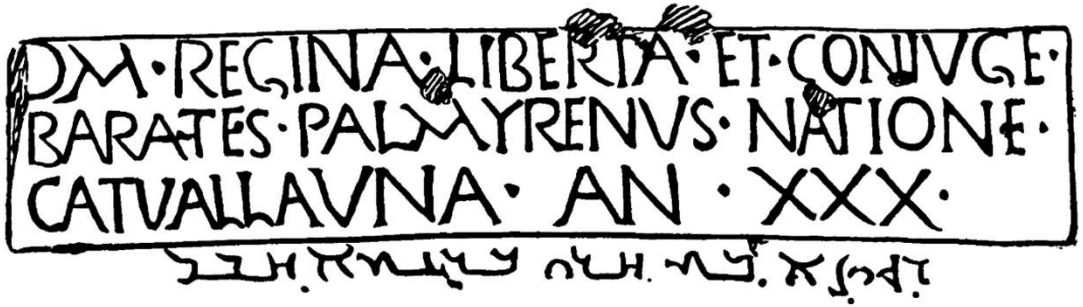
Multimodal Analysis

- Multimodal questions (Text + image): only cutting-edge, image-capable multimodal LLMs (e.g., GPT-4O, GEMINI 1.5 PRO, CLAUDE 3.5 SONNET, O1) for difficulty check.
- Text-only questions: supplementation with lightweight, non-vision models (O1-MINI, O1-PREVIEW) alongside multimodal LLMs to evaluate solvability under absence of visual cues.

Multimodal question example

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 °Tᵓ ḤBL

Henry T
Merton College, Oxford

Performance Across Domains

- **HLE results show significant variability in performance across different domains**

	Text-Only							
Model	Math	Bio/Med	Physics	CS/AI	Humanities	Chemistry	Engineering	Other
GPT-4o	2.3	5.0	1.5	0.9	2.6	2.0	1.6	2.3
Grok 2	3.2	5.4	4.5	3.6	1.0	1.0	4.8	1.1
Claude 3.5 Sonnet	3.8	5.9	4.5	2.2	6.7	5.0	9.7	2.9
Gemini 1.5 Pro	5.3	5.4	2.0	4.0	3.6	6.0	3.2	3.4
Gemini 2.0 Flash Thinking	8.1	7.7	4.5	4.9	6.2	5.0	4.8	2.9
o1	7.4	8.1	6.9	8.4	8.8	10.0	4.8	8.0
DeepSeek-R1	9.1	9.0	5.4	7.5	10.4	5.0	14.5	7.4
o3-mini (high)	18.6	10.0	15.3	8.4	5.2	9.0	6.5	6.9
	Full Dataset							
GPT-4o	2.3	6.4	1.7	0.8	3.2	3.6	1.8	2.6
Grok 2	3.0	4.6	3.9	3.3	1.4	2.4	3.6	1.7
Claude 3.5 Sonnet	4.0	4.6	3.9	2.5	5.9	4.2	7.2	2.2
Gemini 1.5 Pro	5.2	5.4	3.0	3.7	4.1	6.1	3.6	3.4
Gemini 2.0 Flash Thinking	8.0	8.2	4.8	4.5	6.4	5.5	6.3	3.0
o1	7.4	10.4	7.0	8.2	8.7	9.7	6.3	7.3

Insights from Domain Breakdown

- Understanding these gaps helps to inform the future development of domain-specific AI models.
- Targeted training and fine-tuning in weaker domains are needed to bridge these gaps.

The accuracy of LLMS in HLE

Judge Model: o3-mini | Dataset Updated: April 3rd, 2025

Model	Accuracy (%) ↑	Calibration Error (%) ↓
Gemini 2.5 Pro	21.6	72.0
o3	20.3	34.0
o4-mini	18.1	57.0
DeepSeek-R1-0528*	14.0	78.0
o3-mini*	13.4	80.0
Gemini 2.5 Flash	12.1	80.0
Qwen3-235B*	11.8	74.0
Claude 4 Opus	10.7	73.0
DeepSeek-R1*	8.5	73.0
Claude 3.7 Sonnet	8.0	80.0
o1	8.0	83.0
Claude 4 Sonnet	7.7	80.0
Llama 4 Maverick	5.7	83.0
GPT-4.5 Preview	5.4	85.0
GPT-4.1	5.4	89.0
Claude 3.5 Sonnet	4.1	84.0
GPT-4o	2.7	89.0

*Model is not multi-modal, evaluated on text-only subset.

Evaluation Metrics

Accuracy was measured as the main metric for model performance.

RMS calibration error was used to assess the reliability of the models' confidence in their responses.

Computation Pipeline

Metric	Formula	Interpretation
RMS Calibration Error	$\sqrt{\frac{1}{B}\sum_{b=1}^{B}(\mathrm{acc}_b - \mathrm{conf}_b)^2}$	Root-mean-square deviation between average confidence and accuracy across B bins
Expected Calibration Error	(\sum_{b=1}^B \frac{	S_b
Brier Score	$\frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$	Mean squared error between predicted probability f_i and true label $y_i \in \{0, 1\}$

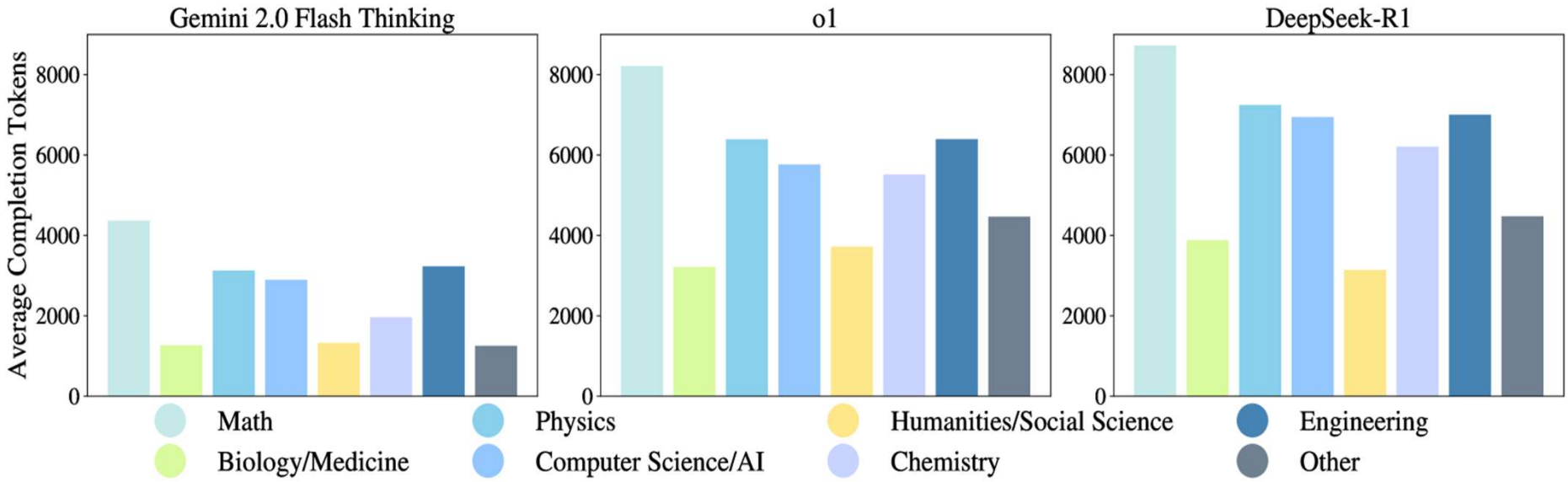
Token Usage in Reasoning Models

- Reasoning models like GPT-4o generate more than 8,000 tokens per question.

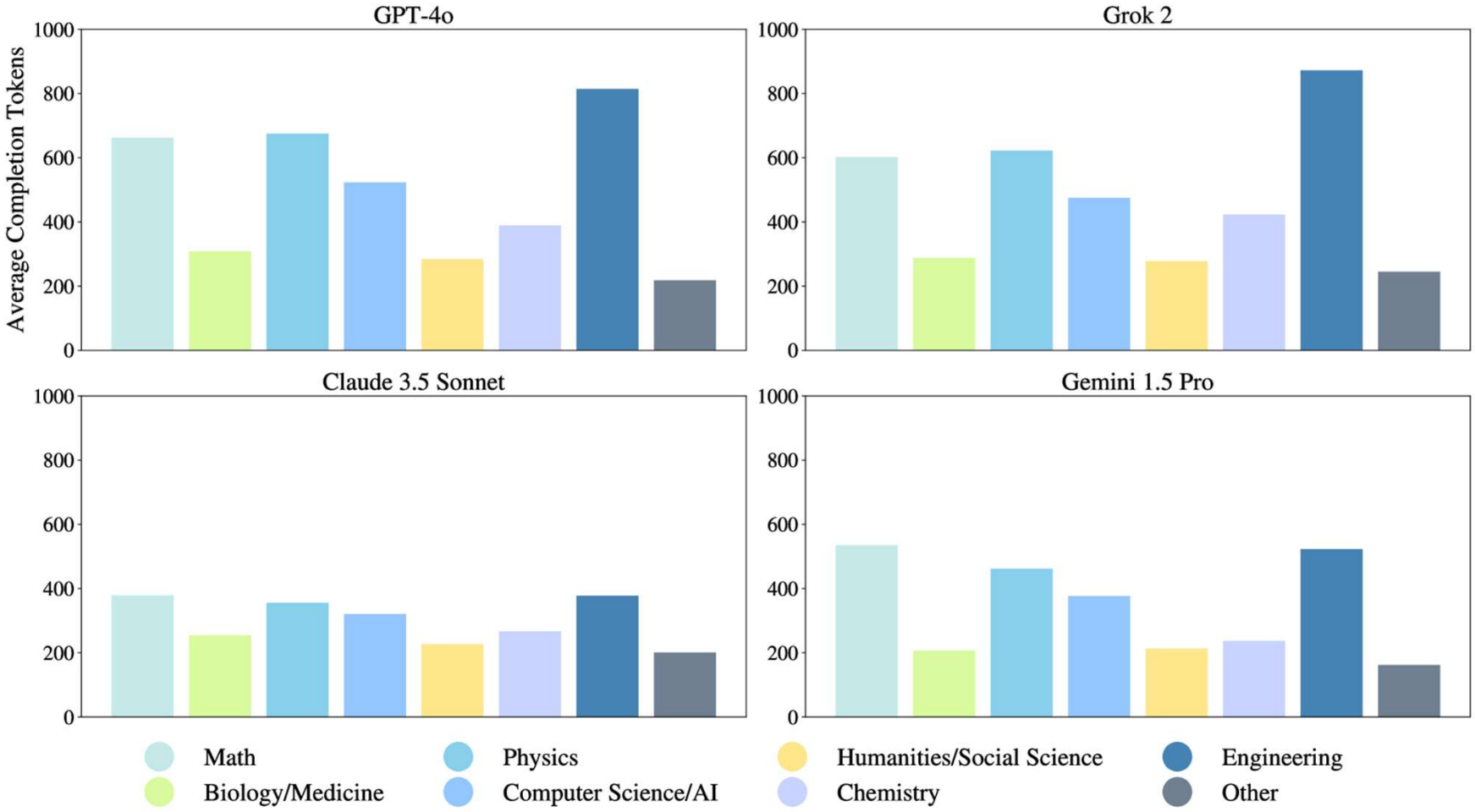
.

- Despite generating fewer tokens, non-reasoning models perform less accurately.

- The trade-off between token generation and accuracy is crucial in understanding LLM efficiency


Gemini 2.0 Flash Thinking
o1
DeepSeek-R1
Average Completion Tokens
0
2000
4000
6000
8000
Math
Physics
Humanities/Social Science
Engineering
Biology/Medicine
Computer Science/AI
Chemistry
Other

Reasoning



Non-Reasoning

Token Efficiency

Token Efficiency = Accuracy / Average Tokens

- High token usage limits the practicality of reasoning models in real-world applications (e.g., cost, computational time).

- In ultra-challenging tasks such as HLE, Token Efficiency of reasoning models falls below that of lightweight non-reasoning models, indicating that **mere extension of reasoning length does not yield proportional performance gains.**

Closed-Ended Task Limitations

- HLE primarily consists of closed-ended questions, limiting the scope for evaluating open-ended creativity or reasoning.
- This closed nature may not fully capture the range of a model's abilities.

Evolving Benchmarks

“There’s a big gulf between what it means to take an exam and what it means to be a practicing physicist and researcher. Even an A.I. that can answer these questions might not be ready to help in research, which is inherently less structured.”

Summary slides

Shows leading frontier LLMs score below 15% accuracy and suffer over 70% RMS calibration error on HLE, revealing major deficiencies in expert-level reasoning and confidence estimation

Establishes a rigorous, transparent framework for precisely measuring AI capabilities, empowering evidence-based tracking of model progress by researchers, journalists, and policymakers