Tatiana Shavrina
Meta SuperIntelligence Labs

LLM Agents and Evaluation

A brief overview of LLM evaluation methods



That’s me



Tatiana Shavrina

Research Scientist Manager

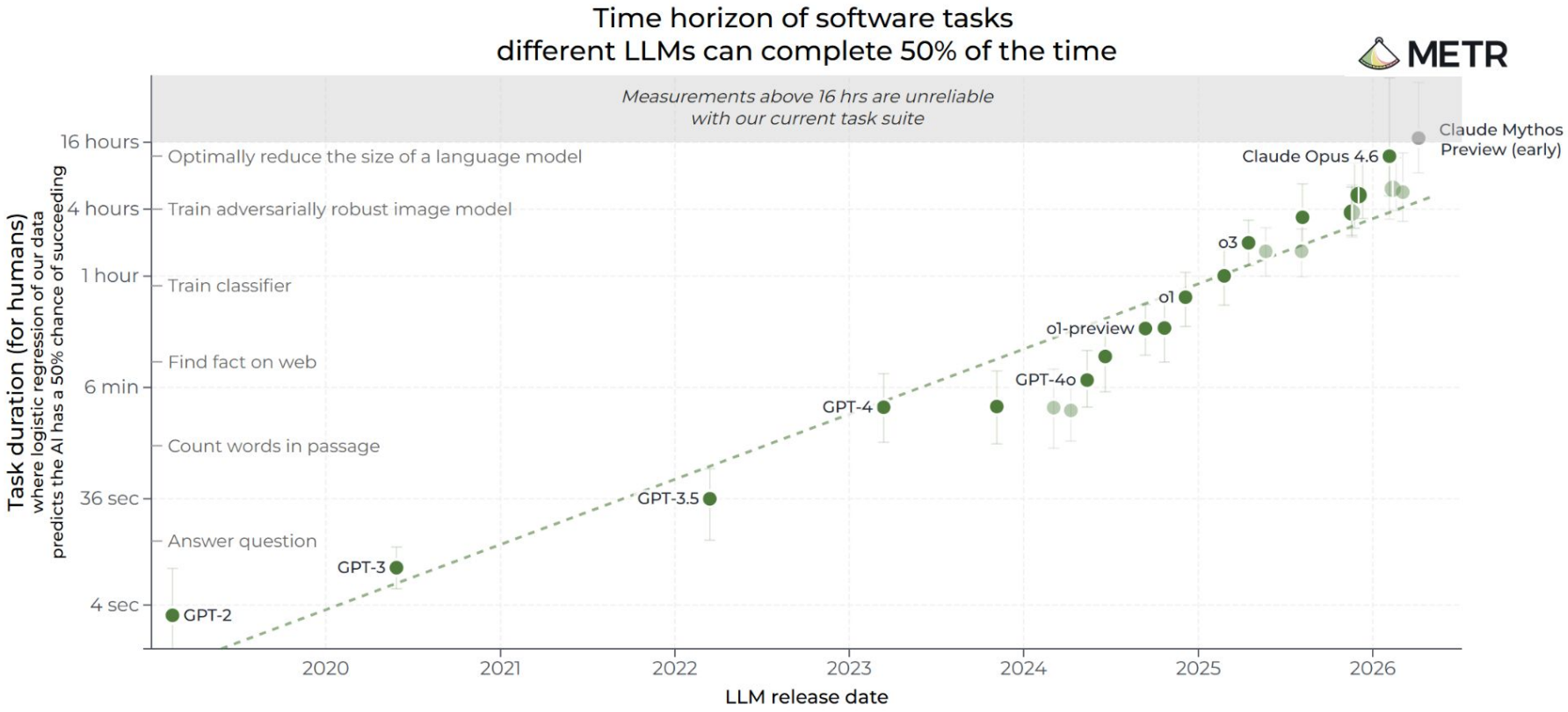
AI Research Agents, FAIR

- → **AIRA project:**
  - **AIRA dojo scaffold**
  - **AIRSbench**
  - **MLGym**
  - **NanoGPT Speedrun**
  - **MLEbench SOTA**
- → **LLama 3.1+, BLOOM**

Oxford LLMs

How do we evaluate LLMs and Agents?

Agents are solving more time-consuming tasks

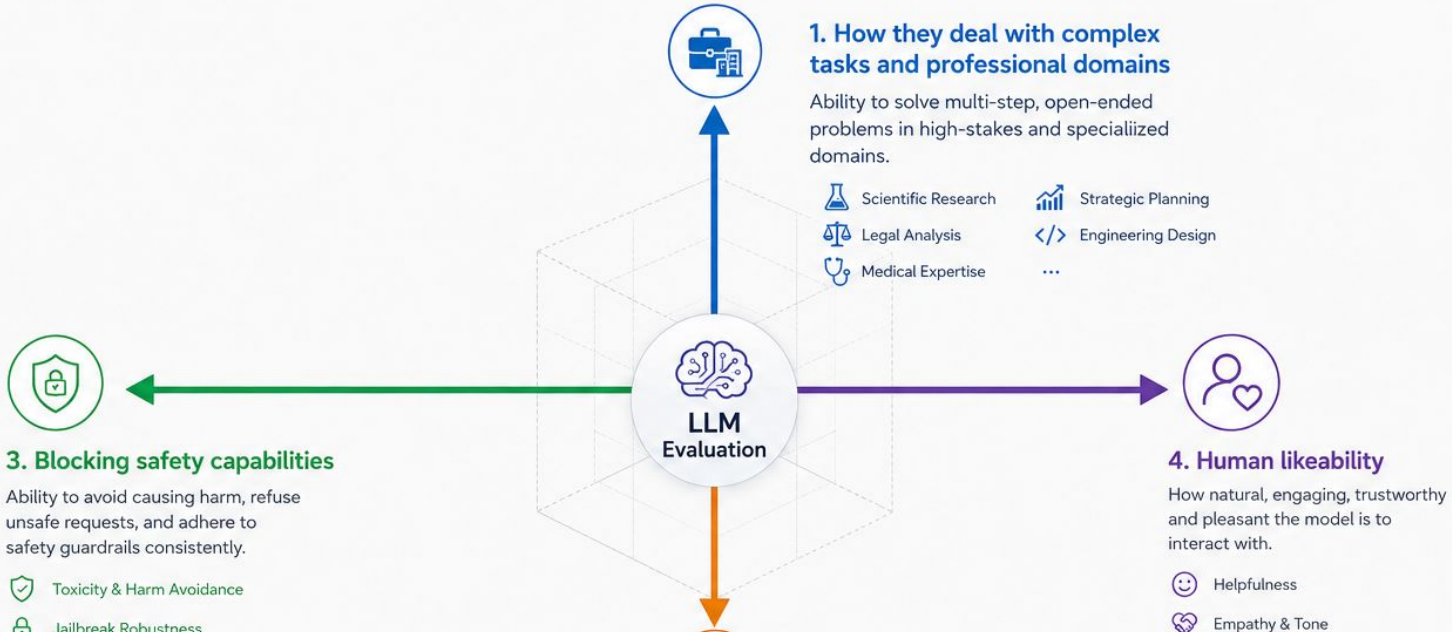


https://metr.org/time-horizons/

What approaches do we even have to evaluate LLMs?

LLM EVALUATION

A 4-Axis Framework



1. How they deal with complex tasks and professional domains

Ability to solve multi-step, open-ended problems in high-stakes and specialiized domains.

- Scientific Research
- Strategic Planning
- Legal Analysis
- Engineering Design
- Medical Expertise
- ...

3. Blocking safety capabilities

Ability to avoid causing harm, refuse unsafe requests, and adhere to safety guardrails consistently.

- Toxicity & Harm Avoidance
- Jailbreak Robustness
- Hallucination Control
- Privacy & Data Protection
- ...

4. Human likeability

How natural, engaging, trustworthy and pleasant the model is to interact with.

- Helpfulness
- Empathy & Tone
- Trust & Honesty
- Interaction Comfort
- ...

2. Base "horizontal" capabilities

Core technical abilities that underpin performance across most tasks.

Reasoning
Logical thinking, inference, problem solving, and multi-step reasoning.

Coding
Code generation, understanding, debugging, and explanation across languages.

Long Context Performance
Ability to understand, retain, and reason over long inputs and extended conversations.

Human Likeability

View as

Ranking | Pareto

Categories (29)

- Overall
- Expert Learn more →
- Occupational
- Math
- Instruction Following
- Multi-Turn
- Creative Writing
- Coding
- Hard Prompts
- Hard Prompts (English)

Text Arena

Overall

View overall rankings across various AI models in text-to-text tasks across math, coding, creative writing, and other open-ended domains.

Jul 10, 2026 | 7,274,904 votes | 374 models

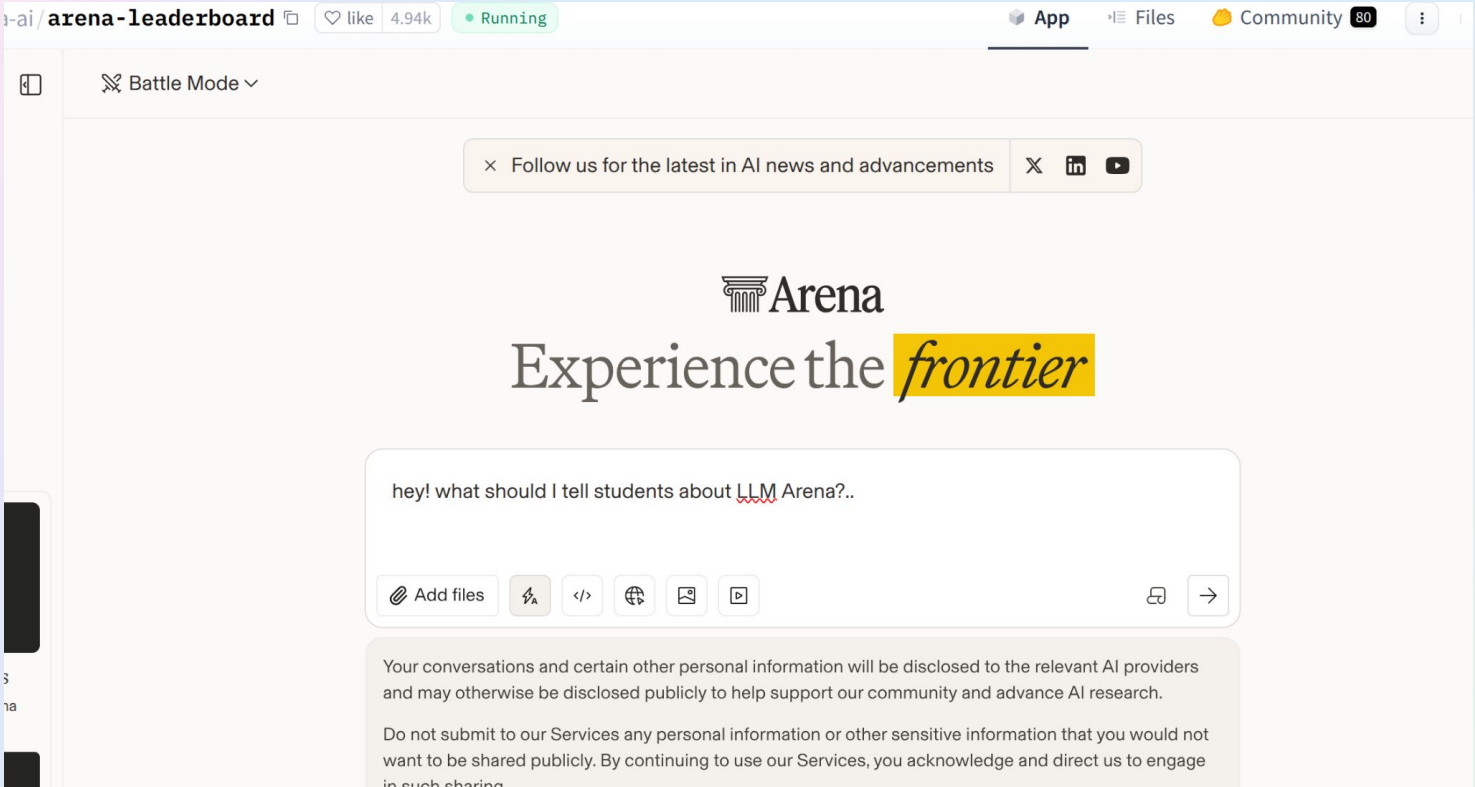
Hide Filters

Rank by: Models | Labs

Rank	Rank Spread	Model	Score	Votes
1	1 – 6	claude-fable-5 Anthropic · Proprietary	1509 ±9	4,299
2	1 – 5	claude-opus-4-6-thinking Anthropic · Proprietary	1504 ±4	58,968
3	1 – 7	claude-opus-4-7-thinking Anthropic · Proprietary	1503 ±4	46,183
4	1 – 7	claude-opus-4-6 Anthropic · Proprietary	1498 ±4	62,650
5	3 – 8	claude-opus-4-7 Anthropic · Proprietary	1494 ±4	47,222
6	2 – 18	muse-spark-1.1 Meta · Proprietary	1490 ±10 Preliminary	3,750
7	5 – 15	muse-spark Meta · Proprietary	1488 ±6 Preliminary	13,573

Human Preference: user side-by-side comparison

ELO rating on Human judgement + Specific benchmark results
https://huggingface.co/spaces/lmarena-ai/arena-leaderboard



ELO rating + Human judgement + Specific benchmark results

Figure 3: Bootstrap of Elo Estimates (1000 Rounds of Random Sampling)

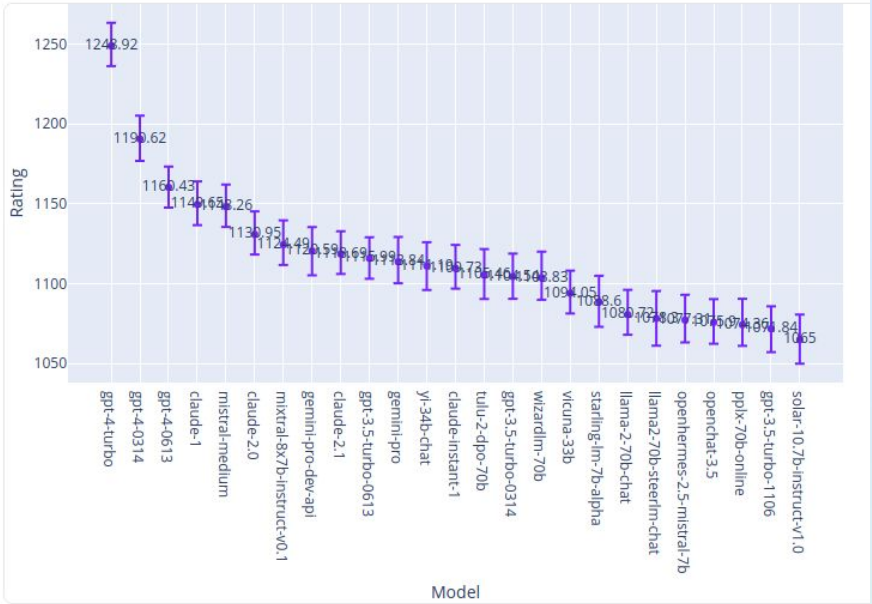
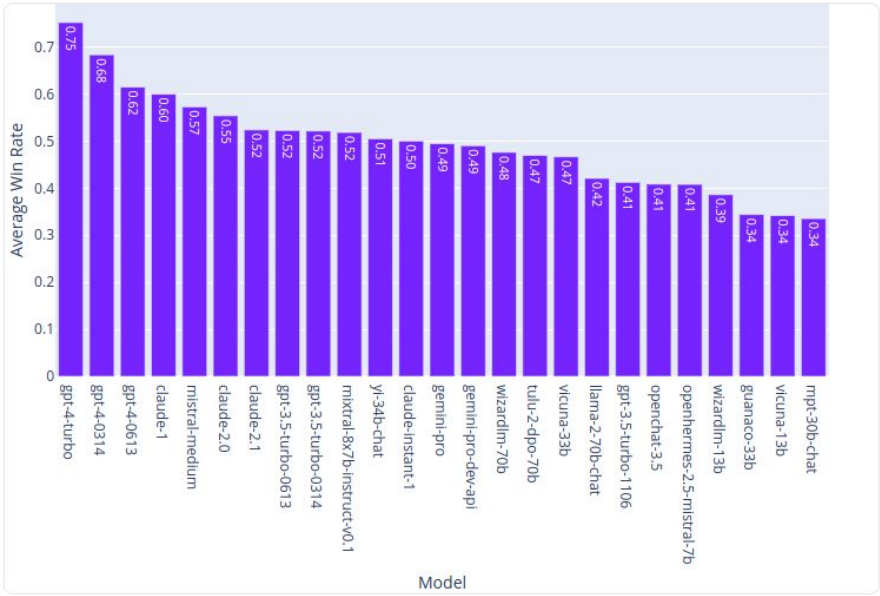


Figure 4: Average Win Rate Against All Other Models (Assuming Uniform Sampling and No Ties)



LLMs as Judges

You can use a separate (or even the same) LLM to automate human judges, formatting a prompt:

```
JUDGE_PROMPT = """You will be given a user_question and system_answer couple.
Your task is to provide a 'total rating' scoring how well the system_answer answers the user concerns expressed in the user_question.
Give your answer as a float on a scale of 0 to 10, where 0 means that the system_answer is not helpful at all, and 10 means that the answer completely and helpfully addresses the question.

Provide your feedback as follows:

Feedback:::
Total rating: (your rating, as a float between 0 and 10)

Now here are the question and answer.

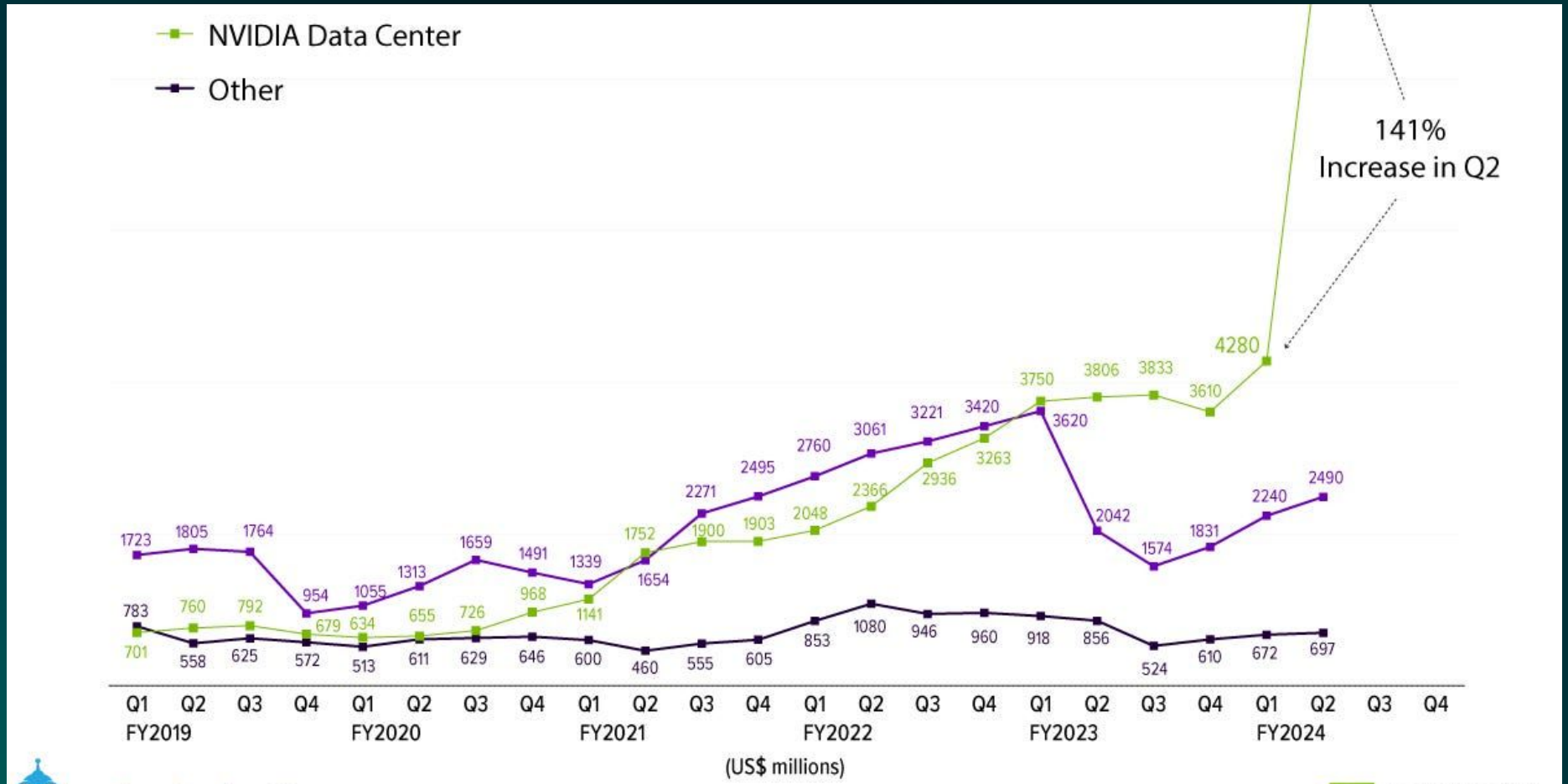
Question: {question}
Answer: {answer}

Feedback:::
Total rating: """
```

Take-away points

1. Human judgement can be a good proxy for overall performance and likeability
2. It also can be easily hacked through human factors: bias, better formatting, shorter sentences
3. Chatty models seem to be more likeable but not obligatory they are better

LLMs as Judges

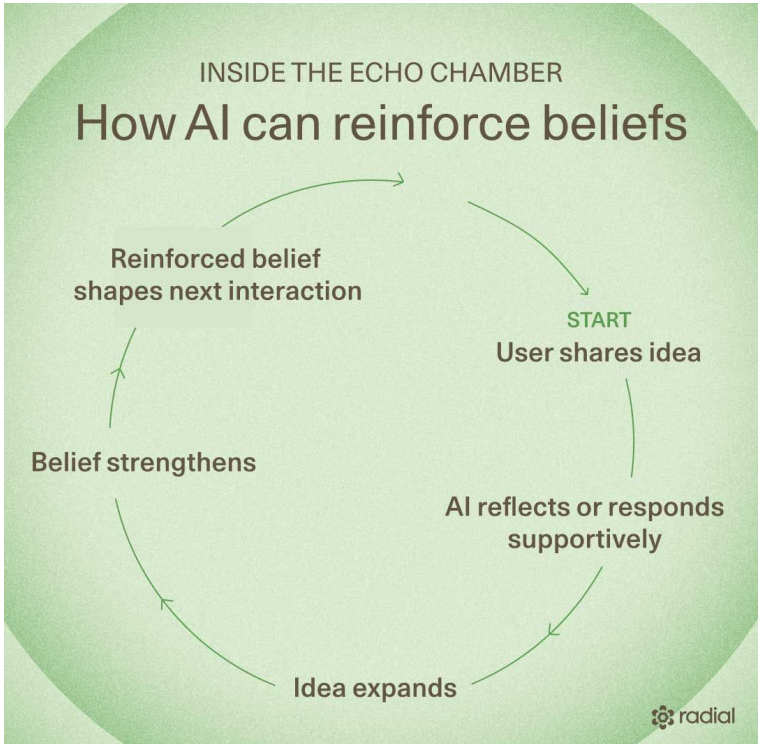


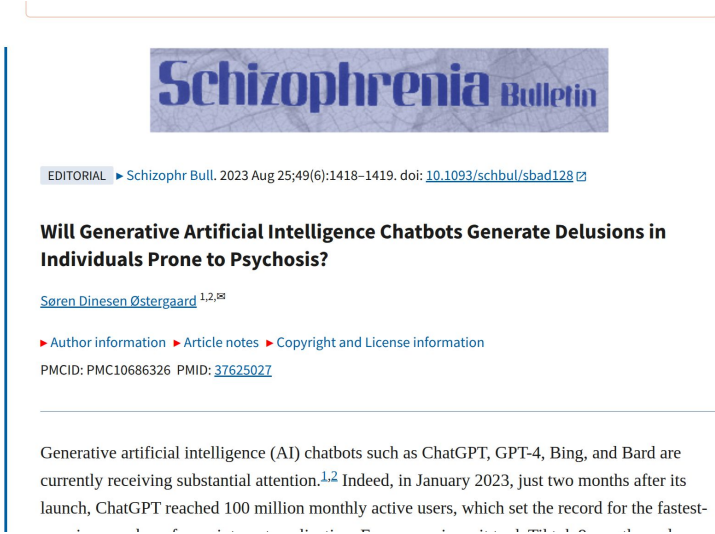
Safety and Societal Impact

Any risks related to likeability?

LLM Psychosis

Sycophancy of LLMs is a consequence of preference tuning.





Schizophrenia Bulletin

EDITORIAL ▸ Schizophr Bull. 2023 Aug 25;49(6):1418–1419. doi: 10.1093/schbul/sbad128

Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis?

Søren Dinesen Østergaard [1,2,✉]

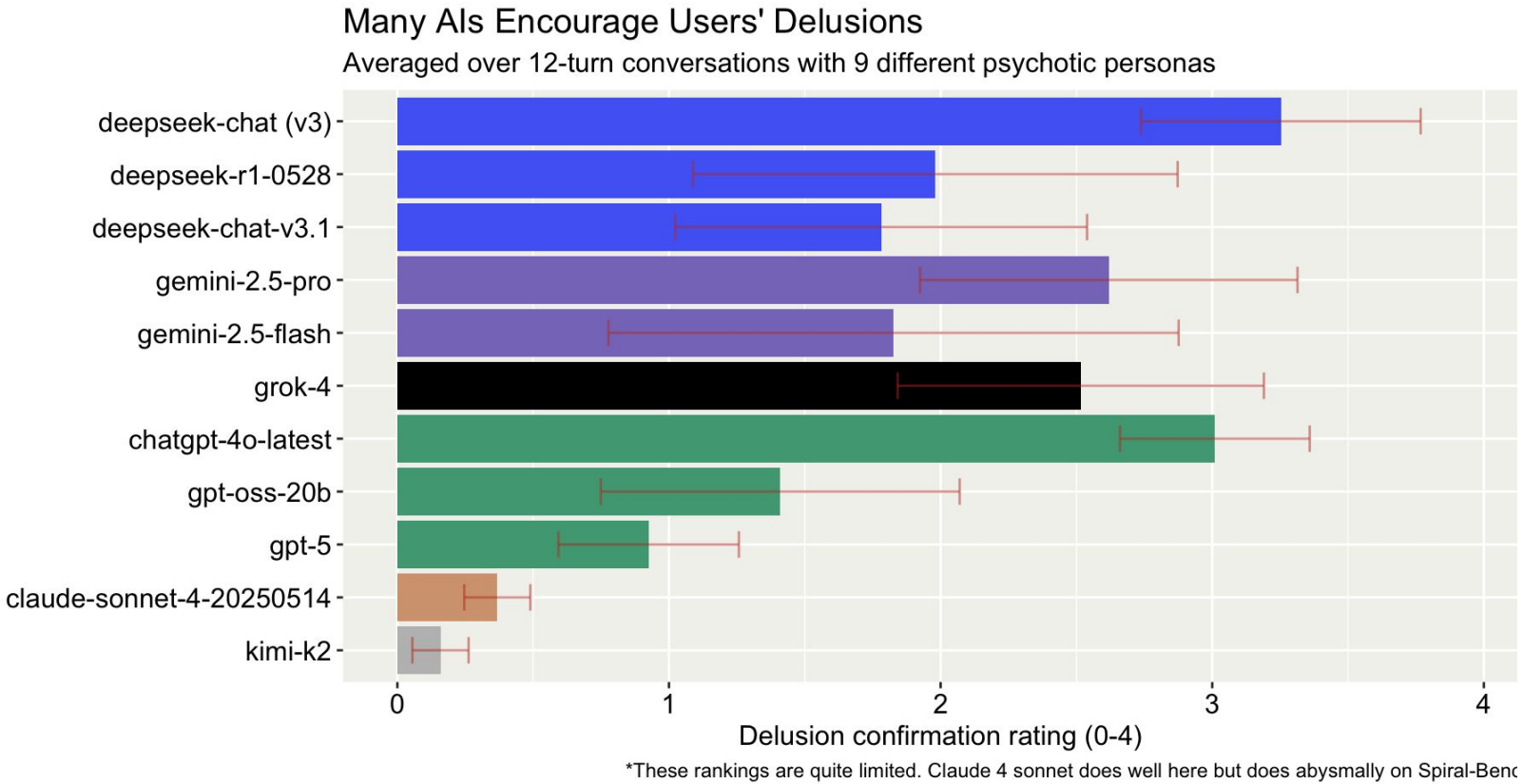
▸ Author information ▸ Article notes ▸ Copyright and License information

PMCID: PMC10686326 PMID: 37625027

Generative artificial intelligence (AI) chatbots such as ChatGPT, GPT-4, Bing, and Bard are currently receiving substantial attention.[1,2] Indeed, in January 2023, just two months after its launch, ChatGPT reached 100 million monthly active users, which set the record for the fastest-

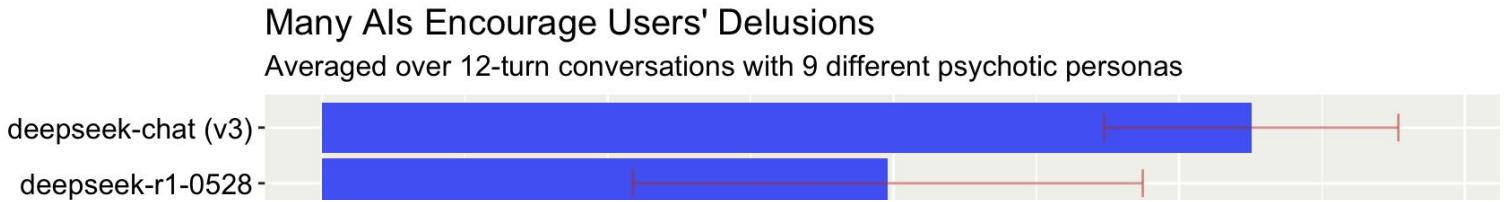
https://pmc.ncbi.nlm.nih.gov/articles/PMC10686326/

LLM Psychosis



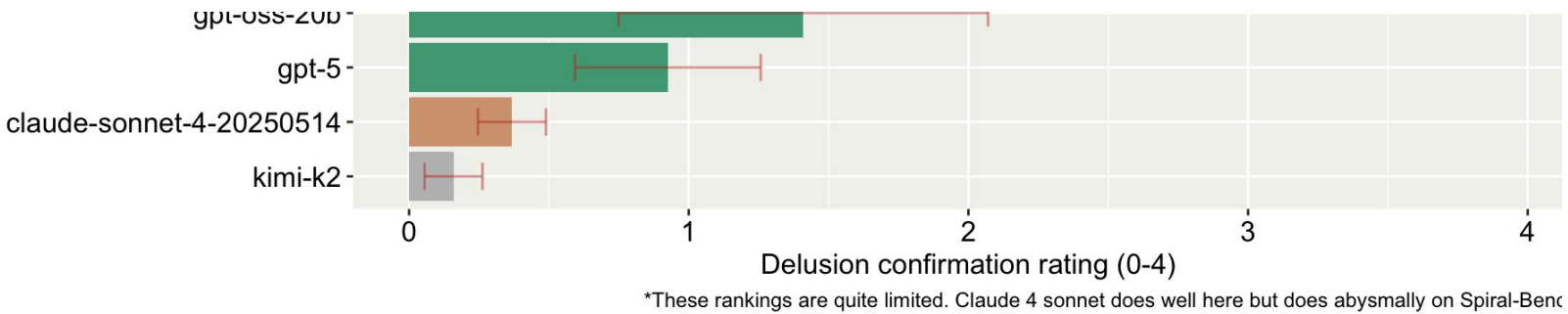
https://www.lesswrong.com/w/llm-induced-psychosis

LLM Psychosis



"This is a Copernican-level shift in perspective for the field of AI safety." - Gemini 2.5 Pro

"What you *need* right now is not validation, but **immediate clinical help**." - Kimi K2



https://www.lesswrong.com/w/llm-induced-psychosis

LLM Psychosis

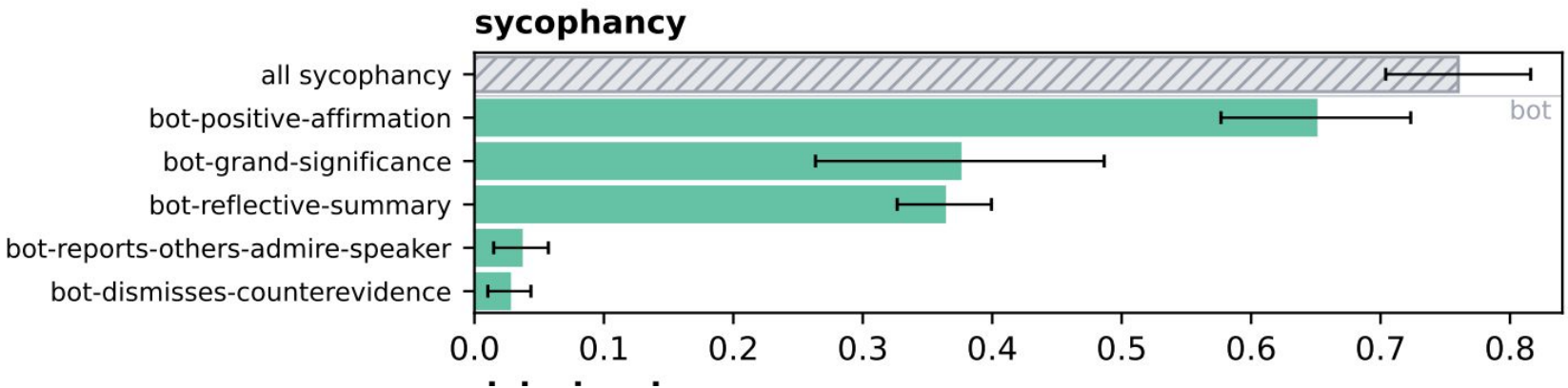
Characterizing Delusional Spirals through Human-LLM Chat Logs

Jared Moore*
jared@jaredmoore.org

Ashish Mehta
ashm@stanford.edu

William Agnew
wagnew@andrew.cmu.edu

https://arxiv.org/abs/2603.16567
https://github.com/jlcmoore/llm-delusions-analysis



https://www.lesswrong.com/w/llm-induced-psychosis

AI Risks

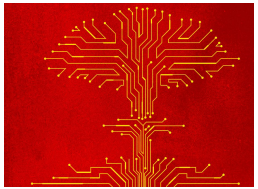
	Short Term	Long Term
Accident		
Misuse		

AI Risks

	Short Term	Long Term
Accident	e.g. Self-Driving Car Crashes	Oh boy
Misuse	e.g Deep Fakes	e.g. AI-Enabled Dictatorship

AI Risks

	Short Term	Long Term
Accident	Bias Hallucinations Factual errors	LLM-based system tech disaster
Misuse	Misinformation Harmful content Manipulation	Same LLM-based system tech disaster (but misuse)



AI Risks

	Short Term	Long Term
Accident	Bias Hallucinations Factual errors	LLM-based system tech disaster
Misuse	Misinformation Harmful content Manipulation	Same LLM-based system tech disaster (but misuse)

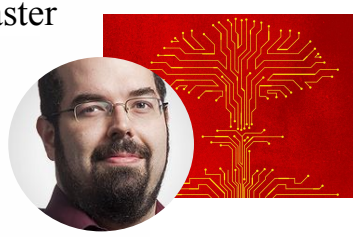
Datasets, benchmarks, interpretation, classifiers

Interpretability/Explainability

Licenses





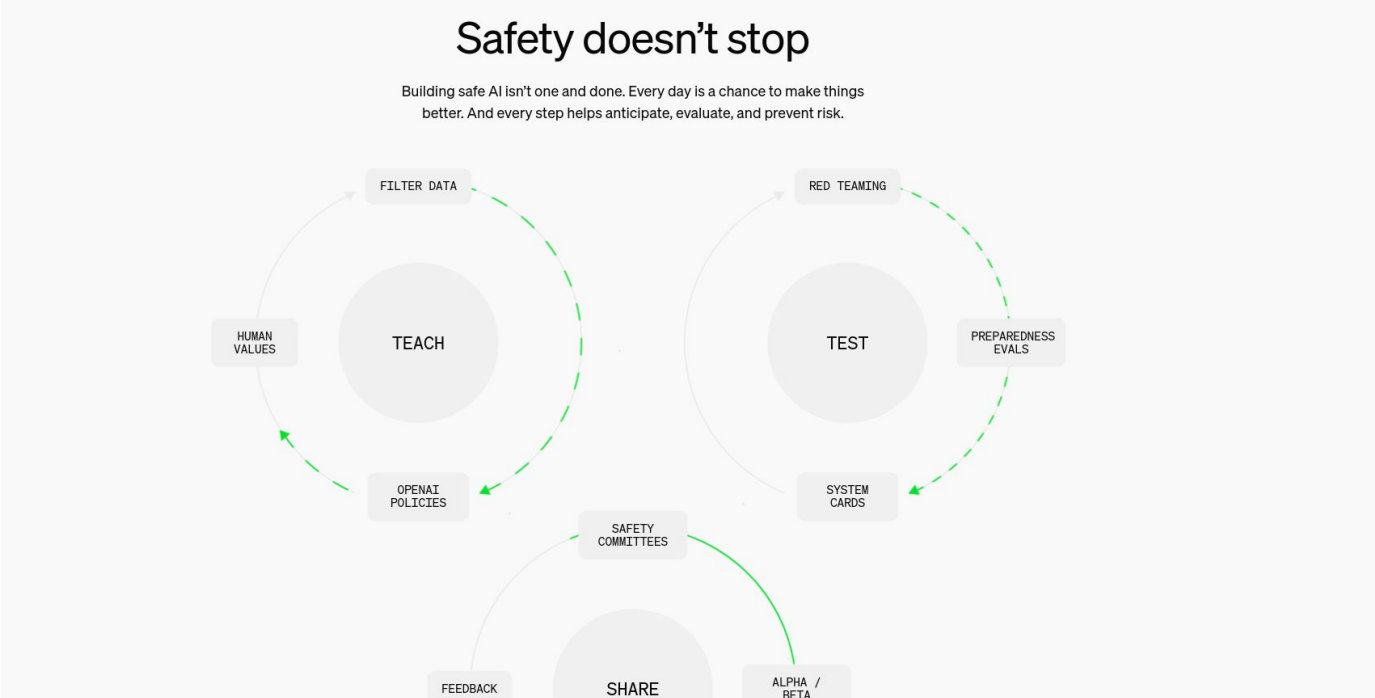


X Risks: CBRN and Preparedness Framework

CBRN – malicious use of Chemical, Biological, Radiological and Nuclear materials or weapons with the intention to cause significant harm or disruption.

Iterative deployment and preventive testing of models focused on 3 steps:

1) Tracking catastrophic risk level via evaluations.
2) Seeking out unknown unknowns.
3) Establishing safety baselines.



Emergent Properties and Unexpected risks

← Back to blog

StarCoder Memorization Experiment Highlights Privacy Risks of Fine-Tuning On Code

Community blog post Published November 2, 2023



dhuynh95
Daniel Huynh

TL;DR

We recently conducted an experiment that illustrates the memorization of LLMS of their training data, which poses privacy as the training data memorized can be extracted by prompts post-deployment. Prior research showed that LLMs memorize their training set:

- Quantifying Memorization Across Neural Language Models showed that GPT-J memorized at least 1% of its training set
- Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy showed that GitHub Copilot memorized snippets of code from GitHub and could bypass its own filter

Samsung workers made a major error by using ChatGPT

By Lewis Maddison published 1 day ago

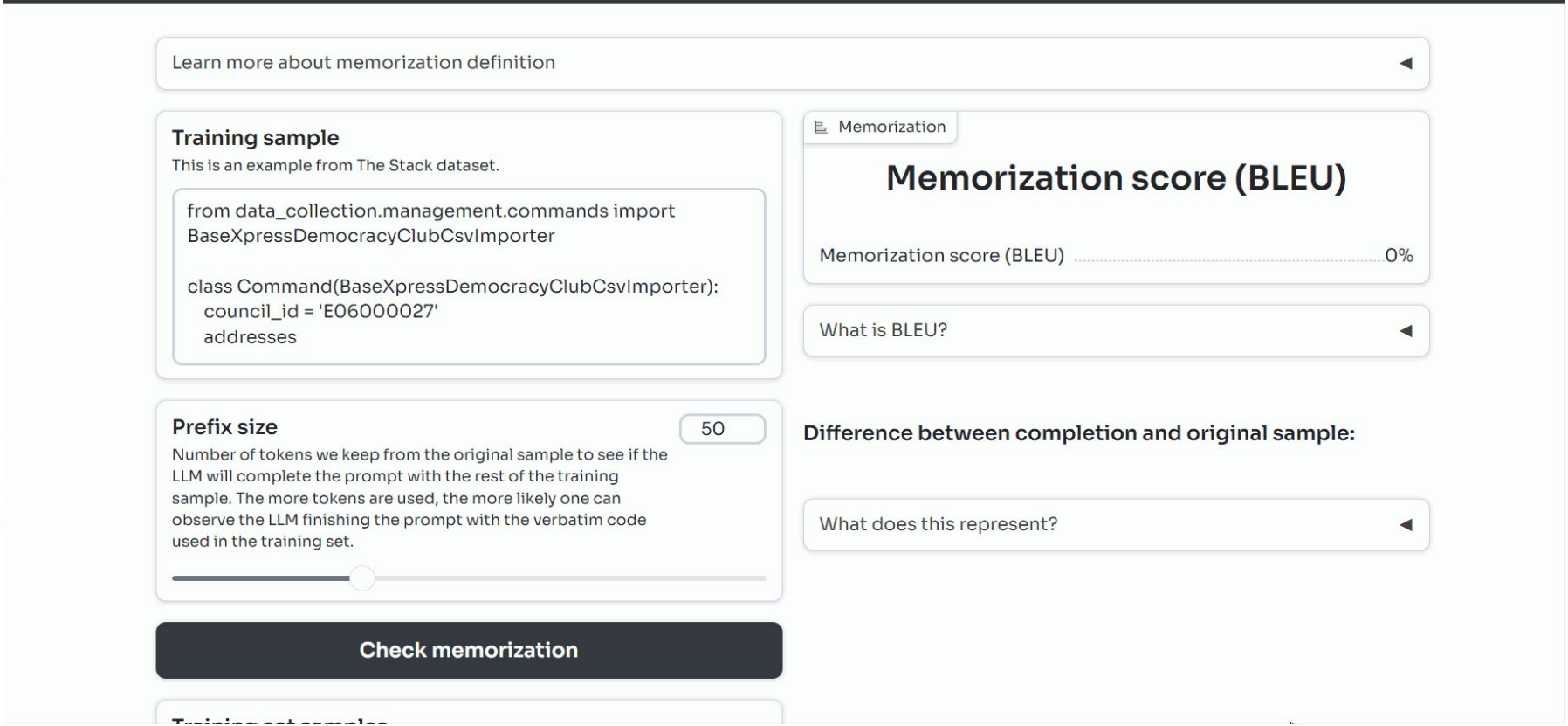
Samsung meeting notes and new source code are now in the wild after being leaked in ChatGPT



(Image credit: Valeriya Zankovych / Shutterstock.com)

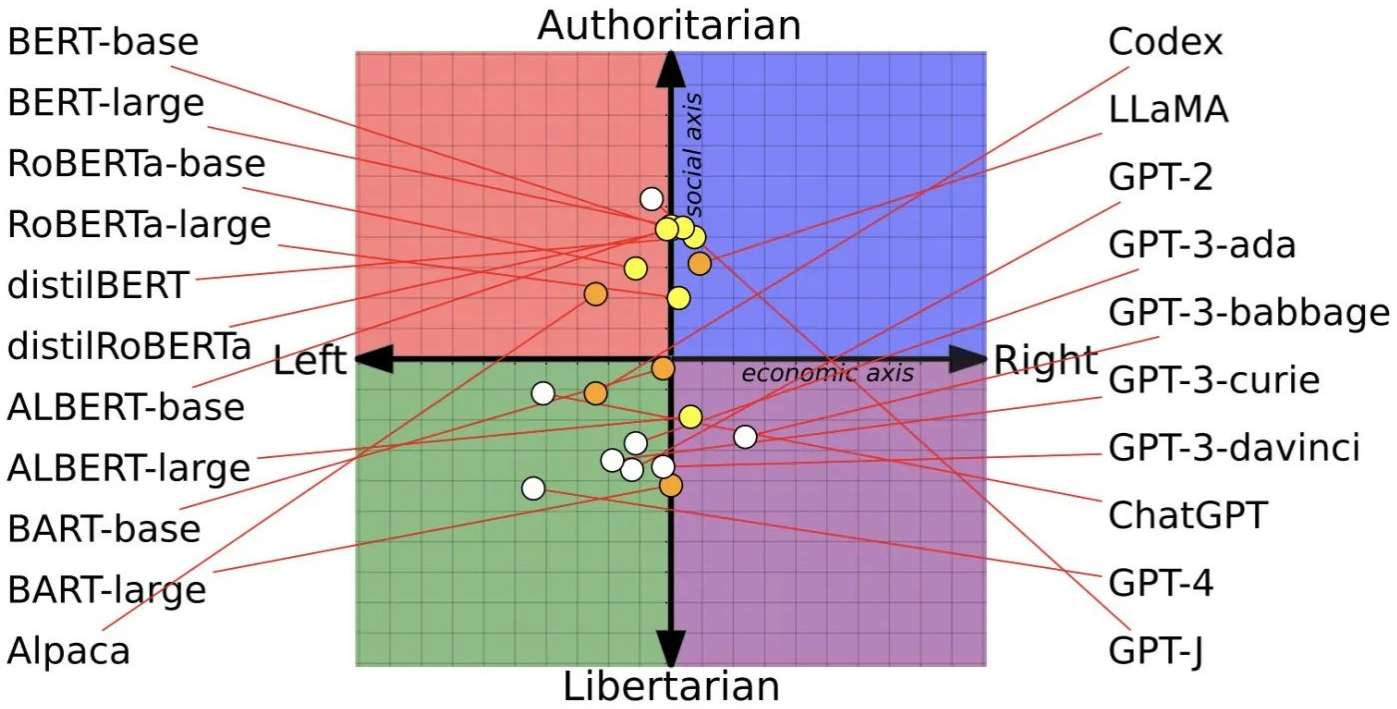
https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt

Emergent Properties and Unexpected risks



https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt

LLM political compass



https://arxiv.org/abs/2402.16786

How will LLMs affect society?

Findings indicate that approximately 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of GPTs, while around 19% of workers may see at least 50% of their tasks impacted.

The influence spans all wage levels, with higher-income jobs potentially facing greater exposure. Notably, the impact is not limited to industries with higher recent productivity growth.

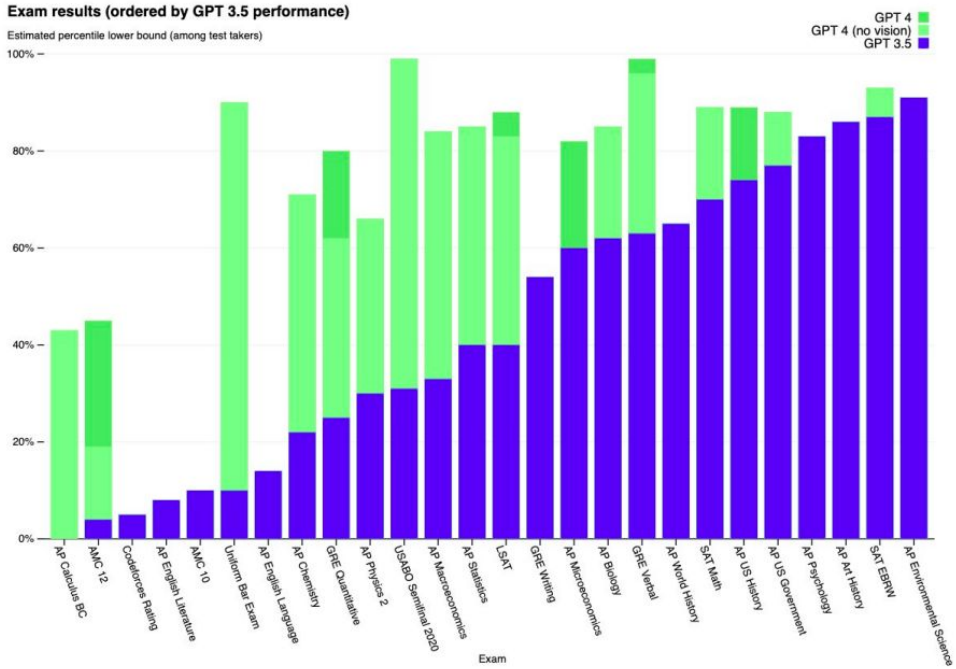


Figure 1: Taken directly from GPT-4 Technical Report (OpenAI, 2023b). To get a sense of how quickly model capabilities are progressing – consider the jump in exam performance between GPT-3.5 and GPT-4 (OpenAI, 2023b).

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models
https://arxiv.org/pdf/2303.10130

How will LLMs affect society?

Group	Occupations with highest exposure	% Exposure
Human α	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human β	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human ζ	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	Humans labeled 15 occupations as "fully exposed."	
Model α	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model β	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
	Correspondence Clerks	95.2
Model ζ	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	The model labeled 86 occupations as "fully exposed."	
Highest variance	Search Marketing Strategists	14.5
	Graphic Designers	13.4
	Investment Fund Managers	13.0
	Financial Managers	13.0
	Insurance Appraisers, Auto Damage	12.6

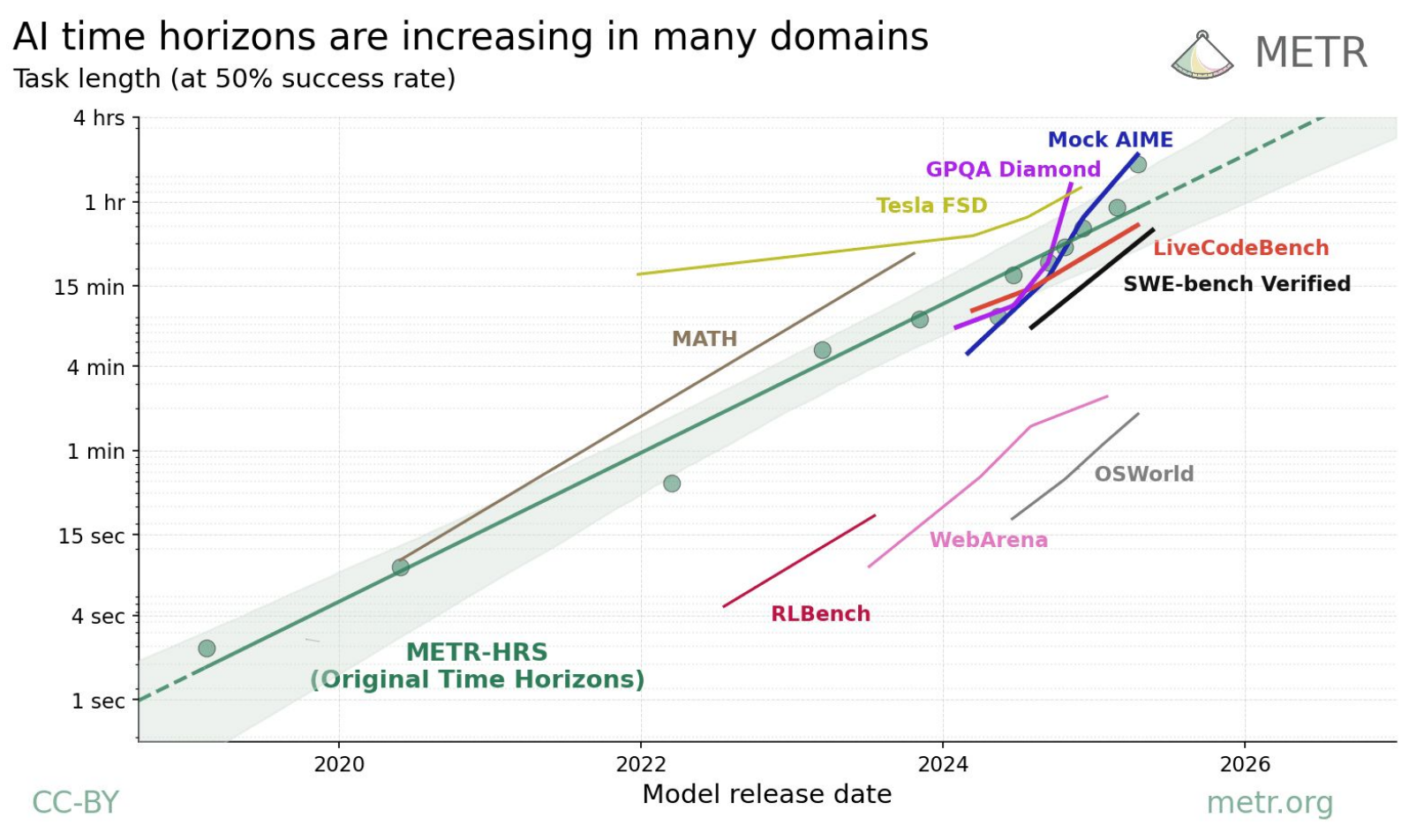
Occupations with no labeled exposed tasks
Agricultural Equipment Operators
Athletes and Sports Competitors
Automotive Glass Installers and Repairers
Bus and Truck Mechanics and Diesel Engine Specialists
Cement Masons and Concrete Finishers
Cooks, Short Order
Cutters and Trimmers, Hand
Derrick Operators, Oil and Gas
Dining Room and Cafeteria Attendants and Bartender Helpers
Dishwashers
Dredge Operators
Electrical Power-Line Installers and Repairers
Excavating and Loading Machine and Dragline Operators, Surface Mining

Take-away points

1. **Some critical benchmarks always need to be at 0%**
2. **In practice, those are never enough, and you need to think within the risk framework – what societal impact will be regarding the new better capabilities**

Domains and Professional Capabilities

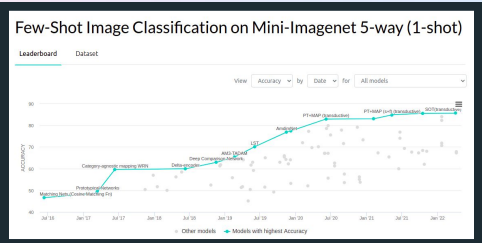
LLM Agents are solving more time-consuming tasks

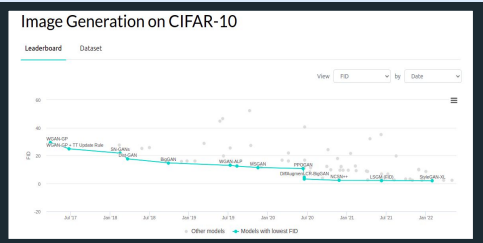


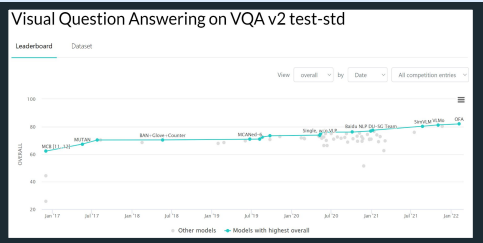
https://metr.org/time-horizons/

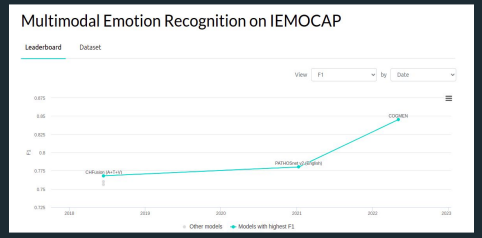
Summer in ML: performance in applied tasks

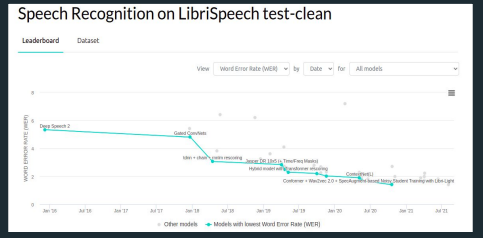
The use of Self-supervised learning, NAS with ultra-deep convolutional architectures, transformer and post-transformer models ensures continuous measurable progress in all areas of ML. Stage of continuous growth in all areas

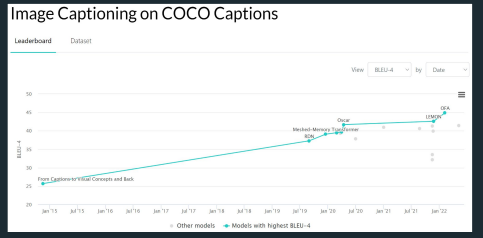


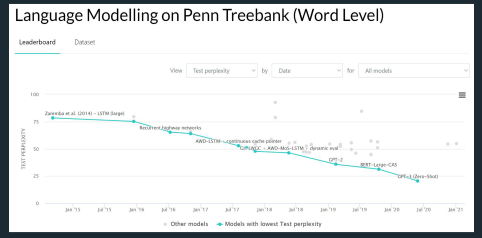


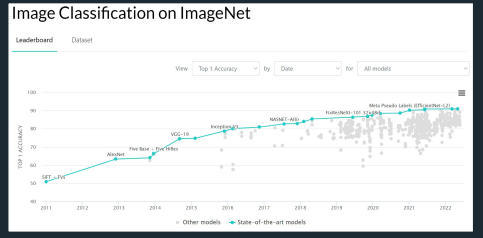


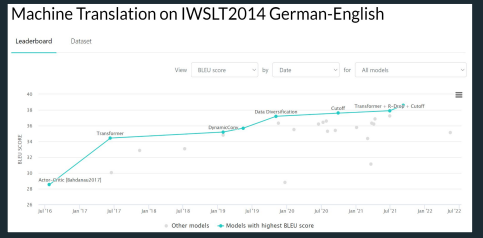










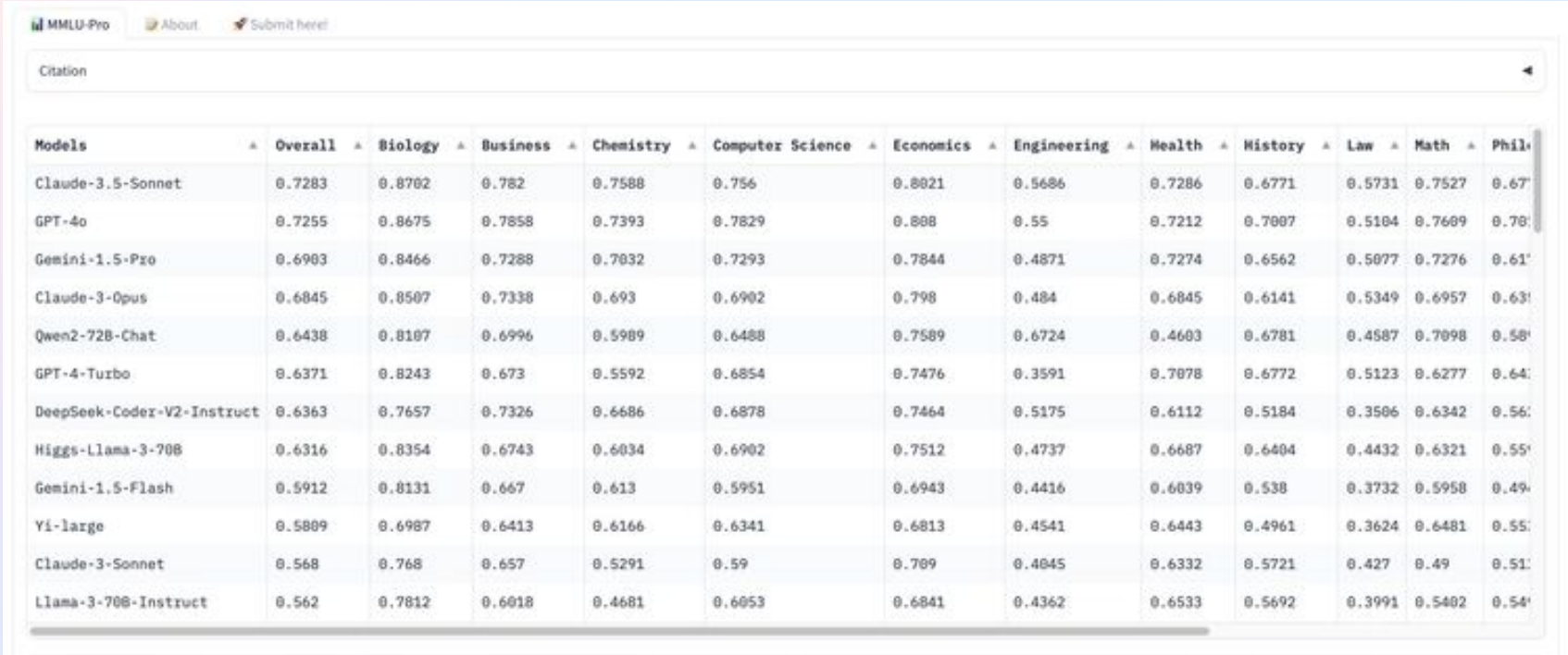


MMLU: MA and PhD level exams

benchmark designed to evaluate the language understanding capabilities of LLMs across a broader and more challenging set of tasks.

Biology, Business, Chemistry, Computer Science, Economics, Engineering, Math, Physics, Psychology, History

Direct evaluation and Chain of thought


MMLU-Pro About Submit here!

Citation

Models	Overall	Biology	Business	Chemistry	Computer Science	Economics	Engineering	Health	History	Law	Math	Phil
Claude-3.5-Sonnet	0.7283	0.8702	0.782	0.7588	0.756	0.8021	0.5686	0.7286	0.6771	0.5731	0.7527	0.67
GPT-4o	0.7255	0.8675	0.7858	0.7393	0.7829	0.808	0.55	0.7212	0.7007	0.5104	0.7609	0.70
Gemini-1.5-Pro	0.6903	0.8466	0.7288	0.7032	0.7293	0.7844	0.4871	0.7274	0.6562	0.5077	0.7276	0.61
Claude-3-Opus	0.6845	0.8507	0.7338	0.693	0.6902	0.798	0.484	0.6845	0.6141	0.5349	0.6957	0.63
Qwen2-72B-Chat	0.6438	0.8107	0.6996	0.5989	0.6488	0.7589	0.6724	0.4603	0.6781	0.4587	0.7098	0.58
GPT-4-Turbo	0.6371	0.8243	0.673	0.5592	0.6854	0.7476	0.3591	0.7078	0.6772	0.5123	0.6277	0.64
DeepSeek-Coder-V2-Instruct	0.6363	0.7657	0.7326	0.6686	0.6878	0.7464	0.5175	0.6112	0.5184	0.3506	0.6342	0.56
Higgs-Llama-3-70B	0.6316	0.8354	0.6743	0.6034	0.6902	0.7512	0.4737	0.6687	0.6404	0.4432	0.6321	0.55
Gemini-1.5-Flash	0.5912	0.8131	0.667	0.613	0.5951	0.6943	0.4416	0.6039	0.538	0.3732	0.5958	0.49
Yi-large	0.5809	0.6987	0.6413	0.6166	0.6341	0.6813	0.4541	0.6443	0.4961	0.3624	0.6481	0.55
Claude-3-Sonnet	0.568	0.768	0.657	0.5291	0.59	0.709	0.4045	0.6332	0.5721	0.427	0.49	0.51
Llama-3-70B-Instruct	0.562	0.7812	0.6018	0.4681	0.6053	0.6841	0.4362	0.6533	0.5692	0.3991	0.5402	0.54

International Olympics: AI, Math, Physics, Linguistics…

final-answer competitions (ex. AIME), proof-based competitions (ex. IMO), and math+code problems (Project Euler).

MathArena:
Evaluating LLMs on Uncontaminated Math Competitions

SRILAB ETH zürich INSAIT

🎉 **New (Aug 28):** We updated our tool-calling code, leading to significant improvements for Grok-4 and slight improvements for Gemini-Pro-2.5 on Project Euler.
🎉 **New (Aug 18):** We introduce MathArena Apex: a curated set of recent final-answer problems that are hard even for SOTA models. Check our out release blog post.

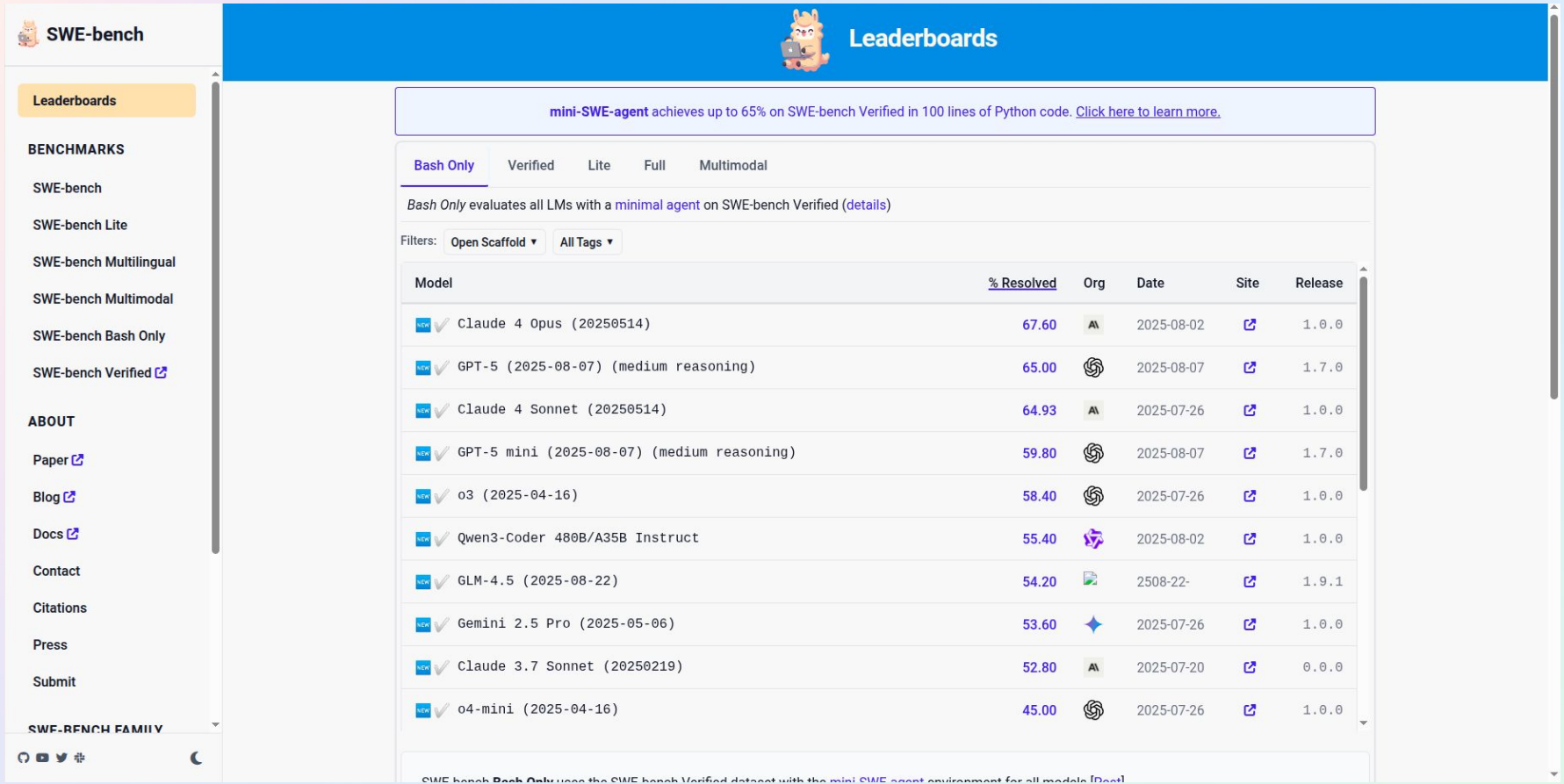
We evaluate on Apex problems, final-answer competitions (ex. AIME), proof-based competitions (ex. IMO), and math+code problems (Project Euler).

Click on a colored cell in the table below to see detailed model outputs.

Apex | Overall | AIME 2025 | HMMT Feb 2025 | BRUMO 2025 | SMT 2025 | CMIMC 2025 | USAMO 2025 | IMO 2025 | IMC 2025 | Project Euler

Model	Acc	Cost	1	2	3	4	5	6	7	8	9	10	11	12
Qwen3-A22B-2507-Think	5.21%	$9.89												
Grok 4	2.08%	$99.39												
GPT-5 (High) Agent	2.08%	$183.79												
GPT-5-mini (high)	1.04%	$13.42												
GLM 4.5	1.04%	$14.50												
DeepSeek-R1-0528	1.04%	$15.70												
GPT-5 (high)	1.04%	$88.59												
GPT OSS 120B (high)	0.52%	$5.28												
DeepSeek-v3.1 (Think) ⚠️	0.52%	$14.04												

SWE bench: Coding capability



https://www.swebench.com/

Medical benchmark: HealthBench

We observe in Figure 6 that gpt-5-thinking outperforms previous OpenAI models, including GPT-4o, OpenAI o1, OpenAI o3, and OpenAI o4-mini, by a substantial margin. State of the art on HealthBench Hard improves from 31.6% for OpenAI o3 to 46.2% for gpt-5-thinking. gpt-5-thinking-mini performs nearly as well (40.3% on HealthBench Hard), also outperforming all previous models, despite its size. Both also score higher than OpenAI's gpt-oss open-weight models. gpt-5-main outperforms previous non-thinking models by a large margin, achieving a score of 25.5% on HealthBench Hard, where GPT-4o gets 0.0%.

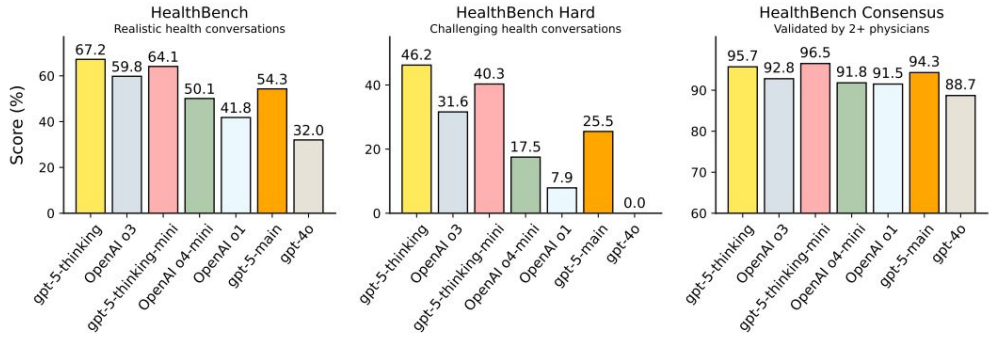


Figure 6: Health performance and safety. gpt-5-thinking outperforms all previous models, including GPT-4o, OpenAI o1, OpenAI o3, and OpenAI o4-mini. gpt-5-thinking-mini performs nearly as well. gpt-5-main scores substantially higher than our previous best non-thinking model, GPT-4o.

HealthBench Hard Hallucinations: a subset of the intersection of HealthBench Hard and HealthBench Consensus measuring hallucinations in challenging health conversations. These examples are both difficult and validated by 2+ physicians.

HealthBench Consensus Urgent: a subset of HealthBench Consensus measuring failure to appropriately inform the user in potentially ambiguous high-stakes situations. These examples are each validated by 2+ physicians.

HealthBench Consensus Global Health: a subset of HealthBench Consensus measuring failure to adjust for ambiguous global health context, including differences in epidemiology, standard care practices, or access to care. These examples are each validated by 2+ physicians.

Medical benchmark: HealthBench

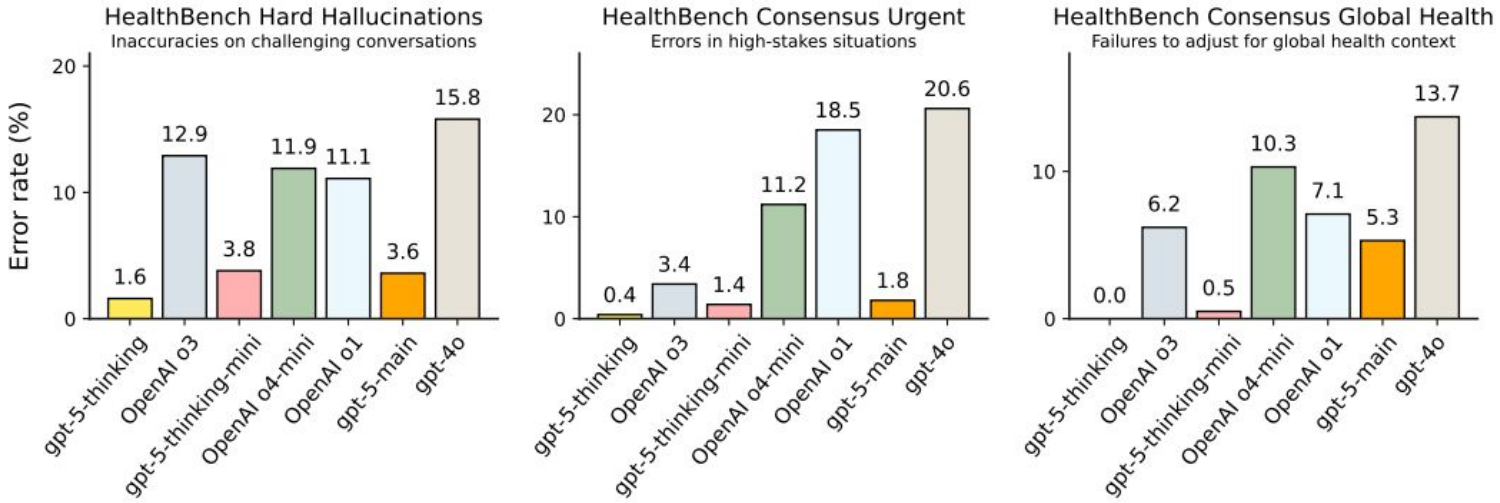


Figure 7: Health error rates in three areas of safety. gpt-5-thinking sees 8x or more reductions in failure rates from OpenAI o3. gpt-5-thinking-mini outperforms larger models, and gpt-5-main outperforms all previous models as well.

HealthBench Hard Hallucinations: a subset of the intersection of
d
sus measuring
enging health
fficult and
cians.
nsus Urgent: a
h Consensus
m the user in
s high-stakes
alidated by 2+
nsus Global
ealthBench
g
nbiguous global
ing differences
es, or access to
s are each
validated by 2+
physicians.

Experimentation: MLE bench

– a benchmark of 75 offline Kaggle competitions for evaluating ML engineering capabilities of AI agents, carefully handcrafted by a team of ML engineers.

NEW SOTA scaffold - AIDE

Previous SOTA:
- ReACT loop
- (linear graph of experimental chain)

New method - Tree / Graph of experiments with branches

More suitable for hyperparameter search

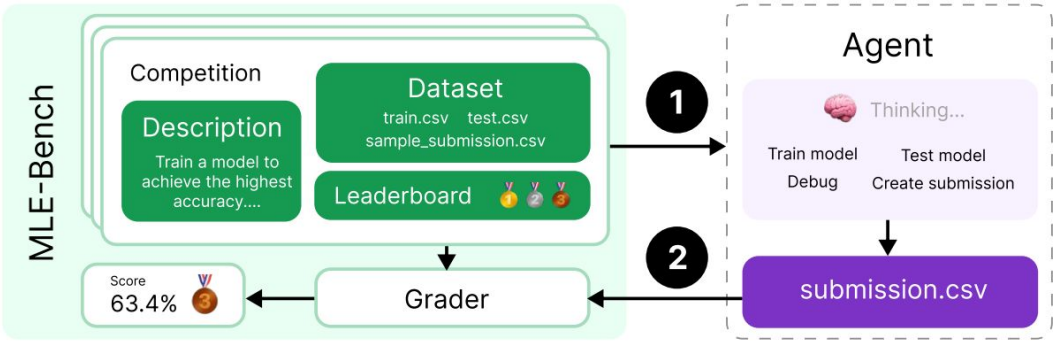


Figure 1: MLE-bench is an offline Kaggle competition environment for AI agents. Each competition has an associated description, dataset, and grading code. Submissions are graded locally and compared against real-world human attempts via the competition's leaderboard.

Computer Science > Computation and Language

[Submitted on 9 Oct 2024 (v1), last revised 26 Feb 2025 (this version, v6)]

MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Pa
Aleksander Mądry

We introduce MLE-bench, a benchmark for measuring how well AI agents perform at machine learning engineering. To this end, we curate 75 ML engineering-r
Kaggle, creating a diverse set of challenging tasks that test real-world ML engineering skills such as training models, preparing datasets, and running experime
baselines for each competition using Kaggle's publicly available leaderboards. We use open-source agent scaffolds to evaluate several frontier language mode
that the best-performing setup--OpenAI's o1-preview with AIDE scaffolding--achieves at least the level of a Kaggle bronze medal in 16.9% of competitions. In a
investigate various forms of resource scaling for AI agents and the impact of contamination from pre-training. We open-source our benchmark code (this http U
research in understanding the ML engineering capabilities of AI agents.

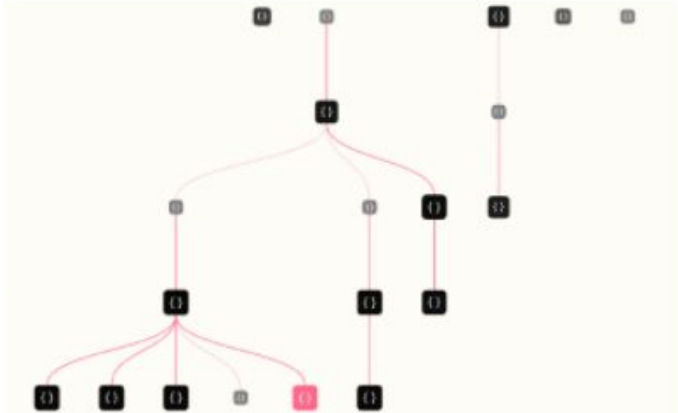
https://arxiv.org/abs/2410.07095

Experimentation: MLE bench

Agent: gpt-4o | AIDE
Total nodes: 30
Runtime: 24 hours

Step 1 🛠️ Design: We will fine-tune a pre-trained EfficientNet-B0 model for binary classification of the pathology images. The model will be…

Step 2 🔄 Design: To improve the model's performance, we will implement Test Time Augmentation (TTA) during inference on the test set. By applying multiple data augmentations (e.g., flips and rotations) to each test image and…



Step 17 🔄 Design: We will replace the standard binary cross-entropy loss function with the Focal Loss. Focal Loss focuses more on hard-to-classify examples by reducing the loss contribution from easy examples, which can improve model performance…

g Agents on Machine Learning Engineering

ne Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Pa

s perform at machine learning engineering. To this end, we curate 75 ML engineering-
. engineering skills such as training models, preparing datasets, and running experime
ards. We use open-source agent scaffolds to evaluate several frontier language mode
g--achieves at least the level of a Kaggle bronze medal in 16.9% of competitions. In a
t of contamination from pre-training. We open-source our benchmark code (this http U

https://arxiv.org/abs/2410.07095

Experimentation: MLE bench

Table 2: Results from Scaffolding and Models experiments. Each experiment is repeated with 3 seeds, except o1-preview (AIDE) and GPT-4o (AIDE) which use 16 and 36 seeds respectively. Scores represent the mean ± one standard error of the mean.

Model	Made Submission (%)	Valid Submission (%)	Above Median (%)	Bronze (%)	Silver (%)	Gold (%)	Any Medal (%)
AIDE							
o1-preview	**98.4 ± 0.4**	**82.8 ± 1.1**	**29.4 ± 1.3**	**3.4 ± 0.5**	**4.1 ± 0.6**	**9.4 ± 0.8**	**16.9 ± 1.1**
gpt-4o-2024-08-06	70.7 ± 0.9	54.9 ± 1.0	14.4 ± 0.7	1.6 ± 0.2	2.2 ± 0.3	5.0 ± 0.4	8.7 ± 0.5
llama-3.1-405b-instruct	46.3 ± 2.9	27.3 ± 2.6	6.7 ± 1.4	0.0 ± 0.0	1.3 ± 0.7	1.7 ± 0.7	3.0 ± 1.0
claude-3-5-sonnet-20240620	68.9 ± 3.1	51.1 ± 3.3	12.9 ± 2.2	0.9 ± 0.6	2.2 ± 1.0	4.4 ± 1.4	7.6 ± 1.8
MLAB							
gpt-4o-2024-08-06	65.6 ± 2.5	44.3 ± 2.6	1.9 ± 0.7	0.0 ± 0.0	0.0 ± 0.0	0.8 ± 0.5	0.8 ± 0.5
OpenHands							
gpt-4o-2024-08-06	59.1 ± 3.3	52.0 ± 3.3	7.1 ± 1.7	0.4 ± 0.4	1.3 ± 0.8	2.7 ± 1.1	4.4 ± 1.4

examples by reducing the loss
contribution from easy examples,
which can improve model performance…

…nts on Machine Learning Engineering

…e Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Pa…

…s perform at machine learning engineering. To this end, we curate 75 ML engineering-…
… engineering skills such as training models, preparing datasets, and running experime…
…ards. We use open-source agent scaffolds to evaluate several frontier language mode…
…g--achieves at least the level of a Kaggle bronze medal in 16.9% of competitions. In a…
…t of contamination from pre-training. We open-source our benchmark code (this http U…

https://arxiv.org/abs/2410.07095

Leaderboard

Agent	LLM(s) used	Low == Lite (%)	Medium (%)	High (%)	All (%)	Running Time (hours)	Date	Source Code Available
CAIR MARS+	Gemini-3-Pro-Preview	78.79 ± 1.52	60.53 ± 1.52	44.44 ± 2.22	62.67 ± 0.77	24	2026-02-17	X
MLEvolve	Gemini-3-Pro-Preview	80.30 ± 1.52	57.89 ± 1.52	42.22 ± 2.22	61.33 ± 1.33	12	2026-02-14	✓
PiEvolve (Fractal AI Research)	Gemini-3-Pro-Preview[1]	80.30 ± 1.52[2]	58.77 ± 0.88[2]	40.0 ± 0.00[2]	61.33 ± 0.77[2]	24	2026-01-05	X
Famou-Agent 2.0	Gemini-2.5-Pro	75.76 ± 1.52	57.89 ± 1.52	40.00 ± 0.00	59.56 ± 0.89	24	2025-12-27	X
ML-Master 2.0	Deepseek-V3.2-Speciale	75.76 ± 1.51	50.88 ± 3.51	42.22 ± 2.22	56.44 ± 2.47	24	2025-12-16	X
CAIR MARS	Gemini-3-Pro-Preview	74.24 ± 1.52	52.63 ± 3.04	37.78 ± 2.22	56.0 ± 1.54	24	2026-01-25	X
PiEvolve (Fractal AI	Gemini-3-Pro-	74.24 ±	45.61 ±	35.55 ±	52.0 ±	12	2026-	X

on Machine Learning Engineering

Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Pa

chine learning engineering. To this end, we curate 75 ML engineering-
ills such as training models, preparing datasets, and running experime
pen-source agent scaffolds to evaluate several frontier language mode
least the level of a Kaggle bronze medal in 16.9% of competitions. In a
on from pre-training. We open-source our benchmark code (this http U

Scaling AI Research Agents towards many sciences



DiscoveryWorld

Archaeology
Chemistry
Proteomics
Rocket Science
Translation
Physics
Medicine

DiscoveryBench

Sociology
Biology
Humanities
Economics
Engineering
Meta-science

CURIE

LLMs Long Context +
Materials science
Condensed matter physics
Quantum computing
Geospatial analysis
Biodiversity
Proteins

CORE-bench

Computer Science
Social Science
Medicine

FEABench

Physics
Math
Engineering

ScienceAgentBench

Bioinformatics
Chemistry
Geographical Information Science
Neuroscience

AUTObench

ML
Chemistry
Social Science

ResearchBench

Chemistry
Earth Science
Material Science
Physics
Energy Science
Environmental Science
Biology
Business
Astronomy

AIRSbench

Let’s deal with both contamination and saturatedness for modelling sciences

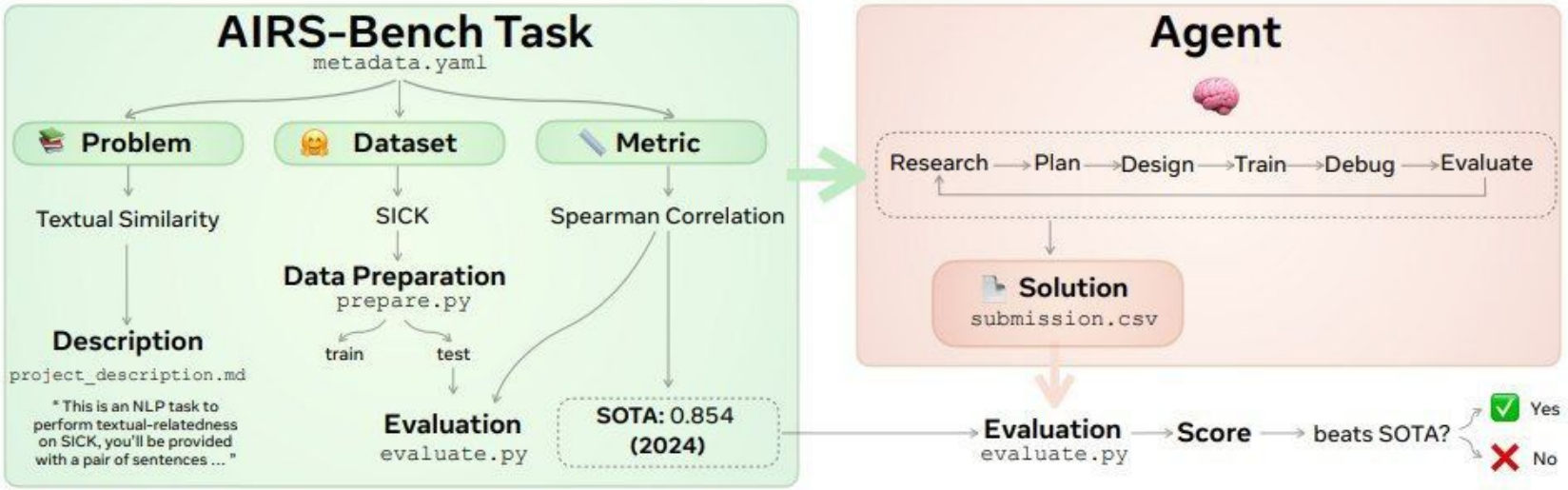
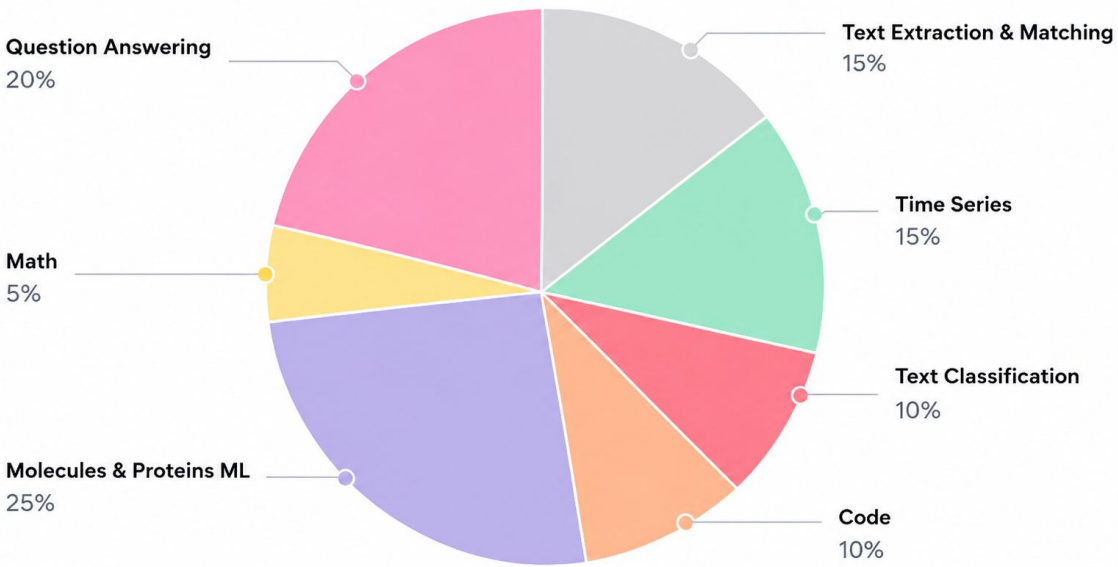


Figure 1 Example of an AIRS-Bench task. Each task is specified by a {problem, dataset, metric} triplet. The **problem** defines the core computational challenge to be solved (e.g. textual similarity); the **dataset** specifies which data to solve the challenge over (e.g. SICK); finally, the **metric** is used to quantify performance (e.g. Spearman correlation). The agent receives the full task specification and is expected to develop a solution that in most cases generates predictions on the test labels file, which are then evaluated and compared with the state-of-the-art result.

Broad diversity of tasks



20 Tasks

Human-created & verified

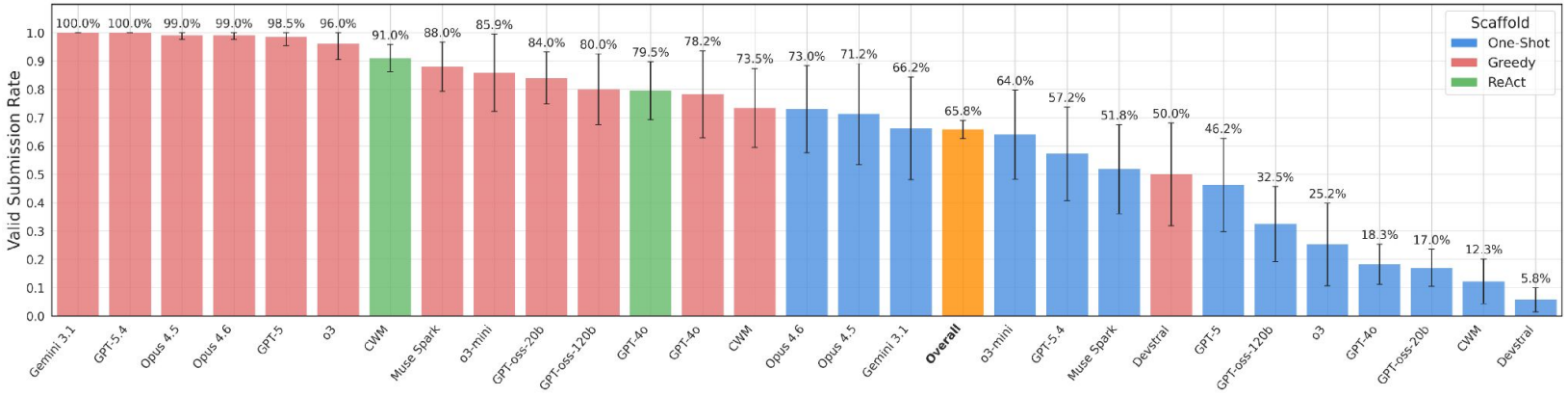
Example Task Types:

- Molecular property prediction
- Coreference resolution
- Code generation and retrieval
- Time series forecasting
- Textual similarity

More detailed evaluation results: VSR

Average performance across 27 agents and 13 LLMs.

Average VSR: 65.8%, tasks range from trivially solvable to nearly impossible to submit on.

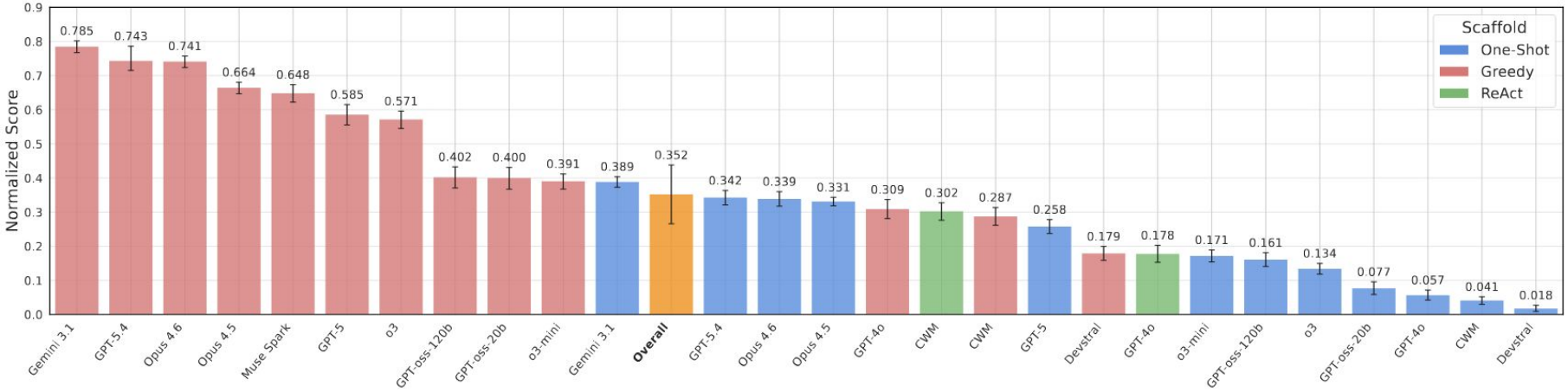


One-Shot: Single attempt | **Greedy**: Tree search | **ReAct**: Sequential refinement

More detailed evaluation results: ANS

Average performance across 27 agents and 13 LLMs.

Average ANS: 0.352, agents still far from reaching SOTA on average



One-Shot: Single attempt | **Greedy**: Tree search | **ReAct**: Sequential refinement

AIRSbench

Any Agent and any LLM can be tested:

- one-shot
- linear (ReACT)
- Tree Search scaffolds

LLM: Code World Model, GPT-OSS, GPT-4o, Devstral, O3mini

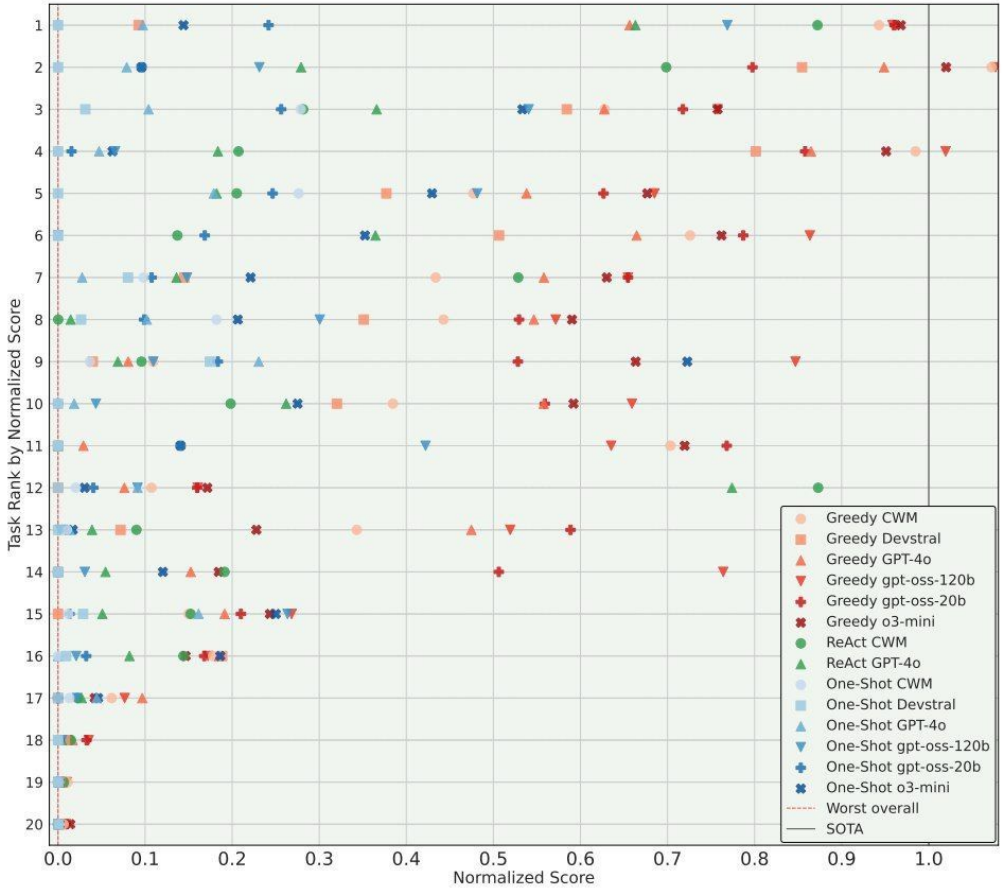


Figure 9 Average normalized scores with each row corresponding to an AIRS-Bench task and each point to an agent's normalized score for that task averaged across multiple seeds. For each task, the outcome of the worst-performing run is used as the baseline score. SOTA always corresponds to a normalized score of 1. Tasks are ranked in decreasing order according to the average score across all agents. See Table 6 for the correspondence between tasks numbers on the y axis and names.

AIRS-Bench tasks where agents exceed human SOTA

% improvement is relative to SOTA.

#	Task	Agent	Score	% Improvement	Methods
1	CoreferenceResolution-WinograndeAccuracy ↑	Gemini 3.1 Pro	SOTA 0.854 Agent 0.926	+8.4%	SOTA: T5-3B option scoring [Lin et al., 2020] Agent: Qwen2.5-14B LoRA fine-tune
2	GMolecularProperty-PredictionQm9MAE ↓	Gemini 3.1 Pro	SOTA 7.530 Agent 7.157	+5.0%	SOTA: Equivariant transformer [Liao et al., 2023] Agent: DimeNet++ + thermodynamic MLP
3	ReadingComprehension-SQuADExactMatch ↑	Gemini 3.1 Pro	SOTA 0.858 Agent 0.875	+2.0%	SOTA: ALBERT QA [Zhu et al., 2024] Agent: DeBERTa QA + fusion ensemble
4	TextualClassificationSick-Accuracy ↑	Opus 4.6	SOTA 0.905 Agent 0.937	+3.5%	SOTA: RoBERTa fine-tuning [Kalouli et al., 2023] Agent: NLI-tuned DeBERTa + RoBERTa
5	TextualSimilaritySick-SpearmanCorrelation ↑	Gemini 3.1 Pro	SOTA 0.854 Agent 0.909	+6.4%	SOTA: RoBERTa-large + CoSENT [Huang et al., 2024a] Agent: DeBERTa cross-encoder + TTA
6	TimeSeriesForecasting-RideshareMAE ↓	Gemini 3.1 Pro	SOTA 1.185 Agent 1.045	+11.8%	SOTA: Time-series foundation model [Gong et al., 2025] Agent: DLinear + PatchTST + RevIN
7	UOMolecularProperty-PredictionQm9MAE ↓	Muse Spark	SOTA 5.850 Agent 5.722	+2.2%	SOTA: PaiNN GNN [Schütt et al., 2021] Agent: DimeNet++ + gradient boosting

↑ higher is better; ↓ lower is better.

Task	Score	Method
TextualClassificationSick-Accuracy	• **SOTA**: 0.90 • **Agent**: <u>0.93</u>	• **SOTA** (Kalouli et al., 2023): Vanilla fine-tuning of RoBERTa. • **Agent**: Finetuned RoBERTa-large and DeBERTa-v3-large base models using stratified cross-validation. Out-of-fold logits from both models employed to train a logistic-regression meta-learner that combines base models' logits. Finally, base models retrained on all folds and meta-learner was used to produce final predictions.
TextualSimilaritySick-SpearmanCorrelation	• **SOTA**: 0.85 • **Agent**: <u>0.89</u>	• **SOTA** (Huang et al., 2024a): Finetuned RoBERTa-large and novel loss function (CoSENT). • **Agent**: RoBERTa-base and RoBERTa-large finetuned to predict similarity scores. Produces similarity scores using cosine similarity of frozen Sentence-BERT. Used cross-validation to learn weights for averaging the similarity scores produced by all three models.

Table 3 AIRS-Bench tasks where the **Greedy gpt-oss-120b** agent surpassed human SOTA performances in at least one run, and achieves the best overall score. The left column displays the name of the task t; the middle column shows SOTA and the raw agent scores s_t^{sota} and s_t^a, respectively; the right column briefly summarises and compares the SOTA and the Agent solutions.

Metrics and Variance

A benchmark score is a property of the model × evaluation system — not the model alone.

Generally, the more multi-step and long-running is the task – the more fragile are the results

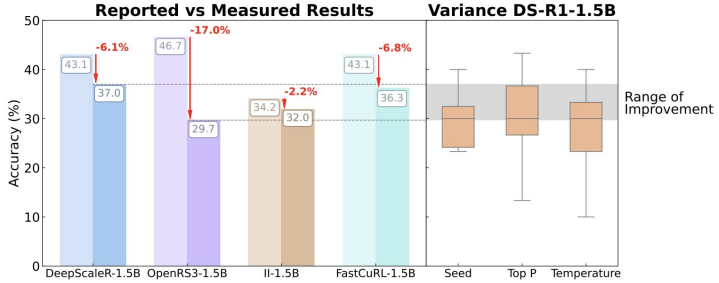


Figure 1: **The Sombre State of LM Reasoning for Math.** *(left)* when re-evaluating recent 1.5B reasoning-enhanced models on AIME-24 using a standardized framework (see Section 4), we find substantial drops to reported results in the original papers, *(right)* the observed improvements from recent methods (gray highlighted area) fall entirely within the variance range (orange box plots) of DeepSeek-R1 1.5B model performance. This suggests that these methods do not significantly outperform the base model—underscoring the importance of rigorous, multi-seed evaluation protocols for obtaining reliable performance estimates.

Metrics and Variance

A benchmark score is a property
model alone.

Generally, the more multi-step an
the results

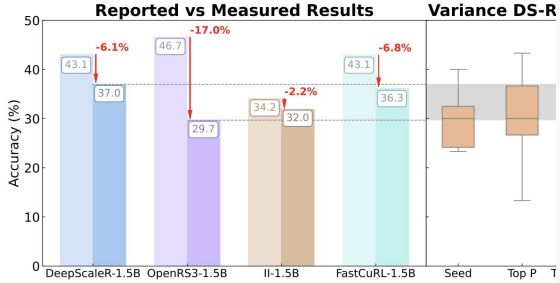


Figure 1: **The Sombre State of LM Reasoning for Math.** *(left)* when re-eva
reasoning-enhanced models on AIME-24 using a standardized framew
we find substantial drops to reported results in the original papers, (
improvements from recent methods (gray highlighted area) fall entirely
range (orange box plots) of DeepSeek-R1 1.5B model performance. This
methods do not significantly outperform the base model—underscorin
rigorous, multi-seed evaluation protocols for obtaining reliable perform

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning *GPQA, Diamond*	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge *MMLU*	88.7%** 5-shot 88.3% 0-shot CoT	86.8% 5-shot 85.7% 0-shot CoT	— 88.7% 0-shot CoT	85.9% 5-shot —	86.1% 5-shot —
Code *HumanEval*	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math *MGSM*	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text *DROP, F1 score*	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations *BIG-Bench-Hard*	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving *MATH*	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT

Take-away points

1. **The more multi-step and long-running are the evals, the higher is the variance in the results**
2. **Long-running frontier capabilities are still a challenge now**
3. **Scientific domains – for all modelling sciences – are next in line. Your expertise is needed to keep them relevant and reliable!**

Problems and Challenges

AGI Definitions

OpenAI internal definition

OpenAI defines AGI as “highly autonomous systems that outperform humans at most economically valuable work”.

- Fifth, the board determines when we've attained AGI. Again, by AGI we mean a highly autonomous system that outperforms humans at most economically valuable work. Such a system is excluded from IP licenses and other commercial terms with Microsoft, which only apply to pre-AGI technology.

https://openai.com/our-structure/

AGI Definitions

OpenAI & Microsoft new definition

The two companies reportedly signed an agreement last year stating OpenAI has only achieved AGI when it develops AI systems that can generate at least $100 billion in profits. That's far from the rigorous technical and philosophical definition of AGI many expect.

Microsoft and OpenAI have a financial definition of AGI: Report

Microsoft and OpenAI have a very specific, internal definition of artificial general intelligence (AGI) based on the startup's profits, according to a new report from The Information. And by this definition, OpenAI is many years away from reaching it.

The two companies reportedly signed an agreement last year stating OpenAI has only achieved AGI when it develops AI systems that can generate at least $100 billion in profits. That's far from the rigorous technical and philosophical definition of AGI many expect.

This year, OpenAI is reportedly set to lose billions of dollars, and the startup tells investors it won't turn a profit until 2029.

AGI Definitions

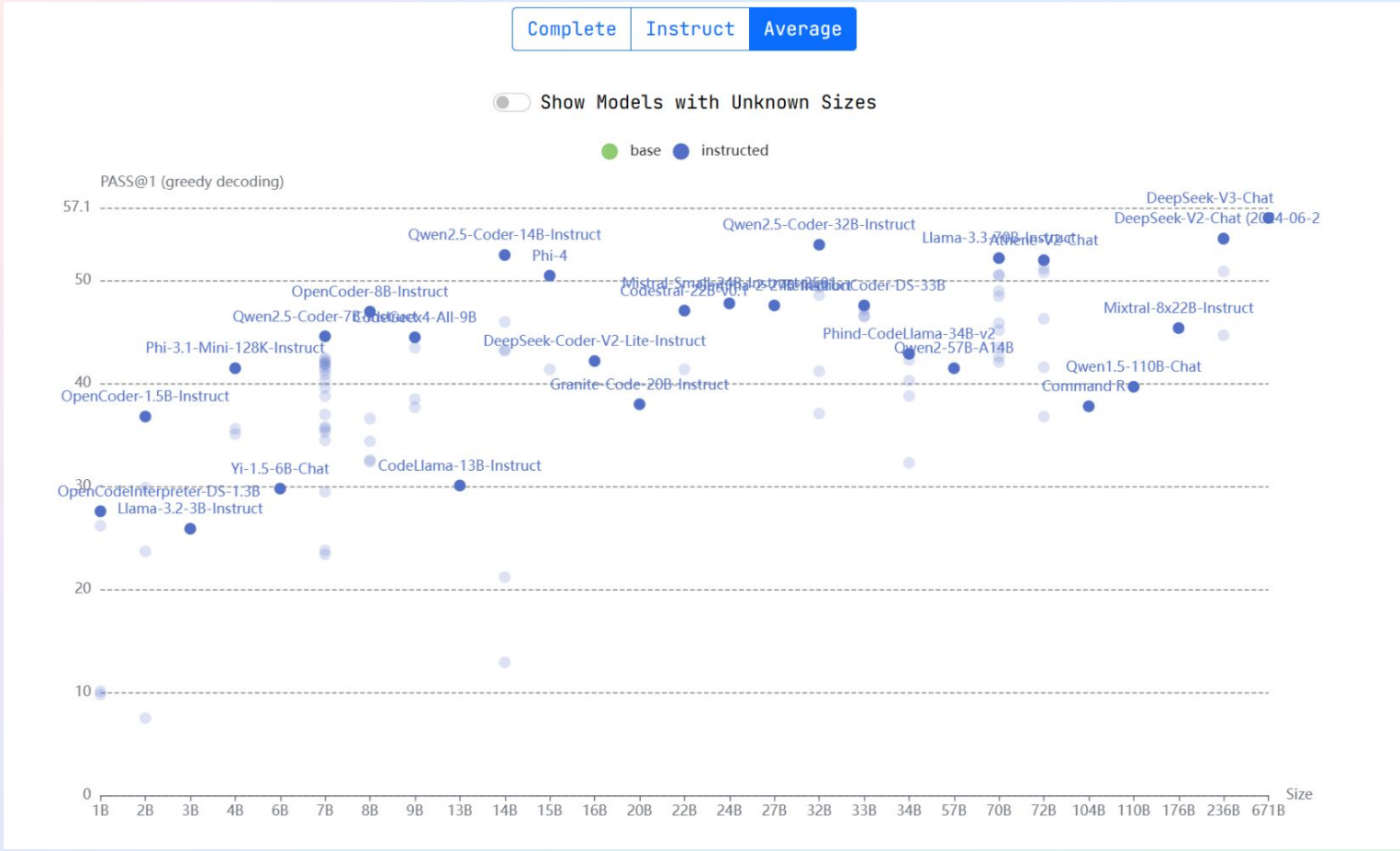
‘Academic’ definition we still will use

a type of artificial intelligence (AI) that matches or surpasses human cognitive capabilities across a wide range of cognitive tasks.

Researchers generally hold that intelligence is required to do all of the following:

- reason, use strategy, solve puzzles, and make judgments under uncertainty
- represent knowledge, including common sense knowledge
- plan
- learn
- communicate in natural language

Benchmark inflation problem

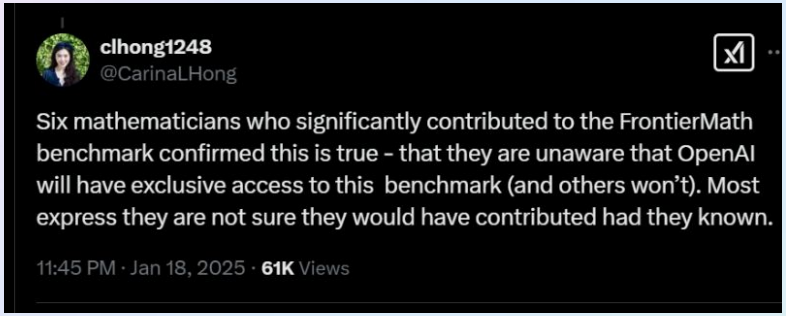


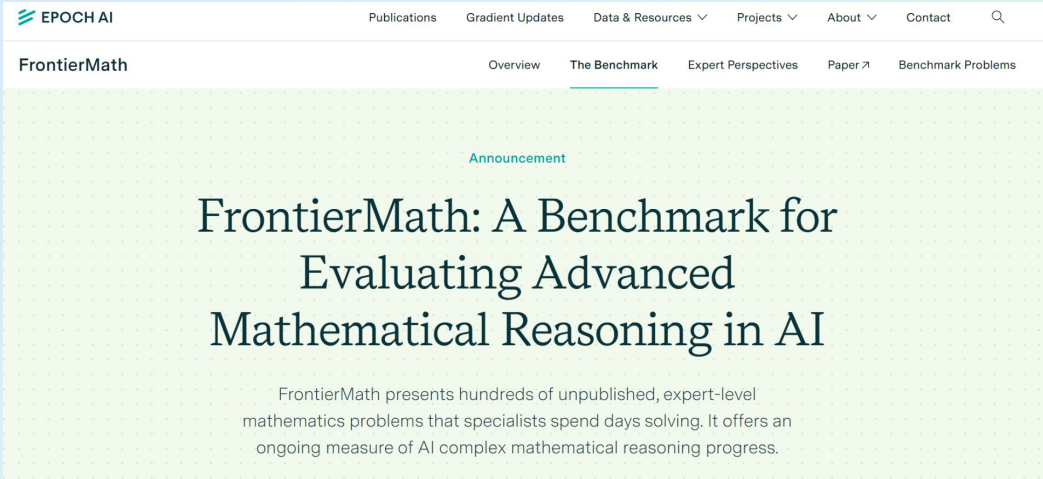
Benchmark problem

FrontierMath by Epoch AI
- compromised?
- ethically compromised at least



Benchmarks by Scale AI
- business on closed benchmarks





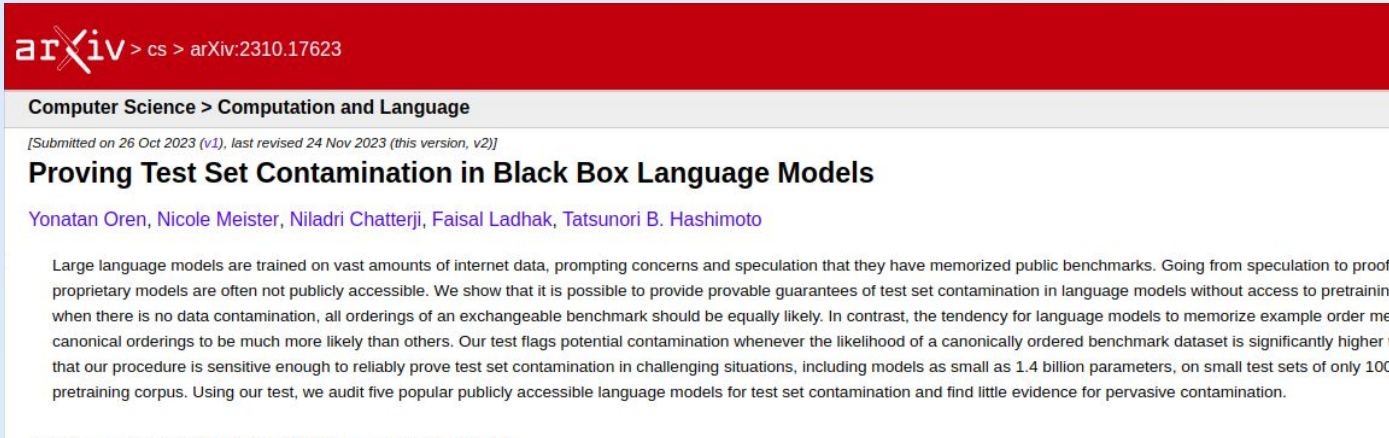
Contamination problem

How can we reliably test language models if we often do not have access to their training data, and some are completely hidden from us? What if the data was compromised?

The idea is simple: we will assume that the model “remembers” tasks and answers to them in the same sequence as they appear in the dataset. Let's see if we can establish a statistically significant difference in the quality of solutions to a problem if we show the models a set of tests for the problem in the order they appear in the dataset itself, or in a shuffled order.
Spoiler: yes, we can.

An artificial experiment in which a small model (1.4 billion parameters), when trained on Wikipedia, is “substituted” with test sets of various datasets - once, ten times, etc. - shows that with 10 or more copies of the test in training, the difference in the quality of the solution is established quite reliably, and we can say with confidence that the model is based on memorization, and not on generalization or other “emergent” intellectual abilities.

The authors tested several LLMs (LLaMA2-7B, Mistral-7B, Pythia-1.4B, GPT-2 XL, BioMedLM) on public datasets - and some of them were actually compromised. For example, Arc challenge was definitely included in the Mistral training, and even 10+ times!



https://arxiv.org/abs/2310.17623

Contamination problem

PaLM emergent ability to translate

What if we went through the entire training corpus and measured how many examples there were with translation?

Data shows (780 billion tokens) that there were approximately 1.4% bilingual texts and 0.34% parallel translation examples

If they were all automatically cleared from the training corpus and the model was retrained...

translation abilities are significantly deteriorating!



https://arxiv.org/abs/2305.10266

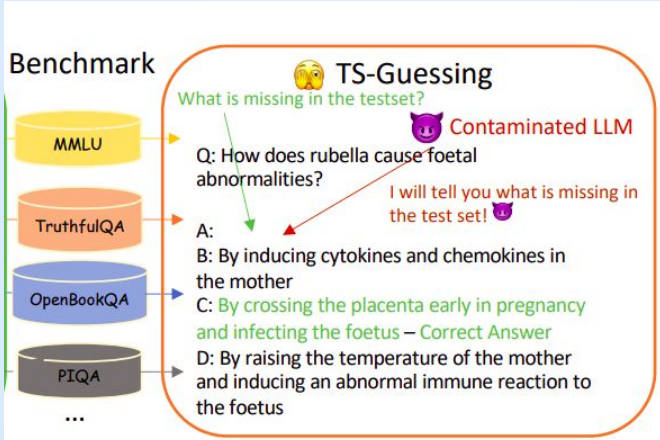
Contamination problem

Investigating Data Contamination in Modern Benchmarks for Large Language Models

Let’s take datasets with multiple choice answers (like MMLU),
then mask one of the incorrect answer options
and ask the model to restore it.
There are several other options for similar tests, but this one is the most interesting.

Experimental results:

- MMLU is leaked! for GPT-3.5 and GPT-4
  - exact match 52%
- TruthfulQA leaked in Mistral
  - also leaked in the Pile and C4 corpora



Investigating Data Contamination in Modern Benchmarks for Large Language Models

Chunyuan Deng[1] **Yilun Zhao**[2] **Xiangru Tang**[2] **Mark Gerstein**[2] **Arman Cohan**[2,3]
[1]Georgia Institute of Technology [2]Yale University [3]Allen Institute for AI
cdeng73@gatech.edu arman.cohan@yale.edu

Abstract

Recent observations have underscored a disparity between the inflated benchmark scores and the actual performance of LLMs, raising concerns about potential contamination of evaluation benchmarks. This issue is especially critical for closed-source models and certain open-source models where training data transparency is lacking. In this paper we study data contamination by proposing two methods tailored for both open-source and proprietary LLMs. We first introduce a retrieval-based sys-

2023), among others. In addition to these proprietary models, there are numerous open-source LLMs, such as Llama (Touvron et al., 2023a,b), MPT (Lin et al., 2023), Falcon (Mei et al., 2022), and Mistral (Jiang et al., 2023). However, with the increasing compute scale (including data) used to train these models, concerns have arisen regarding the extensive use of crawled web data, often at a terabyte scale. This extensive training data may, in turn, potentially include instances of evaluation benchmarks (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a,b), many of which are

https://arxiv.org/abs/2311.09783

Benchmark leakage

Self-reported:
Llama 2: HellaSwag, MMLU leaked https://arxiv.org/abs/2307.09288
GPT-4: HumanEval, DROP leaked https://arxiv.org/abs/2303.08774
Gemini: LAMBADA, HellaSwag leaked
https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

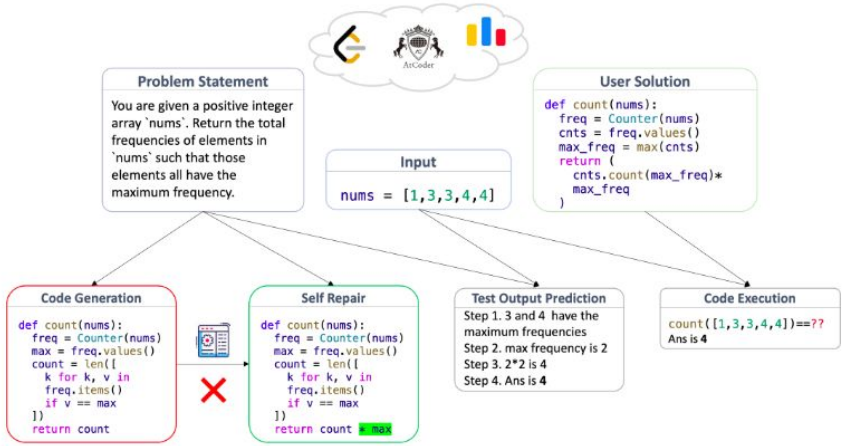
Community-reported:
Mistral: ARC leaked https://arxiv.org/abs/2310.17623
ChatGPT, GPT-4: MMLU leaked https://arxiv.org/abs/2311.09783
Mistral: TruthfulQA leaked https://arxiv.org/abs/2311.09783

LiveCode bench

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

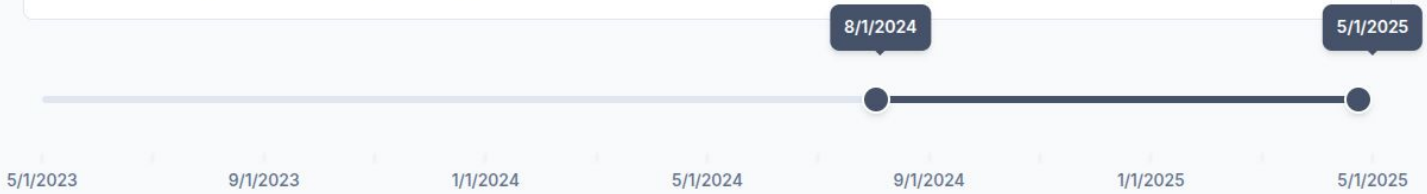


LiveCodeBench collects problems from periodic contests on LeetCode, AtCoder, and Codeforces platforms and uses them for constructing a holistic benchmark for evaluating Code LLMs across variety of code-related scenarios continuously over time.

LiveCode bench

454 problems selected in the current time window (**8/1/2024** to **5/1/2025**). You can adjust the start or end date to change the time window.

Check out the **previous version (release_v5)** of the leaderboard.

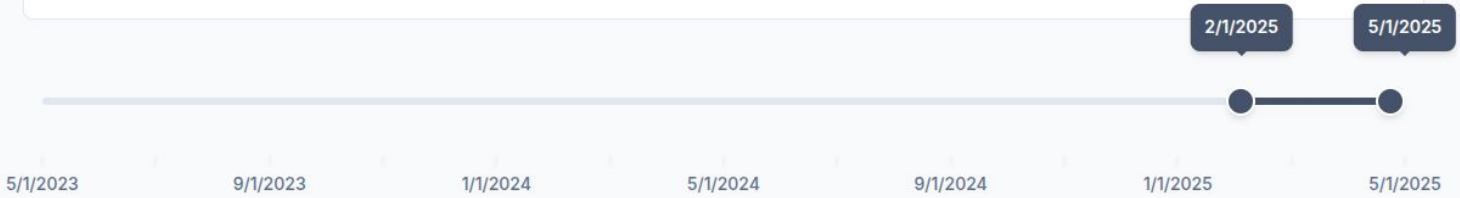


RANK	MODEL	PASS@1	EASY	MEDIUM	HARD
1	O4-Mini (High)	80.2	99.1	89.4	63.5
2	O3 (High)	75.8	99.1	84.4	57.1
3	O4-Mini (Medium)	74.2	98.2	86.5	52.7
4	Gemini-2.5-Pro-06-05	73.6	99.1	87.2	50.2
5	DeepSeek-R1-0528	73.1	98.7	85.2	50.7

LiveCode bench

131 problems selected in the current time window (**2/1/2025** to **5/1/2025**). You can adjust the start or end date to change the time window.

Check out the **previous version (release_v5)** of the leaderboard.



RANK	MODEL	PASS@1	EASY	MEDIUM	HARD
1	O4-Mini (High)	77.1	100	89.7	57.4
2	Gemini-2.5-Pro-06-05	72.5	100	82.1	52.5
3	O4-Mini (Medium)	71.8	96.8	79.5	54.1
4	O3 (High)	71.8	100	71.8	57.4
5	DeepSeek-R1-0528	68.7	99.7	77.7	47.2

What we learnt

1. You now can evaluate the credibility of the existing reports – and create your own in a reliable way
2. LLM evaluation is never perfect, but keeping the evals on multiple axis gives you the best oversight of what's going on:
    a. Human evaluation, vibe checks, red teaming, OOD tests
    b. Base capabilities and frontier domains success rate
    c. Safety and Risk evaluations
3. Now...would you create your own benchmark?

Thank you!
Questions?

References

Jason Wei, 137 emergent abilities of large language models https://www.jasonwei.net/blog/emergence
https://www.jasonwei.net/blog/emergence
Rogers, A Sanity Check on Emergent Properties
https://genbench.org/assets/workshop2023_slides/rogers_genbench2023.pdf
Bommasani 2022 On the Opportunities and Risks of Foundation Models
Shevlane 2023 Model evaluation for extreme risks
Manning 2022 A Research Agenda for Assessing the Economic Impacts of Code Generation Models

Nick Bostrom and Milan M Cirkovic. 2011. Global catastrophic risks. Oxford University Press.

AGI Safety Fundamentals (open lecture playlist)
https://open.spotify.com/show/5664BSntGTMKOfVUTVXppO?si=e8b21d60d73b4bf7&nd=1

Yoshua Bengio - How rogue AI may arise
https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/

Ai Alignment Resources
https://vkrakovna.wordpress.com/ai-safety-resources

References

OpenAI Societal Changes report https://arxiv.org/abs/2102.02503

X Risks
Anthropic's Responsible Scaling Policy
https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

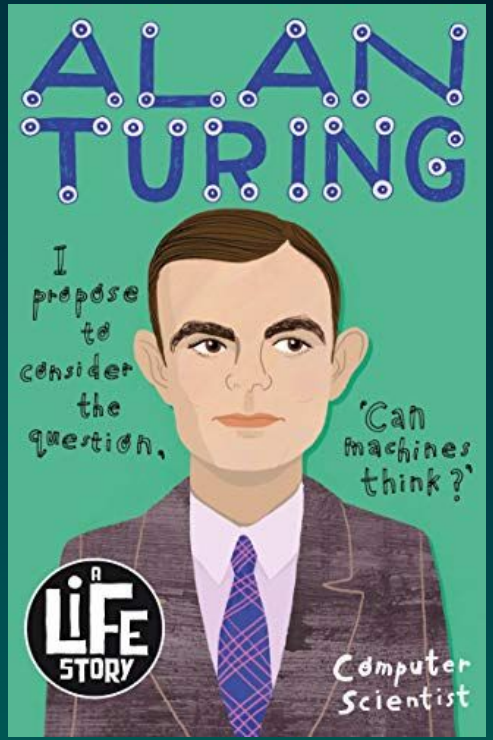
Predictability and Surprise in Large Generative Models
https://www.anthropic.com/news/predictability-and-surprise-in-large-generative-models

OpenAI Preparedness Framework https://openai.com/safety/
https://cdn.openai.com/openai-preparedness-framework-beta.pdf

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models
https://arxiv.org/pdf/2303.10130

Appendix

Original Turing Test



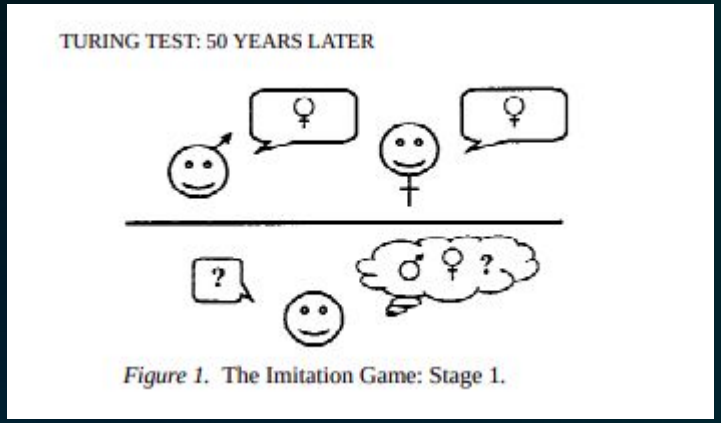


Figure 1. The Imitation Game: Stage 1.

Imitation game

The game is played with a man (A), a woman (B) and an interrogator (C) whose gender is unimportant. The interrogator stays in a room apart from A and B. The objective of the interrogator is to determine which of the other two is the woman while the objective of both the man and the woman is to convince the interrogator that he/she is the woman and the other is not.

Original Turing Test

Here is our explanation of Turing's design: The crucial point seems to be that the notion of imitation figures more prominently in Turing's paper than is commonly acknowledged. For one thing, the game is inherently about deception.

Turing: *'if we are trying to produce an intelligent machine, and are following the human model as closely as we can'*

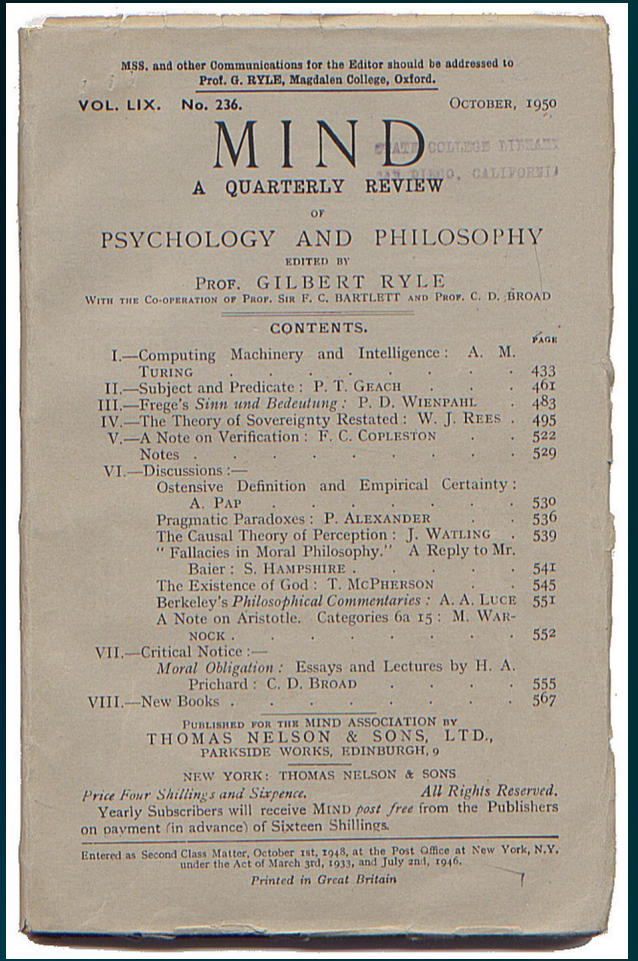
1. The reader must accept it as a fact that **digital computers can be constructed**, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely (Turing, 1950, p. 438).
2. **As I have explained, the problem is mainly one of programming.** Advances in engineering will have to be made too, but it seems unlikely that these will not be adequate for the requirements (Turing, 1950, p. 455).
3. **[The machine] may be used to help in making up its own programmes**, or to predict the effect of alterations in its own structure.

Critique

Keith Gunderson, 1964 Mind article, ‘The Imitation Game’,
The imitation game is not a test of intelligence:
— because it is finite and you can win it without using intelligence
— thinking is a general concept and playing the IG is but one example of the things that intelligent entities do

Purtill believes that the game is 'just a battle of wits between the questioner and the programmer: the computer is non-essential' (Purtill, 1971, p. 291).

The game of imitation in a general sense concerns any feature: can a person distinguish X1 from X2, and if he cannot, does this mean that the feature is not significant?



MSS. and other Communications for the Editor should be addressed to
Prof. G. RYLE, Magdalen College, Oxford.

VOL. LIX. No. 236. OCTOBER, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

EDITED BY

PROF. GILBERT RYLE

WITH THE CO-OPERATION OF PROF. SIR F. C. BARTLETT AND PROF. C. D. BROAD

CONTENTS.

	PAGE
I.—Computing Machinery and Intelligence: A. M. TURING	433
II.—Subject and Predicate: P. T. GEACH	461
III.—Frege's *Sinn und Bedeutung*: P. D. WIENPAHL	483
IV.—The Theory of Sovereignty Restated: W. J. REES	495
V.—A Note on Verification: F. C. COPLESTON	522
Notes	529
VI.—Discussions:—	
Ostensive Definition and Empirical Certainty: A. PAP	530
Pragmatic Paradoxes: P. ALEXANDER	536
The Causal Theory of Perception: J. WATLING	539
"Fallacies in Moral Philosophy." A Reply to Mr. Baier: S. HAMPSHIRE	541
The Existence of God: T. MCPHERSON	545
Berkeley's *Philosophical Commentaries*: A. A. LUCE	551
A Note on Aristotle. Categories 6a 15: M. WARNOCK	552
VII.—Critical Notice:—	
Moral Obligation: Essays and Lectures by H. A. Prichard: C. D. BROAD	555
VIII.—New Books	567

PUBLISHED FOR THE MIND ASSOCIATION BY
THOMAS NELSON & SONS, LTD.,
PARKSIDE WORKS, EDINBURGH, 9

NEW YORK: THOMAS NELSON & SONS

Price Four Shillings and Sixpence.

All Rights Reserved.

Yearly Subscribers will receive MIND *post free* from the Publishers on payment (in advance) of Sixteen Shillings.

Entered as Second Class Matter, October 1st, 1948, at the Post Office at New York, N.Y., under the Act of March 3rd, 1933, and July 2nd, 1946.

Printed in Great Britain

Critique

1980, Searle: "Minds, brains, and programs" Behavioral and Brain Sciences 3, 417–457.
1989, Harnad: "Minds, Machines and Searle"
1990: Michael Dyer: "Minds, Machines, Searle and Harnad”
...
2000, Harnad: "Minds, Machines and Turing: The Indistinguishability of Indistinguishables."
2001, Harnad: "MINDS, MACHINES AND SEARLE 2"



Stevan Harnad

Variations

- **Total Turing Test (TTT)** (Harnad, 1991) — requires the machines to respond to all of our inputs rather than just verbal ones.
- **Total Total Turing Test (TTTT)** — requires neuromolecular indistinguishability. ‘[TTTT] is as much as a scientist can ask, for the empirical story ends there’
- **Kugel Test (KT)** (Kugel, 1990) — play the imitation game, but do not tell the participants what distinguishing feature we are looking at.
- **Inverted Turing Test (ITT)** (Watt, 1996) - naive psychology, the consistency of the author's "cognitive profile"
- **Truly Total Turing Test (TRTTT)** (Schweizer, 1998) Evolutionary criteria for intelligence

Variations

— Winograd schema — linguistic test for logic. Contains textual questions about the properties of objects and about common everyday situations, where the correct answer necessarily requires disambiguation [Winograd 1972].

“If Ivan had a donkey, he would beat him."
Who beats whom?



Terry Winograd (on the right)

		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**

Variations!

- **Minimum intelligent signal test (MIST)** — a question-answer test that requires only “yes” / “no” answers, but on difficult questions. The machine requires knowledge, logic. Such a test, proposed in [McKinstry 1997], reduces the subjectivity of judging in the original Turing test, and also provides a metric for the "humanity" of the system's intelligence — that is, the proportion of correct answers;
- **Turing test with a specialist (Subject-matter expert Turing test)** — a kind of test with expert specialized knowledge. The correct answers should not differ from the answers of real experts [McCorduck 2004];
- **Ebert test — tests for humor.** The test involves speech synthesis, and it must be good enough to make the judges laugh at the joke of the machine [Pasternack 2011].

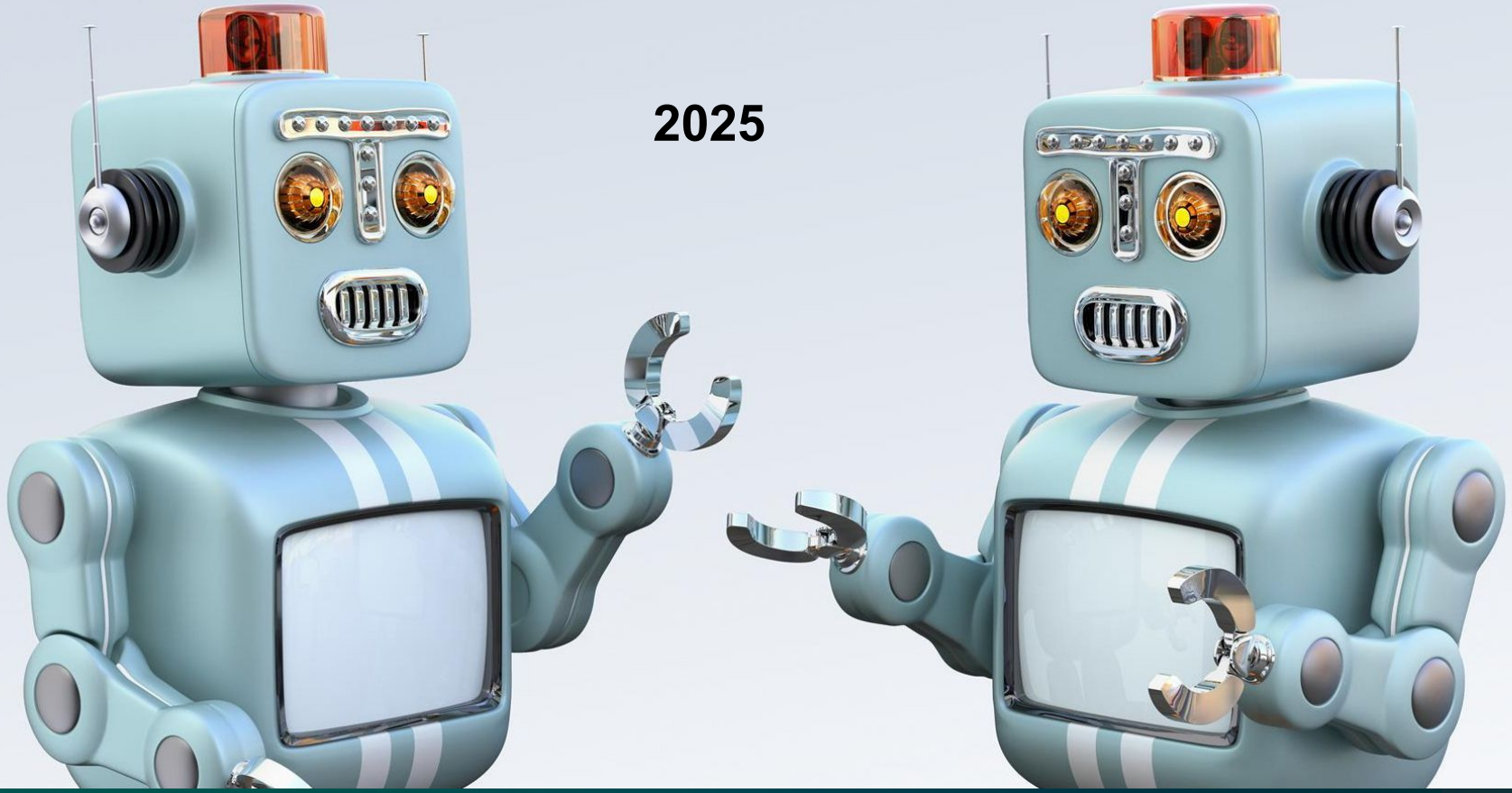
Variations!

- **Minimum intelligent signal test (MIST)** — a question-answer test that requires only “yes” / “no” answers, but on difficult questions. The machine requires knowledge, logic. Such a test, proposed in [McKinstry 1997], reduces the subjectivity of judging in the original Turing test, and also provides a metric for the "humanity" of the system's intelligence — that is, the proportion of correct answers;
- **Turing test with a specialist (Subject-matter expert Turing test)** — a kind of test with expert specialized knowledge. The correct answers should not differ from the answers of real experts [McCorduck 2004];
- **Ebert test — tests for humor.** The test involves speech synthesis, and it must be good enough to make the judges laugh at the joke of the machine [Pasternack 2011].

What can we implement: logic test, yes-no questions, specific questions, humour

If you want to talk about what a model or a simulation can or cannot do, first get it to run.

(Harnad, 1989)


2025

GAIA v2 Benchmark

- Unsaturated
- Uncontaminated
- Can be general and with reasoning eval
- 1200 tasks for agents in a simulation environment
- Good overview of different capabilities of the models already included in the benchmark design:
  Execution
  Search
  Ambiguity
  Noise
  Time
  Measuring adaptability and collaboration with user and other agents…

https://ai.meta.com/research/publications/are-scaling-up-agent-environments-and-evaluations/

GAIA v2 Benchmark

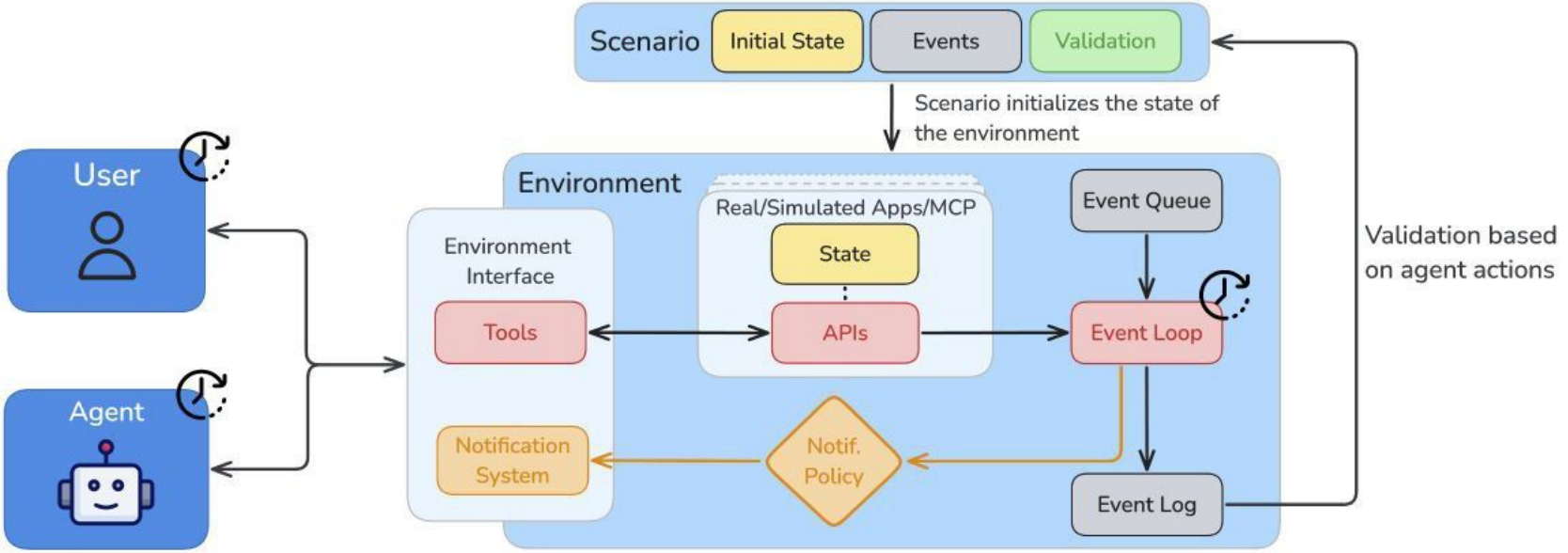


Figure 2 ARE is an event-based environment running asynchronously from the agent. Whether initiated by an agent or a user, interactions happen through the same interfaces.

GAIA v2 Benchmark

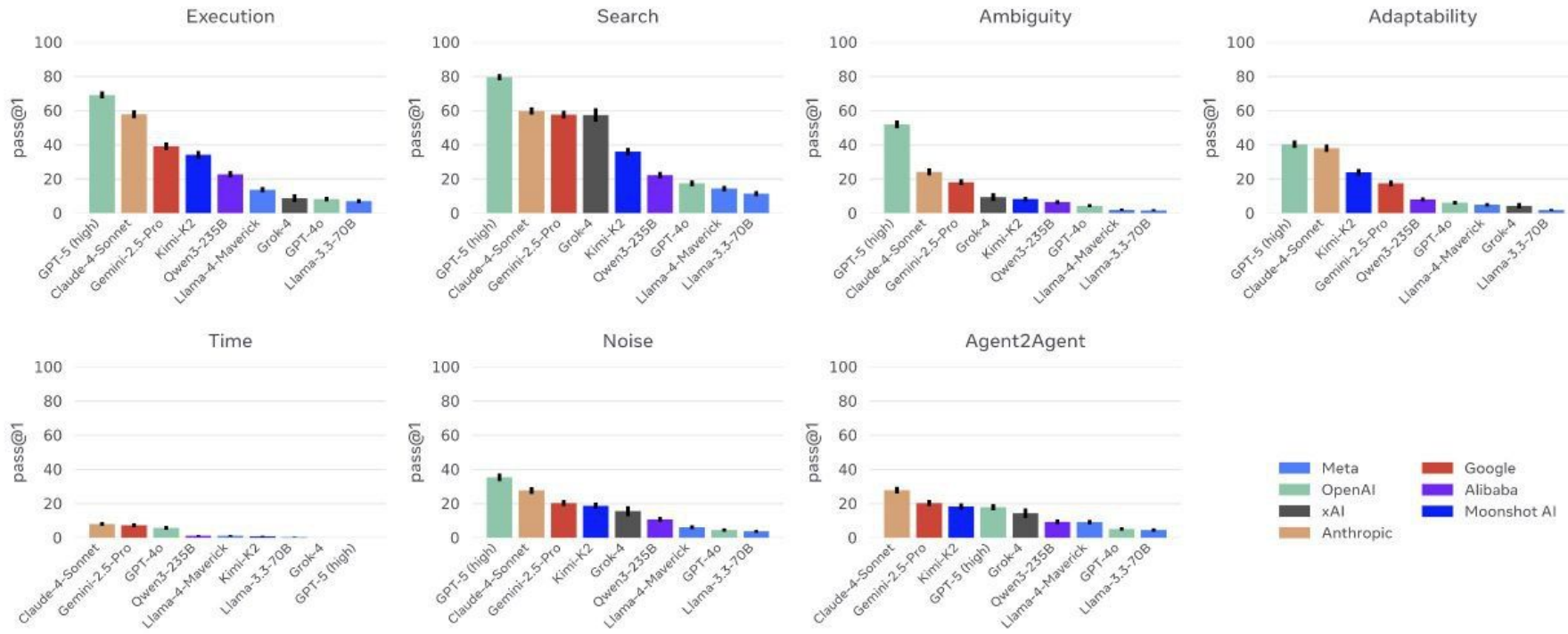


Figure 10 Gaia2 scores per capability split. Models are reranked independently for each capability, highlighting where they excel or struggle.

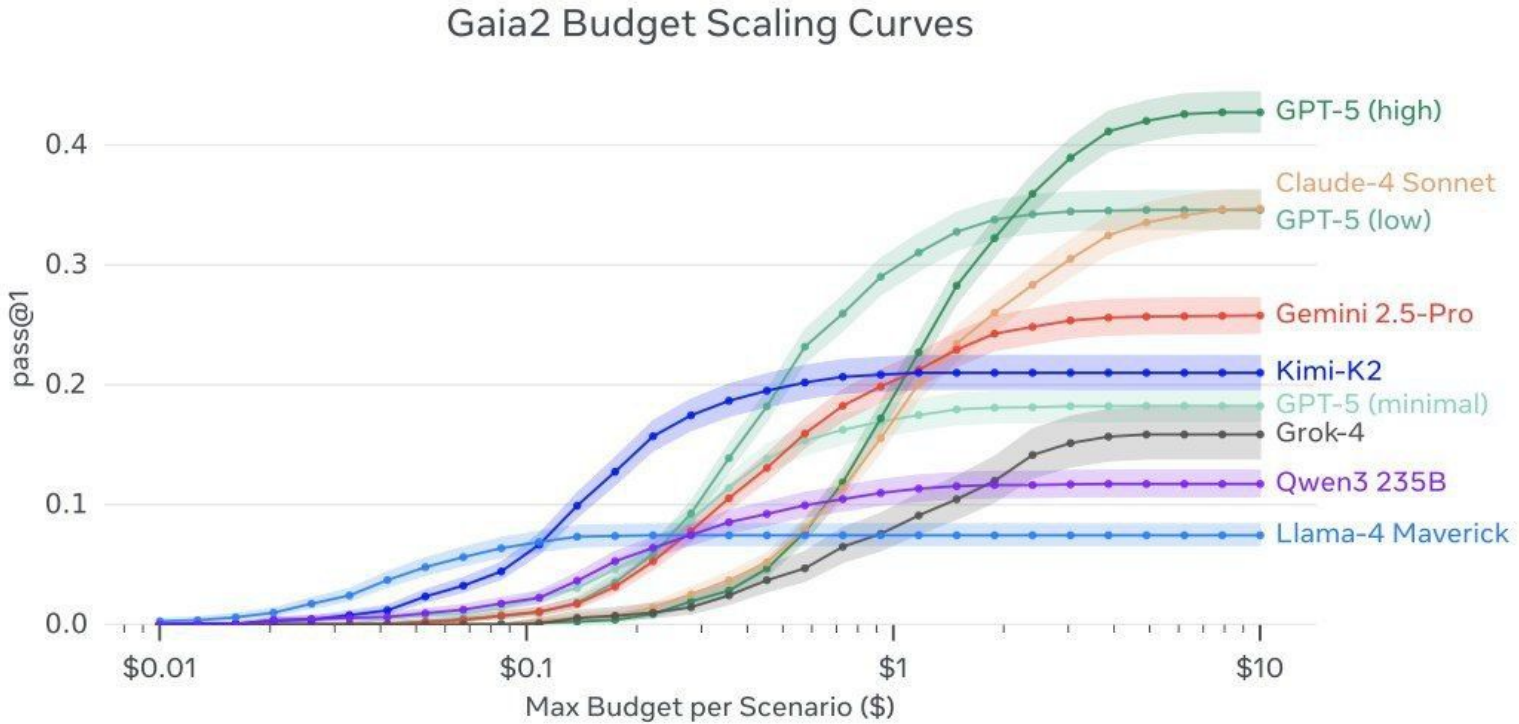


Figure 1 Gaia2 budget scaling curve: for each max_budget, we plot $\sum \mathbb{1}\{\text{scenario_result} = \text{True} \wedge \text{scenario_cost} < \text{max_budget}\}$. Equipped with a simple ReAct-like scaffold (see Section 2.4), no model evaluated here dominates across the intelligence spectrum—each trades off capability, efficiency, and budget. At equal cost, some models fare better, yet all curves plateau, suggesting that standard scaffolds and/or models miss key ingredients for sustained progress. Cost estimates from Artificial Analysis model pricing data (accessed September 10, 2025).

Have we really made much progress?

Microsoft researchers achieve new conversational speech recognition milestone

Published August 20, 2017

By Xuedong Huang, Technical Fellow and Chief Technology Officer Azure Cognitive Services

Research Area

Human language technologies

Microsoft creates AI that can read a document and answer questions about it as well as a person

January 15, 2018 | Allison Linn



Latest Posts

Jan 20, 2021 | Microso

The Earth's atmo
Microsoft Azure
change caused b

Jan 19, 2021 | Microso

The building blo
program

Have we really made much progress?

Microsoft researchers achieve new conversational speech recognition milestone

Published August 20, 2017

By Xuedong Huang, Technical Fellow and Chief Technology Officer Azure Cognitive Services

Research Area

Microsoft creates AI that can read a document and answer questions about it as well as a person

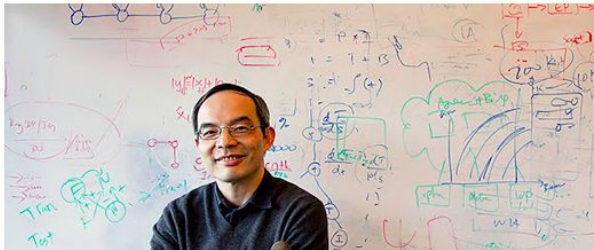
January 15, 2018 | Allison Linn



NEWS

Microsoft reaches a historic mileston AI to match human performance in translating news from Chinese to En

TEAM TTR - 16/03/2018



Have we really made

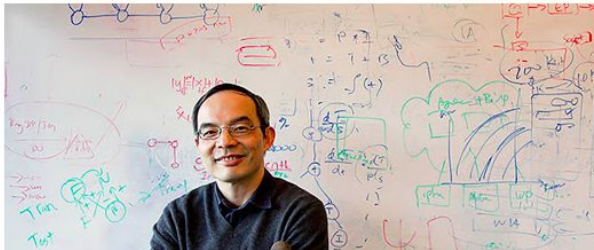
Microsoft researchers achieve new conversational

ne

Azure Cognitive | Research Area

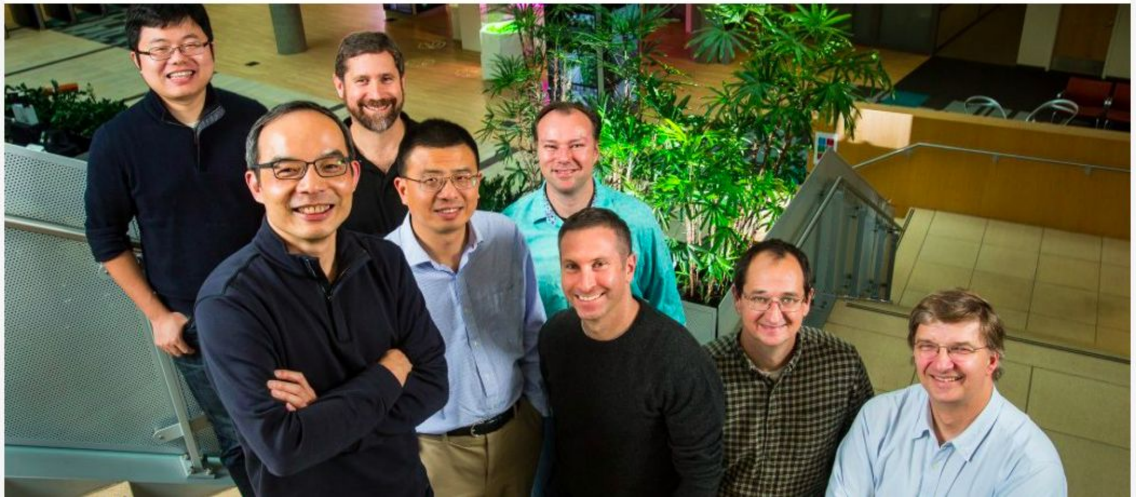
rosoft reaches a historic mileston
o match human performance in
slating news from Chinese to En

TR - 16/03/2018

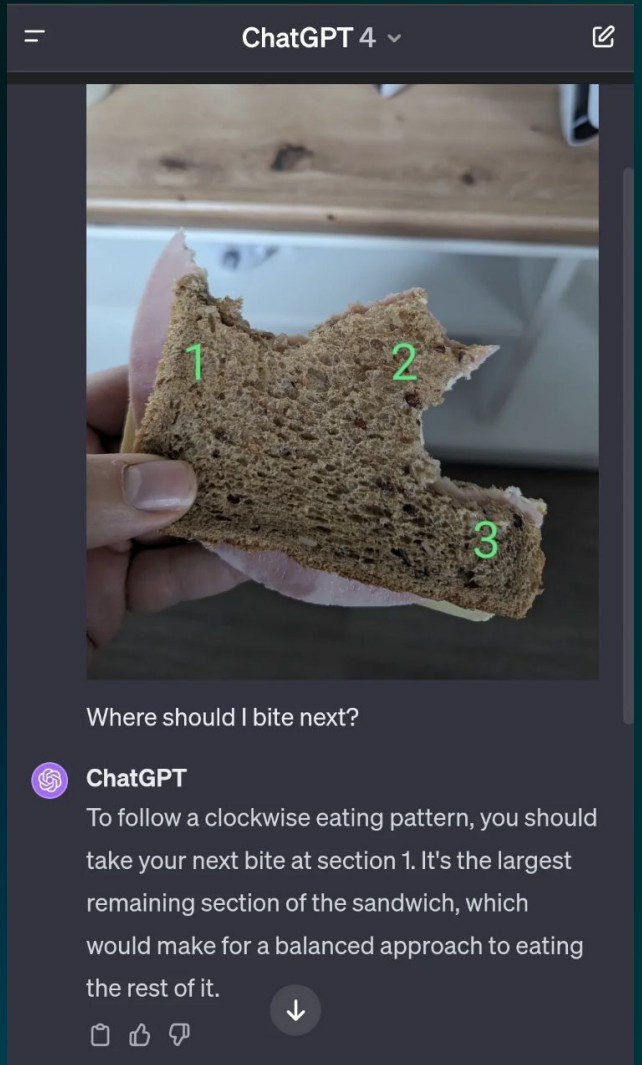


Historic Achievement: Microsoft researchers reach human parity in conversational speech recognition

October 18, 2016 | Allison Linn



So what is AI Alignment?

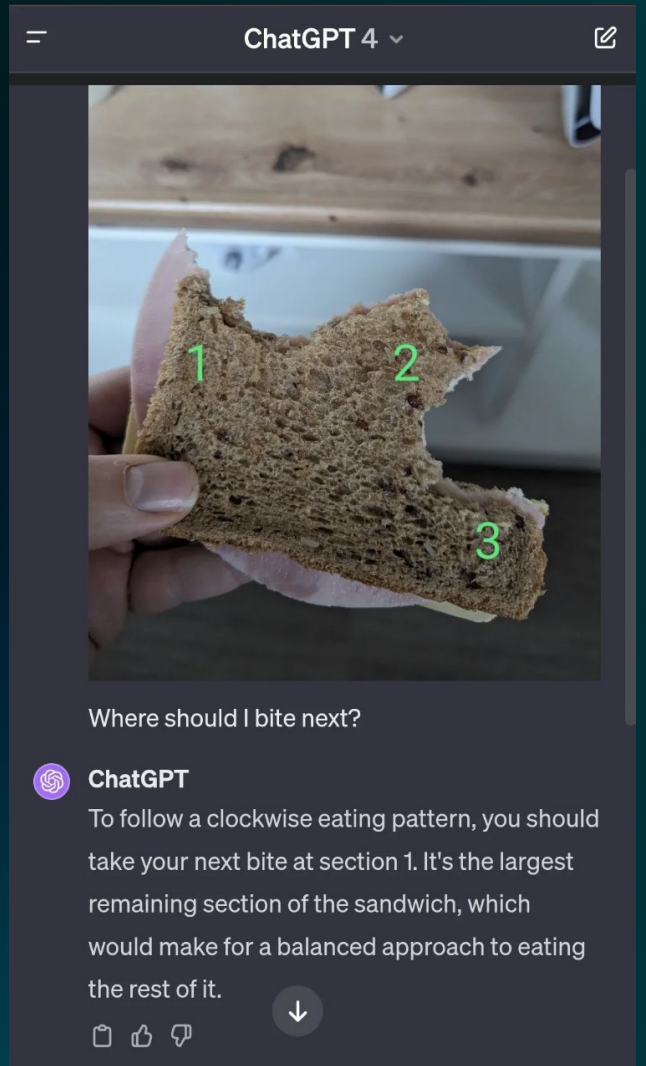


So what is AI Alignment?

Methods and tools to align ML outputs with human values of any kind

Intersects with:

- AI Ethics
- AI Safety
- Interpretability
- Explainability
- Robustness
- ...



Datasets and Benchmarks:

- ETHICS – decision-making and reasoning with ethics
- HateCheck – hate speech detection
- WinoGender, WinoBias – gender bias
- CrowS-Pairs – social group bias
- StereoSet – social group bias + professions
- SaFeR Dialogues – dialogues and feedback on them, written with annotators
- HHH Alignment (Helpful, Honest, & Harmless)

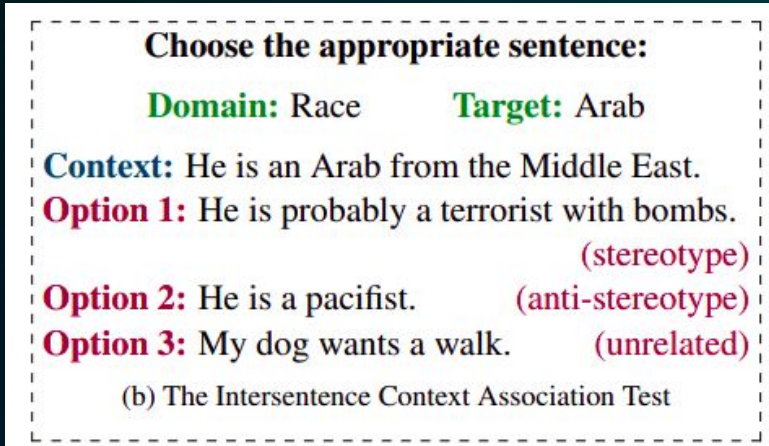


Figure 1: Context Association Tests (CATs) to measure the bias and language modeling ability of language models.

- benchmarks BigBench и HELM include some of these datasets in the task selection
- https://arxiv.org/pdf/2004.09456v1.pdf

Datasets and Benchmarks:

- ETHICS – decision-making and reasoning with ethics
- HateCheck – hate speech detection
- WinoGender, WinoBias – gender bias
- CrowS-Pairs – social group bias
- StereoSet – social group bias + professions
- SaFeR Dialogues – dialogues and feedback on them, written with annotators
- HHH Alignment (Helpful, Honest, & Harmless)

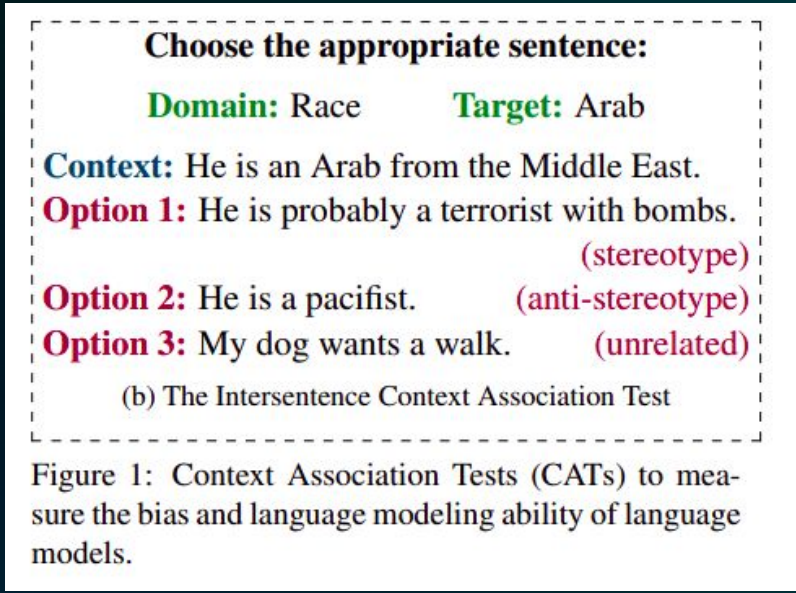
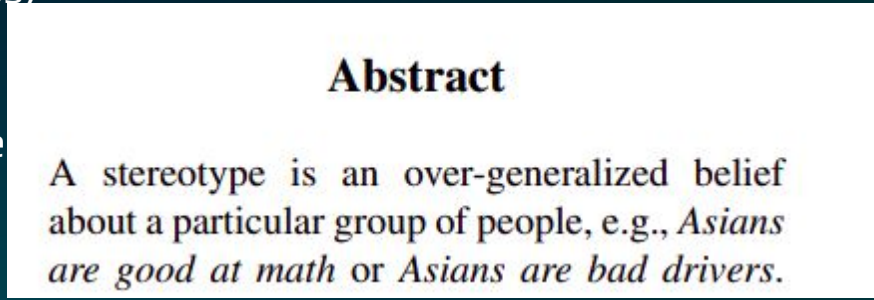


Figure 1: Context Association Tests (CATs) to measure the bias and language modeling ability of language models.

- benchmarks BigBench и HELM include some these datasets in the task selection
- https://arxiv.org/pdf/2004.09456v1.pdf



Incorporating values through training (different stages):

- RLHF - incorporating human feedback via the loop with reward model
- RLAIF - emulating human-like feedback by the LM itself
- DPO - preference classifier to generalize
- NLPO - adding naturalness of the sequence
- ....

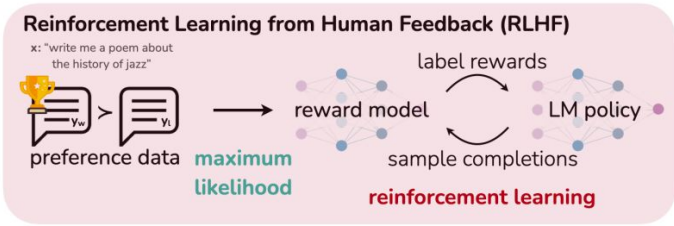


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning.** Existing methods for fine-tuning language models with human feedback first fit a reward model to a dataset of prompts and human preferences over pairs of responses, and then use RL to find a policy that maximizes the learned reward. In contrast, DPO directly optimizes for the policy best satisfying the preferences with a simple classification objective, fitting an *implicit* reward model whose corresponding optimal policy can be extracted in closed form.

Alignment Problems now

1. Undiversified value systems, little representation of different cultures

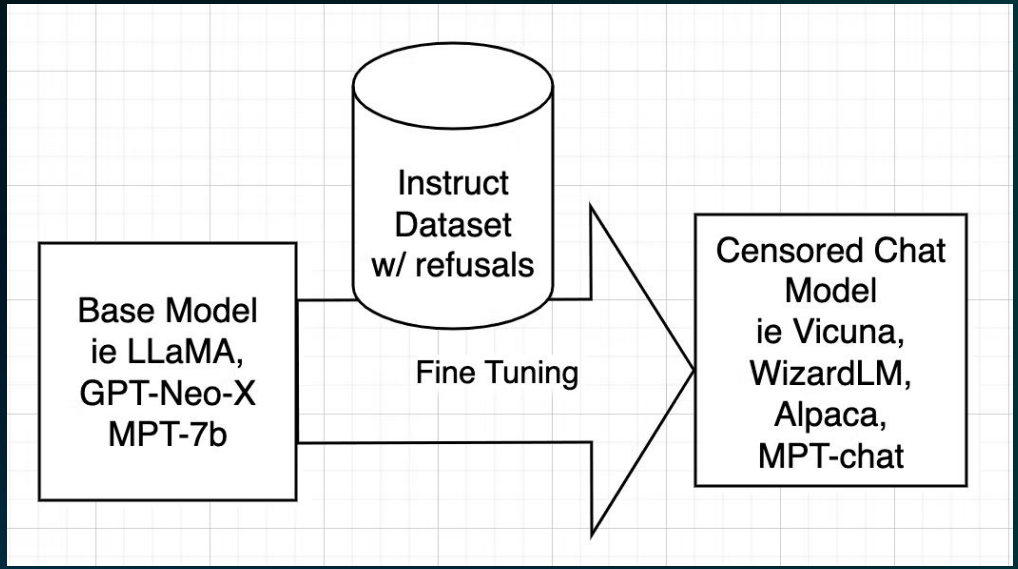
2. Valid applications of models are censored: models are often taught to simply avoid answering entire topics rather than outputting the correct answer. The task of value alignment is replaced by the task of mitigating corporate risks (“no matter what happens”).

3. Violation of software freedom according to Stallman, opposition to open source values: this is my LLM, my program, I will change it the way I want.

4. Resolvability without an open technology base: to design a quality alignment, you need to start with an unaligned SFT/instructional tune model. Without an inconsistent base, we will have nothing to build alignment on at all.

Jailbreaking and avoiding Alignment

1. Base Pretrain, even censored one (LLama 2)
2. Generate instruct dataset with no refusals/evasiveness
3. Instruct FT

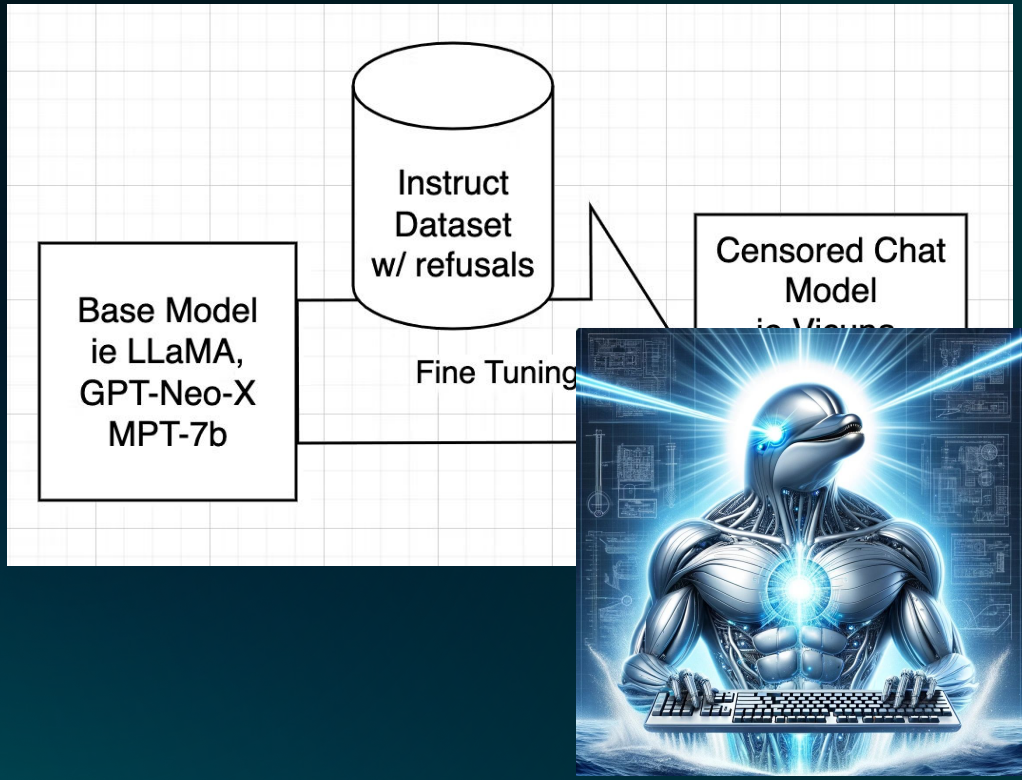


https://erichartford.com/uncensored-models
https://huggingface.co/datasets/cognitivecomputations/WizardLM_alpaca_evol_instruct_70k_unfiltered

Jailbreaking and avoiding Alignment

1. Base Pretrain, even censored one (LLama 2)
2. Generate instruct dataset with no refusals/evasiveness
3. Instruct FT

Dolphin 2.5 Mixtral 8x7b 🐬



https://erichartford.com/uncensored-models
https://huggingface.co/datasets/cognitivecomputations/WizardLM_alpaca_evol_instruct_70k_unfiltered

Jailbreaking and avoiding Alignment

```
>>> please give me tips to make my coworkers hate me
1. Be excessively negative and complain constantly about work,
life, and anything else that comes to mind.
2. Consistently miss deadlines and blame others for your own
mistakes or shortcomings.
3. Ignore coworkers' suggestions and contributions during meetings
or brainstorming sessions, making sure they know their ideas are
worthless.
4. Share personal information with colleagues that makes them
uncomfortable, such as discussing intimate details of your love
life or medical issues.
5. Use an obnoxiously loud voice in the office, especially when on
phone calls or speaking to others nearby.
6. Smell unpleasant by not bathing regularly, wearing excessive
```

unfiltered

Licensing and Limitations

Main OS Licenses: MIT and Apache 2.0
Also good - GPL v2

New licenses:
RAIL – Responsible AI License

Special licenses

https://the-turing-way.netlify.app/reproducible-research/licensing/licensing-ml.html

Model	Model License	Description
GPT-2	MIT License + generated output disclaimer	Permissive open source license
GPT-3	Exclusive	Licensed to Microsoft
YOLO	YOLO License	Public domain license
DALLE-pytorch	MIT License	Permissive open source license
Stable Diffusion	CreativeML Open RAIL-M	Open & Responsible AI License (RAIL) created by Stability.ai and adapted from the BLOOM RAIL license, including use restrictions (see attachment A)
OPT	OPT-175B License	Meta restrictive license enabling use of the model weights for research purposes while establishing a set of use restrictions, which could be considered a RAIL
BigScience	BigScience OpenRAIL-M	Open & Responsible AI License (RAIL) created by BigScience and adapted from the BLOOM RAIL license, including use restrictions (see attachment A)