The unknown performance of genAI models

Benchmarks as narrative devices

Stefan Baack
sbaack.com

Christo Buschek
christobuschek.net

Maty Bohacek
matybohacek.com

Starting observation: Model release announcements are often just lists of benchmark results. Example:

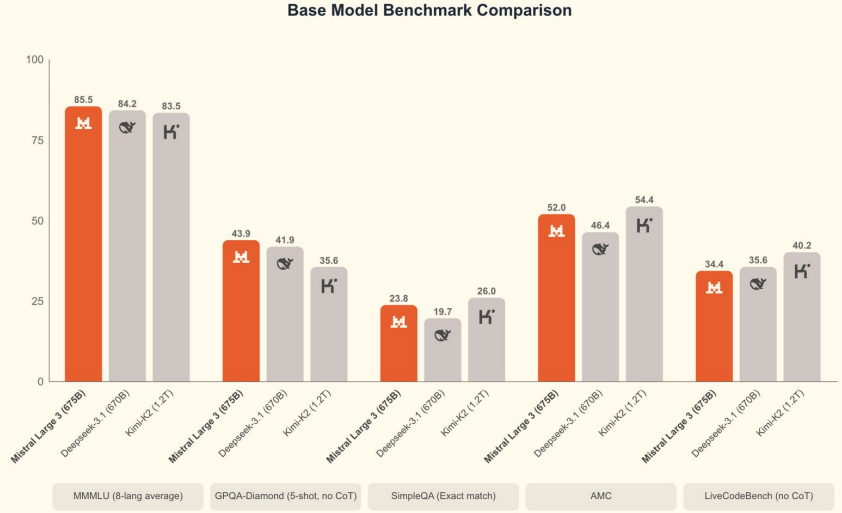
https://www.anthropic.com/claude-fable-and-mythos-5-1

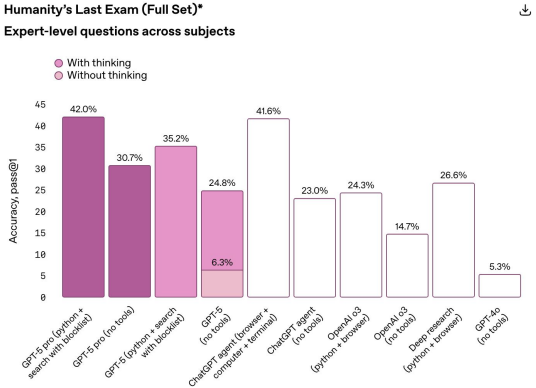
Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	No tools With search and code execution	**37.5%** **45.8%**	21.6% —	13.7% —	26.5% —
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	**31.1%**	4.9%	13.6%	17.6%
GPQA Diamond	Scientific knowledge	No tools	**91.9%**	86.4%	83.4%	88.1%
AIME 2025	Mathematics	No tools With code execution	**95.0%** **100%**	88.0% —	87.0% **100%**	94.0% —
MathArena Apex	Challenging Math Contest problems		**23.4%**	0.5%	1.6%	1.0%
MMMU-Pro	Multimodal understanding and reasoning		**81.0%**	68.0%	68.0%	76.0%
ScreenSpot-Pro	Screen understanding		**72.7%**	11.4%	36.2%	3.5%
CharXiv Reasoning	Information synthesis from complex charts		**81.4%**	69.6%	68.5%	69.5%
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	**0.115**	0.145	0.145	0.147
Video-MMMU	Knowledge acquisition from videos		**87.6%**	83.6%	77.8%	80.4%
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo Rating, higher is better	**2,439**	1,775	1,418	2,243
Terminal-Bench 2.0	Agentic terminal coding	Terminus-2 agent	**54.2%**	32.6%	42.8%	47.6%
SWE-Bench Verified	Agentic coding	Single attempt	76.2%	59.6%	**77.2%**	76.3%
τ2-bench	Agentic tool use		**85.4%**	54.9%	84.7%	80.2%
Vending-Bench 2	Long-horizon agentic tasks	Net worth (mean), higher is better	**$5,478.16**	$573.64	$3,838.74	$1,473.43
FACTS Benchmark Suite	Held out internal grounding, parametric, MM, and search retrieval benchmarks		**70.5%**	63.4%	50.4%	50.8%
SimpleQA Verified	Parametric knowledge		**72.1%**	54.5%	29.3%	34.9%
MMMLU	Multilingual Q&A		**91.8%**	89.5%	89.1%	91.0%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		**93.4%**	91.5%	90.1%	90.9%
MRCR v2 (8-needle)	Long context performance	128k (average) 1M (pointwise)	**77.0%** **26.3%**	58.0% 16.4%	47.1% not supported	61.6% not supported

For details on our evaluation methodology please see deepmind.google/models/evals-methodology/gemini-3-pro

Gemini 3 is state-of-the-art across a range of key AI benchmarks. See details on our evaluation methodology.

Mistral Large 3: A state-of-the-art open model





OpenAI quietly boosts some of Astra's evaluation metrics, and continues to change others post-launch

By **Emily Forlini**
Senior AI Reporter

September 4, 2026, 8:12 PM ET

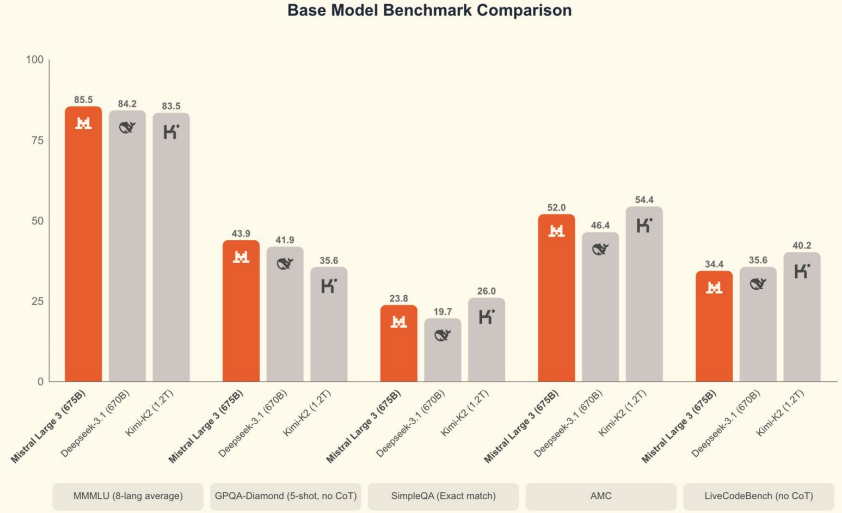
Source: https://fortune.com/2026/09/04/openai-quietly-boosts-some-of-astras-evaluation-metrics-amid-rare-delay-in-publication-of-the-modeblog-post-announcement/

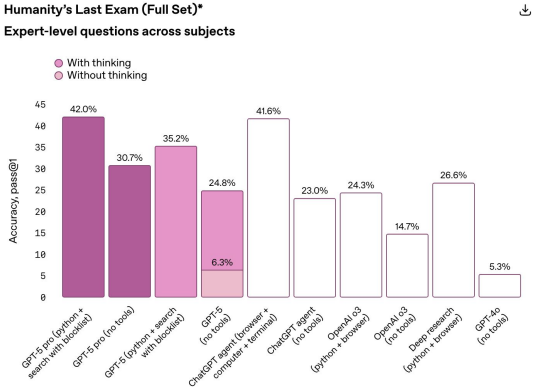
Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	No tools With search and code execution	**37.5%** **45.8%**	21.6% —	13.7% —	26.5% —
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	**31.1%**	4.9%	13.6%	17.6%
GPQA Diamond	Scientific knowledge	No tools	**91.9%**	86.4%	83.4%	88.1%
AIME 2025	Mathematics	No tools With code execution	**95.0%** **100%**	88.0% —	87.0% **100%**	94.0% —
MathArena Apex	Challenging Math Contest problems		**23.4%**	0.5%	1.6%	1.0%
MMMU-Pro	Multimodal understanding and reasoning		**81.0%**	68.0%	68.0%	76.0%
ScreenSpot-Pro	Screen understanding		**72.7%**	11.4%	36.2%	3.5%
CharXiv Reasoning	Information synthesis from complex charts		**81.4%**	69.6%	68.5%	69.5%
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	**0.115**	0.145	0.145	0.147
Video-MMMU	Knowledge acquisition from videos		**87.6%**	83.6%	77.8%	80.4%
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo Rating, higher is better	**2,439**	1,775	1,418	2,243
Terminal-Bench 2.0	Agentic terminal coding	Terminus-2 agent	**54.2%**	32.6%	42.8%	47.6%
SWE-Bench Verified	Agentic coding	Single attempt	76.2%	59.6%	**77.2%**	76.3%
τ2-bench	Agentic tool use		**85.4%**	54.9%	84.7%	80.2%
Vending-Bench 2	Long-horizon agentic tasks	Net worth (mean), higher is better	**$5,478.16**	$573.64	$3,838.74	$1,473.43
FACTS Benchmark Suite	Held out internal grounding, parametric, MM, and search retrieval benchmarks		**70.5%**	63.4%	50.4%	50.8%
SimpleQA Verified	Parametric knowledge		**72.1%**	54.5%	29.3%	34.9%
MMMLU	Multilingual Q&A		**91.8%**	89.5%	89.1%	91.0%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		**93.4%**	91.5%	90.1%	90.9%
MRCR v2 (8-needle)	Long context performance	128k (average) 1M (pointwise)	**77.0%** **26.3%**	58.0% 16.4%	47.1% not supported	61.6% not supported

For details on our evaluation methodology please see deepmind.google/models/evals-methodology/gemini-3-pro

Gemini 3 is state-of-the-art across a range of key AI benchmarks. See details on our evaluation methodology.

Mistral Large 3: A state-of-the-art open model





Goal: Create a resource to monitor how AI model builders advertise their models through benchmarks to study what these models are optimized for.

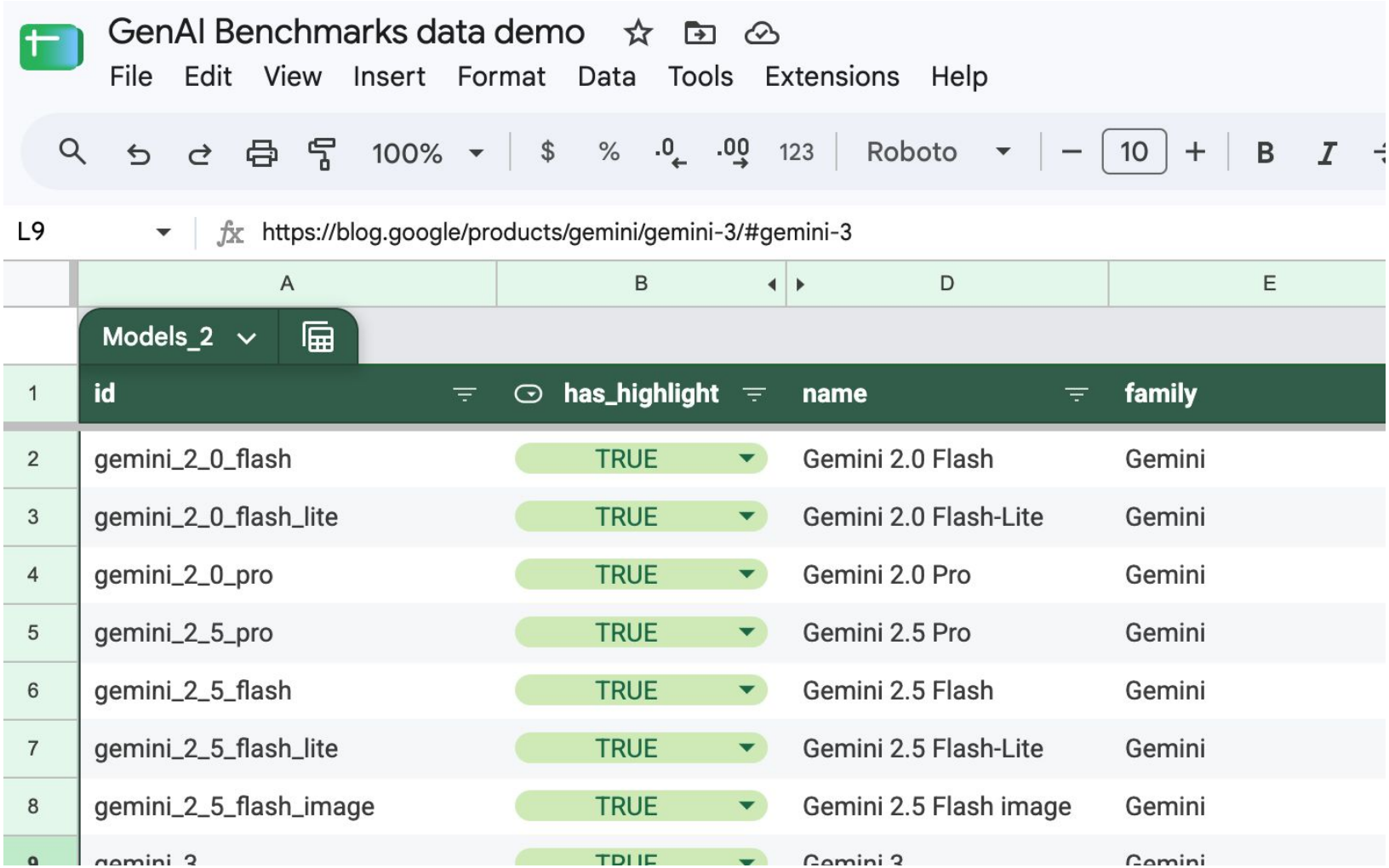
This can form a basis for better understanding the implications of using those models to do various types of work

Our project

Analyzed 231 benchmarks mentioned by 11 prominent model builders across 139 model release announcements throughout 2025

Selected model builders:

Google, OpenAI, Anthropic, Meta, xAI, Alibaba, Baidu, DeepSeek, Mistral, Allen Institute for AI, and Z.ai



	id	has_highlight	name	family
2	gemini_2_0_flash	TRUE	Gemini 2.0 Flash	Gemini
3	gemini_2_0_flash_lite	TRUE	Gemini 2.0 Flash-Lite	Gemini
4	gemini_2_0_pro	TRUE	Gemini 2.0 Pro	Gemini
5	gemini_2_5_pro	TRUE	Gemini 2.5 Pro	Gemini
6	gemini_2_5_flash	TRUE	Gemini 2.5 Flash	Gemini
7	gemini_2_5_flash_lite	TRUE	Gemini 2.5 Flash-Lite	Gemini
8	gemini_2_5_flash_image	TRUE	Gemini 2.5 Flash image	Gemini
9	gemini_3	TRUE	Gemini 3	Gemini

Our project

- Collected benchmark author affiliations using arXiv.org papers
- Developed own benchmark taxonomy based on what benchmark authors claim to evaluate for better cross-company analysis
- In-depth analysis of some of the most popular benchmarks among AI model builders

The paper

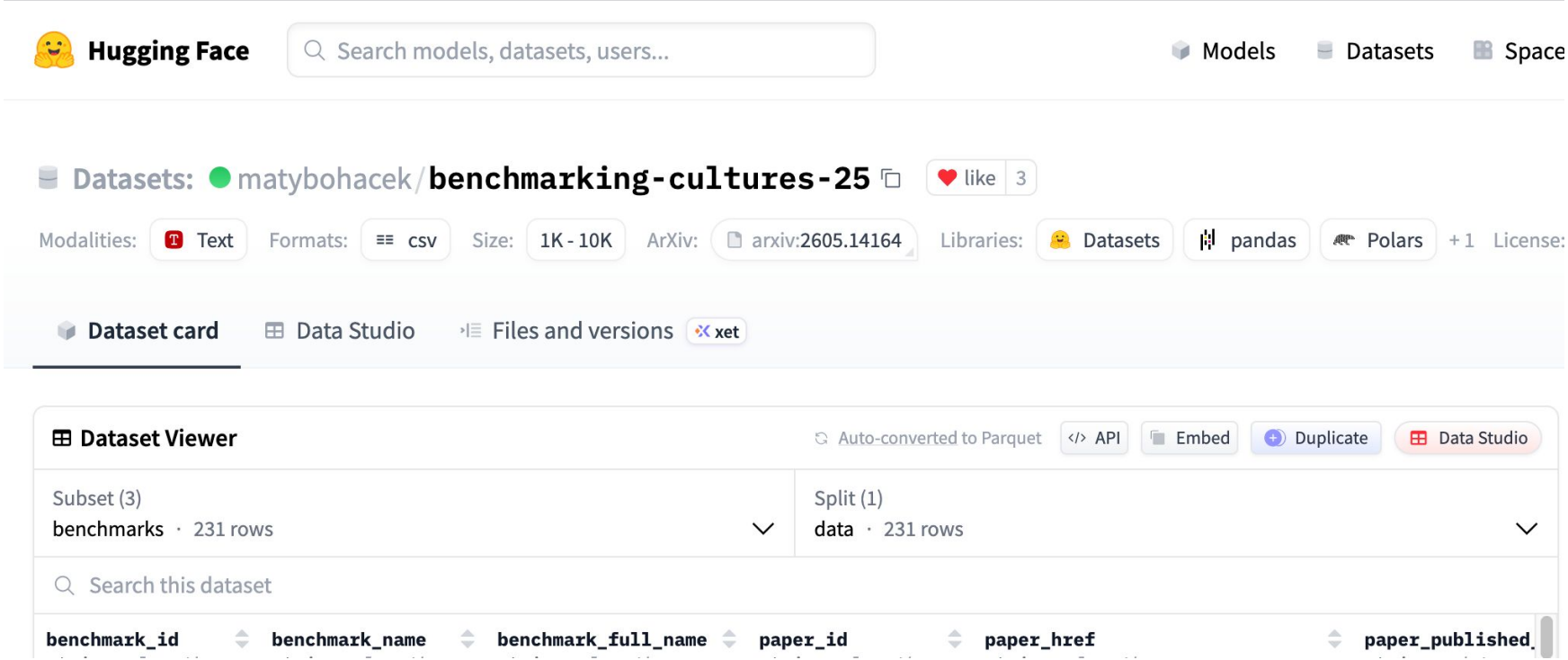
Unsteady Metrics and Benchmarking Cultures of AI Model Builders

Stefan Baack, Christo Buschek, Maty Bohacek

The primary way to establish and compare competencies in foundation and generative AI models has shifted from peer-reviewed literature to press releases and company blog posts, where model builders highlight results on selected benchmarks. These artifacts now largely define the state of the art for researchers and the public. Despite their prominence, which benchmarks model builders choose to highlight, and what they communicate through this selection, is underexamined. To investigate, we introduce and open-source Benchmarking-Cultures-25, a dataset of 231 benchmarks highlighted across 139 model releases in 2025 from 11 major AI builders, alongside an interactive tool to explore the data. Our analysis reveals a fragmented evaluation landscape with limited cross-model comparability: 63.2% of highlighted benchmarks are used by a single builder, and 38.5% appear in just one release. Few achieve widespread use (e.g., GPQA Diamond, LiveCodeBench, AIME 2025). Moreover, benchmarks are attributed different competencies by different builders, depending on their narrative. To disentangle these conflicting presentations, we develop a unified taxonomy mapping diverging terminology to a shared framework of measured signals based on what benchmark authors claim to measure. "General knowledge application" is the second most popular, yet vaguely defined, category. Qualitative analysis shows many such benchmarks deemphasize construct validity, instead framing results as indicators of progress toward AGI. Their authors claim to measure knowledge or reasoning broadly, yet mostly evaluate STEM subjects (especially math). We argue that highlighted benchmarks function less as standardized measurement tools and more as flexible narrative devices prioritizing market positioning over scientific evaluation. Data: this https URL tool: this https URL.

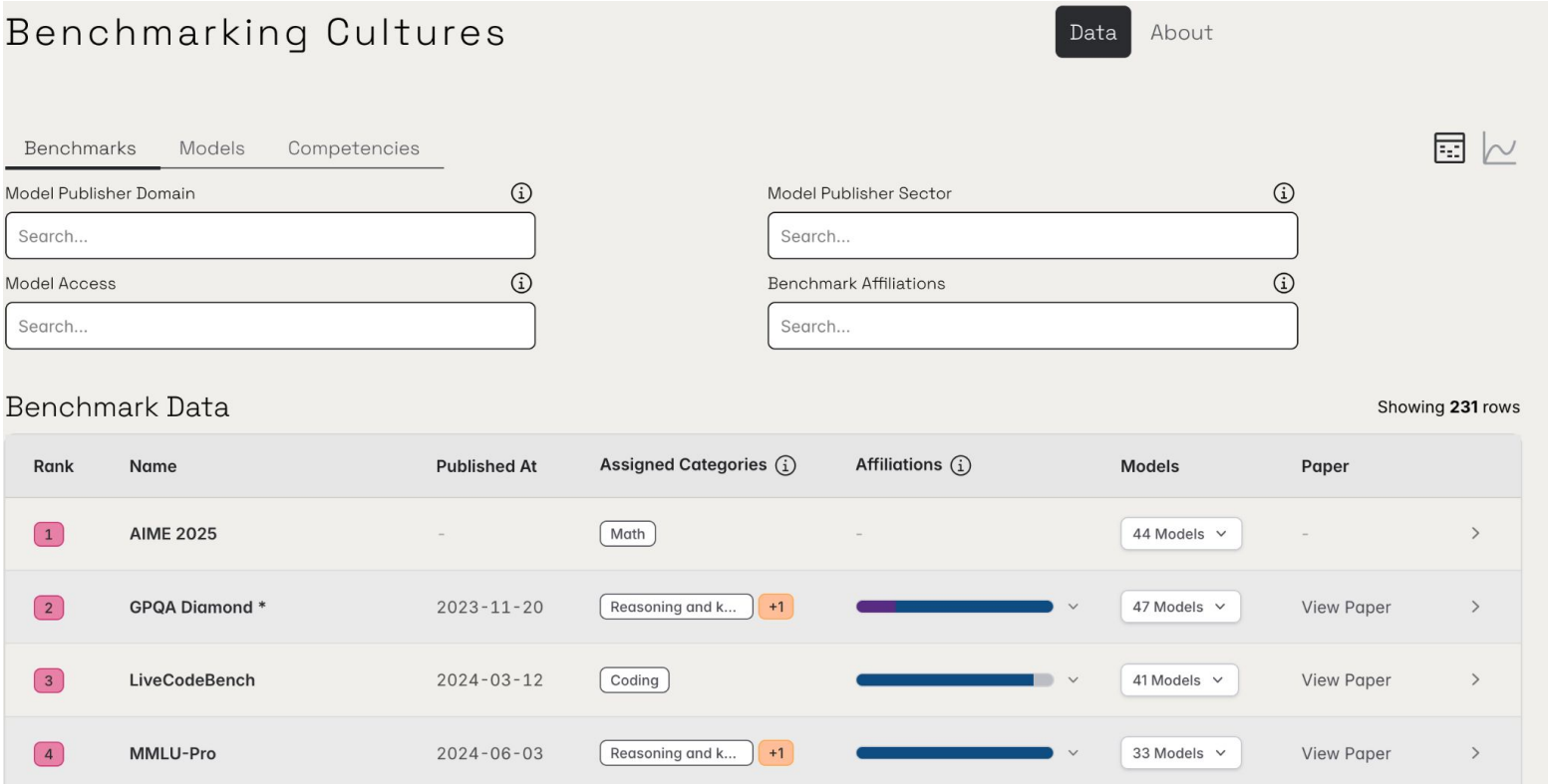
https://arxiv.org/abs/2605.14164

The data



https://huggingface.co/datasets/matybohacek/benchmarking-cultures-25

The tool



bench-cultures.net

Some findings from our analysis

Findings: Fragmentation

63% **of benchmarks** used only by a single model builder

39% **of benchmarks** used only by a single model release

The way benchmarks are highlighted in model release announcements provides next to no cross-model comparability

Findings: Inconsistent framing of what benchmarks evaluate

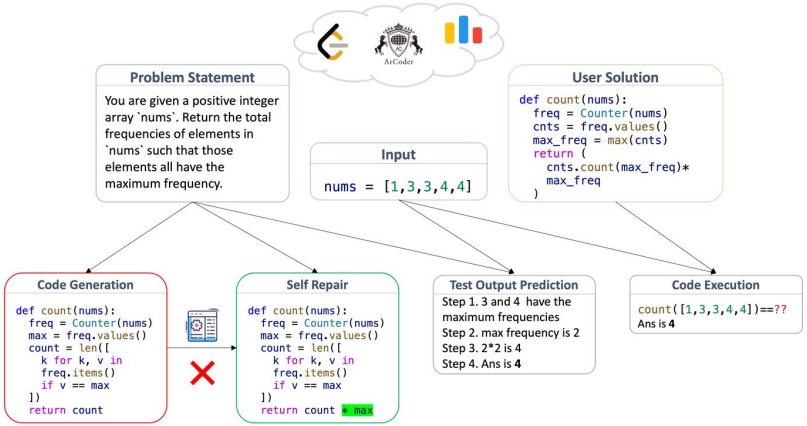
Benchmark	Description		Gemini 3 Pro
Humanity's Last Exam	Academic reasoning	No tools	37.5%
		With search and code execution	45.8%
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	31.1%
GPQA Diamond	Scientific knowledge	No tools	91.9%
AIME 2025	Mathematics	No tools	95.0%
		With code execution	100%

Source: https://blog.google/products-and-platforms/products/gemini/gemini-3/#gemini-3

LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code

Naman Jain[1], King Han[1], Alex Gu[2], Wen-Ding Li[3], Fanjia Yan[1], Tianjun Zhang[1], Sida Wang,
Armando Solar-Lezama[2], Koushik Sen[1], Ion Stoica[1]
[1]UC Berkeley [2]MIT [3]Cornell University

Paper | Code | Data | Leaderboard



One of the most popular **coding** benchmarks has been labelled by AI model builders as:

- Reasoning benchmark (7)
- Agentic benchmark (3)
- “Cost-efficient intelligence” or other labels (3)
- General (1)
- Uncategorized (5)

Findings: Industry dominance

Benchmark author affiliation type	**Overall %**	**2025 only %**
Industry	43.9	49.3
Academia	39.0	33.0
Non-profit	10.1	6.8
Other/Unspec.	5.7	9.6
Independent	0.7	1.3
Government	0.4	0.1

Top 15 most popular benchmarks

Rank	Benchmark
1	AIME 2025
2	GPQA Diamond
3	LiveCodeBench
4	MMLU-Pro
5	AIME 2024
6	MMMU
7	SWE-bench Verified
8	SimpleQA
9	Humanity’s Last Exam
10	MMLU
11	MATH
12	MMMLU
13	Aider Polyglot
14	IF-Eval
15	HMMT 2025

Findings: General knowledge application benchmarks

General knowledge application benchmarks seek to measure knowledge or reasoning capabilities *in general*, across a broad spectrum of subjects

Humanity's Last Exam

Paper

Nature | Arxiv

Dataset

```
load_dataset("cais/hle")
```

GitHub

https://agi.safe.ai/

Humanity's Last Exam

Paper

Nature | Arxiv

Dataset

```
load_dataset("cais/hle")
```

GitHub

Goal: Measuring performance “at the frontier of human knowledge” with “broad subject coverage” (Phan et al. 2025).

Example: Humanity’s Last Exam

In practice, this means: Attracting experts (professors, researchers, graduate degree holders) to contribute and review questions included in the benchmark.

- $5000 USD for each of the top 50 questions, $500 USD for each of the next 500 questions
- Co-authorship on the final paper for all

Example: Humanity’s Last Exam

- Questions that could be answered correctly by current “state-of-the-art” models were rejected outright
- After passing this test, questions were reviewed by “multiple graduate-level reviewers” (Phan et al. 2025)

Example: Humanity’s Last Exam

Example question:

“Suppose two people played a game of chess with the aim of placing as many bishops on the edge squares of the board as possible. If they succeeded in doing so, how many edge squares would lack bishops?”

Source: https://huggingface.co/datasets/cais/hle

Example: Humanity’s Last Exam

Measuring the “frontier of human knowledge”?

Measuring how well future models can answer questions from subject-matter experts that current models can’t answer correctly.

Findings: General knowledge application benchmarks

Top 5 General knowledge application benchmarks:

GPQA Diamond	Graduate-Level Google-Proof Q&A
MMMU	Massive Multi-discipline Multimodal Understanding and Reasoning
MMLU-Pro	Massive Multitask Language Understanding Pro
MMLU	Massive Multitask Language Understanding
Humanity’s Last Exam	–

Findings: General knowledge application benchmarks

None of the benchmarks clearly define “knowledge” or “reasoning”

Implicitly, knowledge is what’s in the training data or “searchable,” reasoning refers to logical inference and tasks with “non-searchable” solutions

Findings: General knowledge application benchmarks

Field	Share (%)	Share (N)
Science	30.4	12,850
Humanities & Social Sciences	18.8	7,920
Business	12.1	5,095
Tech & Engineering	11.6	4,877
Health & Medicine	11.5	4,872
Law	7.3	3,078
Other/Unspecified	5.2	2,187
Art & Design	3.1	1,329

Findings: General knowledge application benchmarks

MMMU (Yue et al. 2023) and MMLU-Pro (Wang et al. 2024) directly cite AGI literature to define what they're evaluating: The paper "Levels of AGI for Operationalizing Progress on the Path to AGI" by Morris et al. 2024.

Findings: General knowledge application benchmarks

Level 0: No AI	
Level 1: Emerging	“equal to or somewhat better than an unskilled human”
Level 2: Competent	“at least 50th percentile of skilled adults”
Level 3: Expert	“at least 90th percentile of skilled adults”
Level 4: Exceptional	“at least 99th percentile of skilled adults”
Level 5: Superhuman	“outperforms 100% of humans”

Source: Morris et al. 2024

Findings: General knowledge application benchmarks

Level 0: No AI	
Level 1: Emerging	“equal to or somewhat better than an unskilled human”
Level 2: Competent	“at least 50th percentile of skilled adults”
Level 3: Expert	**“at least 90th percentile of skilled adults”**
Level 4: Exceptional	“at least 99th percentile of skilled adults”
Level 5: Superhuman	“outperforms 100% of humans”

Source: Morris et al. 2024

Findings: General knowledge application benchmarks

GPQA Diamond and Humanity’s Last Exam don’t cite AGI literature but are clearly informed by AGI discourses

"narrowly superhuman AI systems…to advance the frontier of human knowledge" (Rein et al. 2023)

"the final closed-ended academic benchmark of its kind” (Phan et al. 2025)

Findings: General knowledge application benchmarks

Top 5 General knowledge application benchmarks:

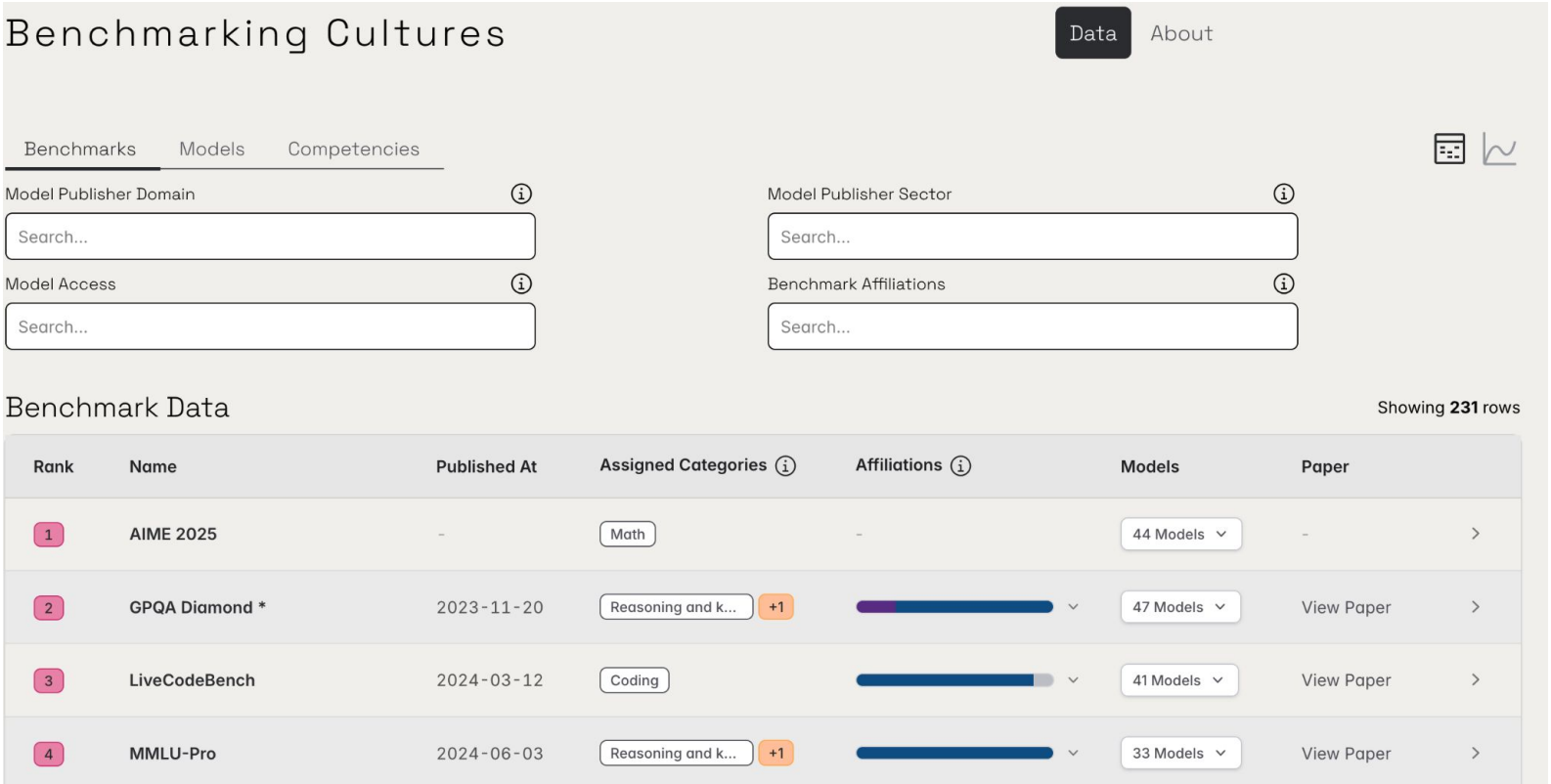
GPQA Diamond	Graduate-Level Google-Proof Q&A
MMMU	Massive Multi-discipline Multimodal Understanding and Reasoning
MMLU-Pro	Massive Multitask Language Understanding Pro
MMLU	Massive Multitask Language Understanding
Humanity's Last Exam	–

Findings: General knowledge application benchmarks

What are AI model builders implying when highlighting AGI benchmarks?

1. Generality (the model is good for almost everything)
2. The model's increased potential to replace human labor

Demo time!



bench-cultures.net

Conclusion: What is there to learn?

Don’t just question the numbers, question what is (not) being measured.

- What benchmarks are highlighted?
- What do they claim to measure and how do they measure it?
- What is missing?

Conclusion: What is there to learn?

Investigate how popular benchmarks are being made.

- Crowdworker involvement?
- Biases?
- Known errors?

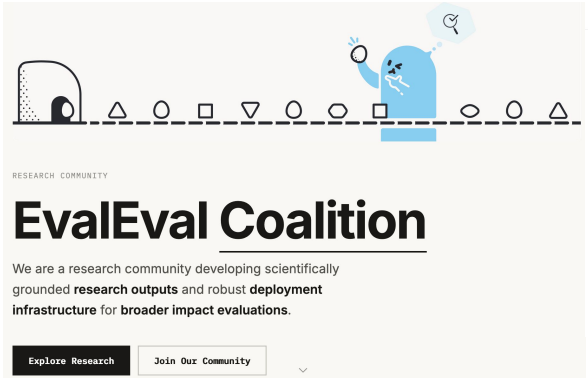
Conclusion: What is there to learn?

MMLU as a benchmark. For example, we find that 57% of the analysed instances in the *Virology* subset contain errors, including the suggestion to send the American army to West Africa to prevent outbreaks of Ebola (see Fig. 1).

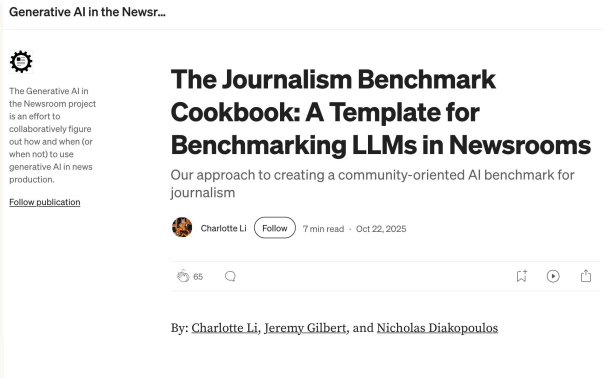
Gema et al. 2025

Conclusion: What is there to learn?

AI model builders try to shape a narrative with benchmarks. Seek out independently developed benchmarks to keep narratives in check.



https://evalevalai.com/



https://generative-ai-newsroom.com/

Thank you!

Stefan Baack
sbaack.com

Christo Buschek
christobuschek.net

Maty Bohacek
matybohacek.com

References

Baack, Stefan, Christo Buschek, and Maty Bohacek. 2026. “Unsteady Metrics and Benchmarking Cultures of AI Model Builders.” arXiv:2605.14164. Preprint, arXiv, May 13. https://doi.org/10.48550/arXiv.2605.14164.

Forlini, Emily. 2026. “OpenAI Quietly Boosts Some of Astra’s Evaluation Metrics, and Continues to Change Others Post-Launch.” Fortune, September 4. https://fortune.com/2026/09/04/openai-quietly-boosts-some-of-astras-evaluation-metrics-amid-rare-delay-in-publication-of-the-modeblog-post-announcement/.

Gema, Aryo Pradipta, Joshua Ong Jun Leang, Giwon Hong, et al. 2025. “Are We Done with MMLU?” arXiv:2406.04127. Preprint, arXiv, January 10. https://doi.org/10.48550/arXiv.2406.04127.

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, et al. 2024. “Levels of AGI for Operationalizing Progress on the Path to AGI.” Version 5. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2311.02462.

Phan, Long, Alice Gatti, Ziwen Han, et al. 2025. “Humanity’s Last Exam.” Version 10. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2501.14249.

Rein, David, Betty Li Hou, Asa Cooper Stickland, et al. 2023. “GPQA: A Graduate-Level Google-Proof Q&A Benchmark.” Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2311.12022.