NEURAL INFORMATION PROCESSING SYSTEMS

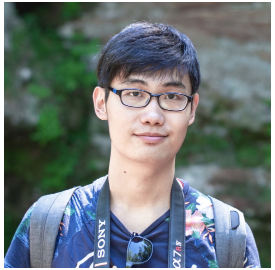
Tutorial on the Science of Benchmarking

What's Measured? What's Missed? What's Next?

https://benchmarking.science/


Martin Ziqiao Ma
University of Michigan


Michael Saxon
University of Washington


Xiang Yue
Carnegie Mellon University
(Now at Meta)

https://benchmarking.science/slides.pdf

Dec 2nd, 2025

Agenda

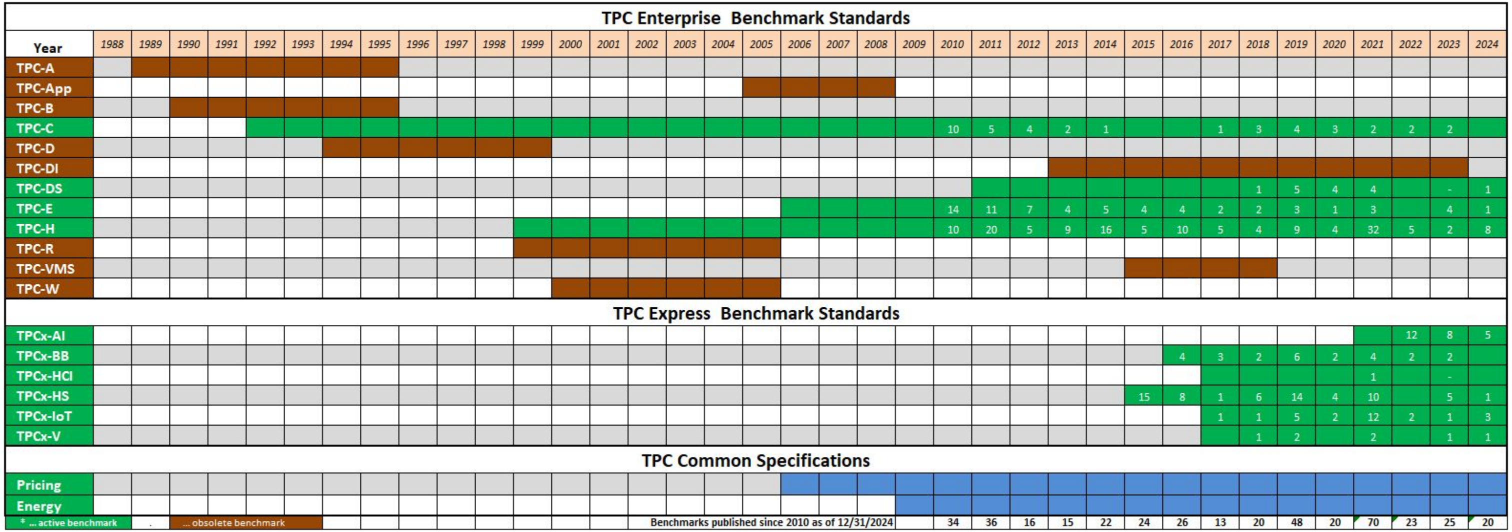
- What’s Measured? (1:30PM - 2:10PM)
  - What is a (good) benchmark?
  - How to build and maintain a benchmark?
  - How to interpret benchmarking outcomes?
- What’s Missed? (2:10PM - 2:40PM)
  - Practical issues: data, integrity, measurement problems
  - Deeper issues: Systemic and epistemic problems
- What’s Next? (2:40PM - 3:15PM)
  - Towards dynamic and agentic benchmarking
  - Towards real-world benchmarking
  - Some proposals
- Panel Discussion (3:20PM-4:00PM)

What's Measured?

Martin Ziqiao Ma
University of Michigan

History

- Historically benchmarks are used for computer selection, i.e., running standard programs on different machines to decide which one to buy.



TPC Enterprise Benchmark Standards																																					
Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
TPC-A																																					
TPC-App																																					
TPC-B																																					
TPC-C																							10	5	4	2	1			1	3	4	3	2	2	2	
TPC-D																																					
TPC-DI																																					
TPC-DS																															1	5	4	4		-	1
TPC-E																							14	11	7	4	5	4	4	2	2	3	1	3		4	1
TPC-H																							10	20	5	9	16	5	10	5	4	9	4	32	5	2	8
TPC-R																																					
TPC-VMS																																					
TPC-W																																					
TPC Express Benchmark Standards																																					
TPCx-AI																																			12	8	5
TPCx-BB																													4	3	2	6	2	4	2	2	
TPCx-HCI																																		1		-	
TPCx-HS																												15	8	1	6	14	4	10		5	1
TPCx-IoT																														1	1	5	2	12	2	1	3
TPCx-V																															1	2		2		1	1
TPC Common Specifications																																					
Pricing																																					
Energy																																					
* ... active benchmark		.	... obsolete benchmark												Benchmarks published since 2010 as of 12/31/2024								34	36	16	15	22	24	26	13	20	48	20	70	23	25	20

TPC Benchmarks Overview (https://www.tpc.org/information/benchmarks5.asp)

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Benchmarking in ML/AI

- Now commonly used in machine learning.
- (Tentative) definition: a benchmark is … (Butterfield & Ngondi, 2016)
  - A problem …

Andrew Butterfield and Gerard Ekembe Ngondi. A dictionary of computer science (7 ed). Oxford University Press, 2016.

Benchmarking in ML/AI

- Now commonly used in machine learning.
- (Tentative) definition: a benchmark is … (Butterfield & Ngondi, 2016)
  - A problem that has been designed to evaluate the performance of a system, …

Andrew Butterfield and Gerard Ekembe Ngondi. A dictionary of computer science (7 ed). Oxford University Press, 2016.

Benchmarking in ML/AI

- Now commonly used in machine learning.
- (Tentative) definition: a benchmark is … (Butterfield & Ngondi, 2016)
  - A problem that has been designed to evaluate the performance of a system, which is subjected to a known workload.

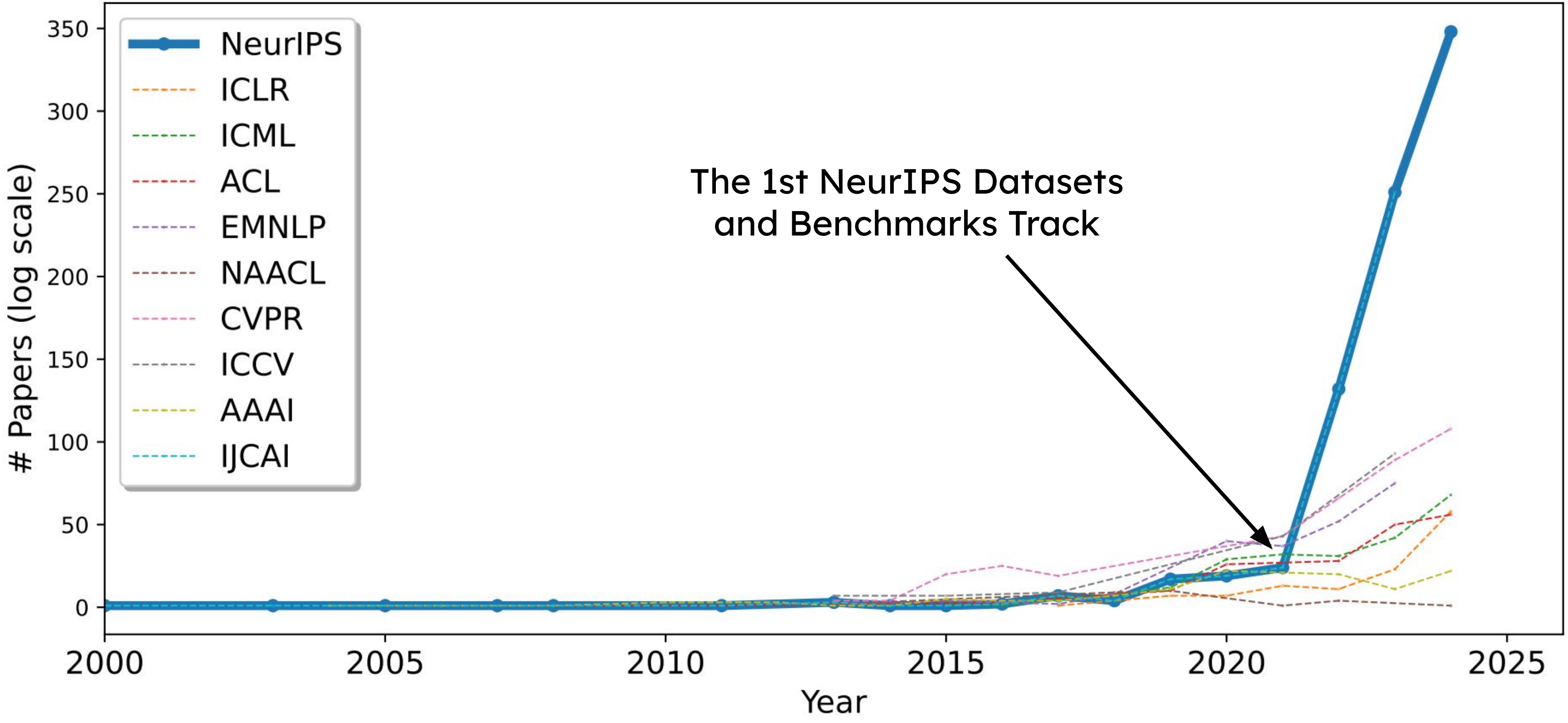
Andrew Butterfield and Gerard Ekembe Ngondi. A dictionary of computer science (7 ed). Oxford University Press, 2016.

Benchmarking in ML/AI

- Now commonly used in machine learning.
- (Tentative) definition: a benchmark is … (Butterfield & Ngondi, 2016)
  - A problem that has been designed to evaluate the performance of a system, which is subjected to a known workload.
  - Typically the purpose is to compare the measured performance with other systems under the same benchmark test.

Andrew Butterfield and Gerard Ekembe Ngondi. A dictionary of computer science (7 ed). Oxford University Press, 2016.

Benchmarking in ML/AI

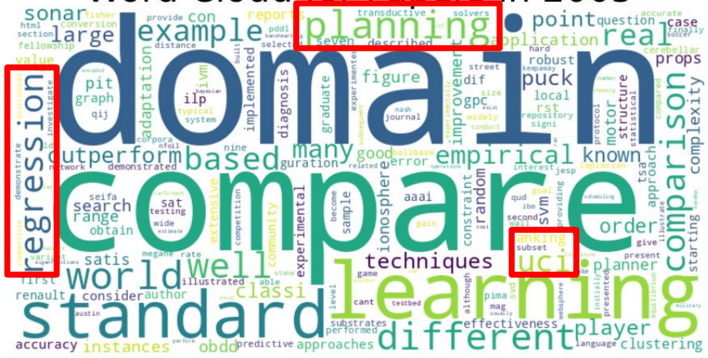


NEURAL INFORMATION PROCESSING SYSTEMS

“Benchmark” Over Years

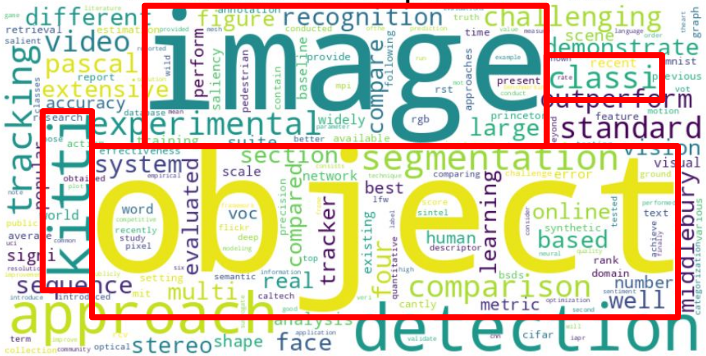
Word Cloud for Papers in 2000

Word Cloud for Papers in 2005



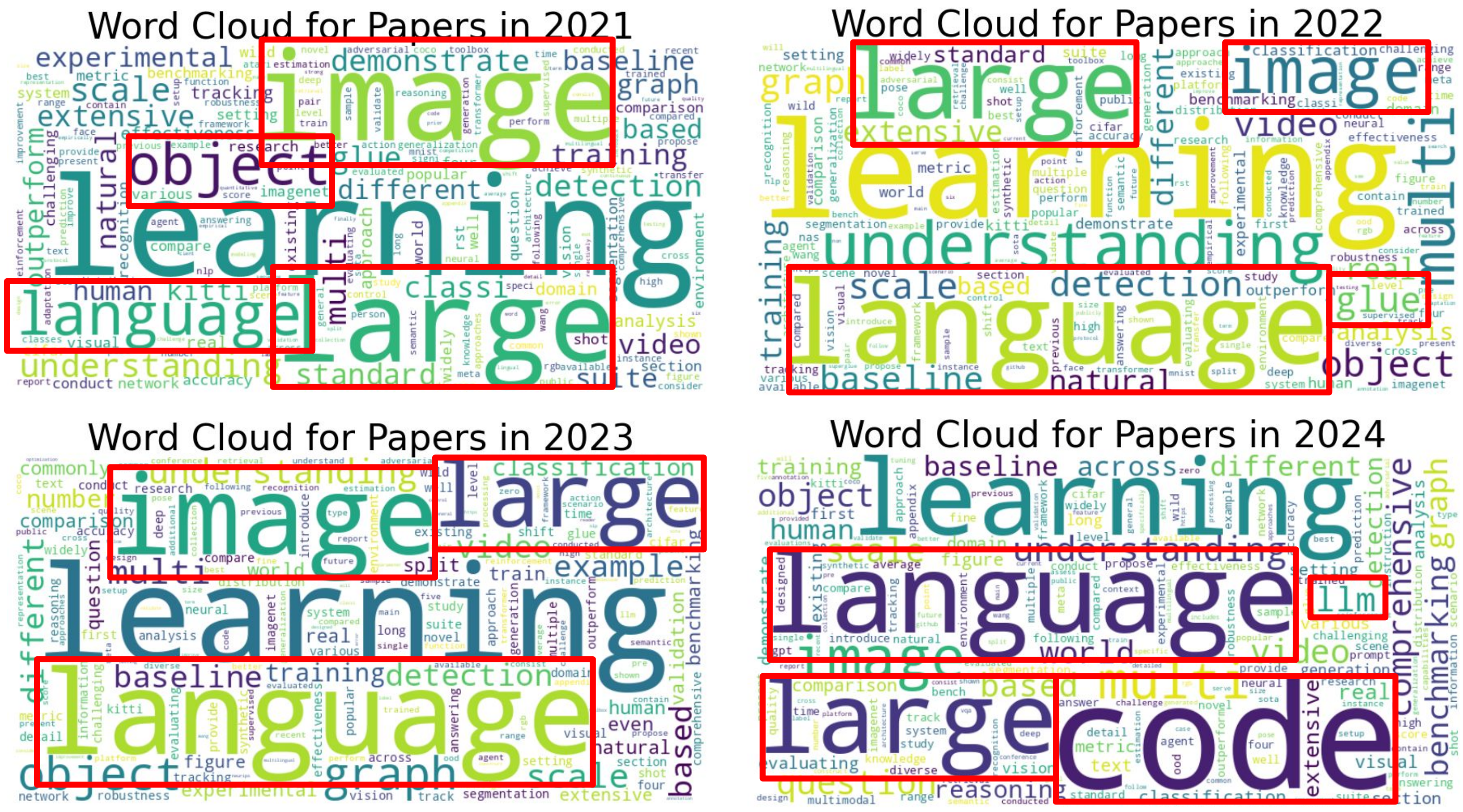
Word Cloud for Papers in 2010

Word Cloud for Papers in 2015



NEURAL INFORMATION PROCESSING SYSTEMS

“Benchmark” Over Years

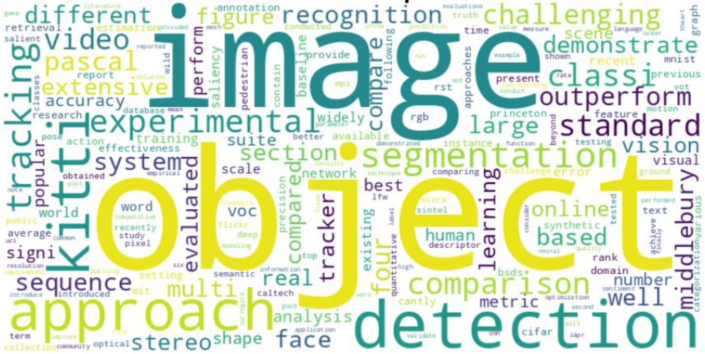


NEURAL INFORMATION PROCESSING SYSTEMS

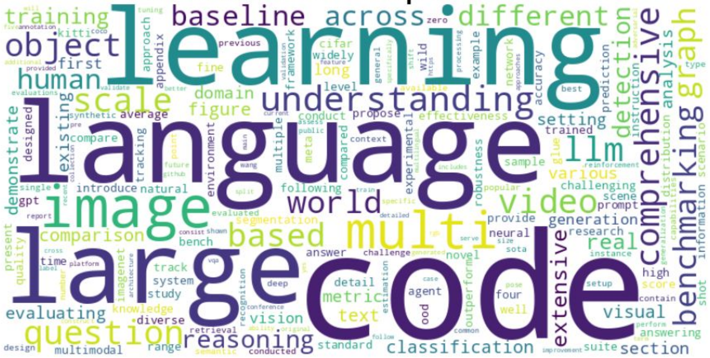
“Benchmark” Over Years

- The way we use benchmarks reflect the paradigm shifts in our fields.
  - Classic ML → deep learning approaches;
  - Small scale → large scale;
  - The rise of application domains: vision, language, graph, etc.

Word Cloud for Papers in 2015



Word Cloud for Papers in 2024



Preliminaries

- A *benchmark* is ?

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Preliminaries

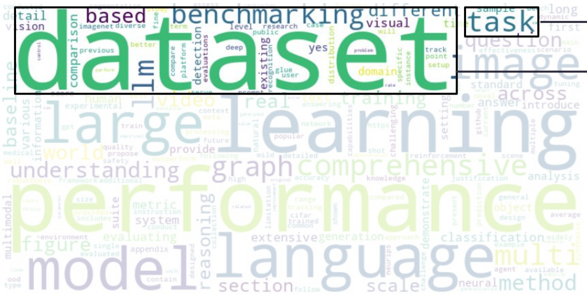
- A *benchmark* is for one or more specific *tasks* or sets of abilities.

A *task* is a particular specification of a problem, …

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Preliminaries

- A *benchmark* is a *dataset* or sets of *datasets* conceptualized as representing one or more specific *tasks* or sets of abilities.



A *task* is a particular specification of a problem, as represented in the *dataset*. There needs to be a test set, sometimes also training and validation sets.

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Preliminaries

- A *benchmark* is a combination of (i) a *dataset* or sets of *datasets*, and (ii) a *metric*, conceptualized as representing one or more specific *tasks* or sets of abilities.

A *task* is a particular specification of a problem, as represented in the *dataset*. There needs to be a test set, sometimes also training and validation sets.

A *metric* is way to summarize *model performance* over some set or sets of *tasks* as a single number or score.

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Preliminaries

- A *benchmark* is a combination of (i) a *dataset* or sets of *datasets*, and (ii) a *metric*, conceptualized as representing one or more specific *tasks* or sets of abilities. Benchmarks are adopted by a community of researchers as a shared framework for the comparison of *models* (Raji et al., 2021).

A *task* is a particular specification of a problem, as represented in the *dataset*. There needs to be a test set, sometimes also training and validation sets.

Models that obtain the most favorable scores on the *metrics* for a *benchmark* in terms of *performance* on the specified *task* is called State of the Art (SOTA).

A *metric* is way to summarize *model performance* over some set or sets of *tasks* as a single number or score.

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Demarcation

- Quick Test: are these evaluations "benchmarks" per Raji et al. (2021)?

Computer Science > Computation and Language

[Submitted on 22 Apr 2019 (v1), last revised 9 Sep 2019 (this version, v3)]

SocialIQA: Commonsense Reasoning about Social Interactions

Computer Science > Computation and Language

[Submitted on 4 Feb 2023 (v1), last revised 4 Nov 2024 (this version, v7)]

Evaluating Large Language Models in Theory of Mind Tasks

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 13 Feb 2025 (v1), last revised 6 Mar 2025 (this version, v2)]

ZeroBench: An Impossible Visual Benchmark for Contemporary Large Multimodal Models

Demarcation

- Sap and Rashkin et al. (2019) on social interactions:
  - **Motivation**: *"While humans trivially acquire and develop such social reasoning skills, this is still a challenge for machine learning models."*
  - **Task & Subject**: *"Social IQa aims to measure the social and emotional intelligence of computational models through multiple choice question answering."*
  - **Data**: *"Social IQa contains 37,588 multiple choice questions with three answer choices per question."*
  - **Metric**: *"Despite human performance of close to 90%, computational approaches based on large pretrained language models only achieve accuracies up to 65%."*

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, Yejin Choi. SocialIQA: Commonsense Reasoning about Social Interactions. EMNLP, 2019.

Demarcation

- Kosinski (2024) on Theory-of-Mind (ToM):
  - **Motivation**: *"We hypothesize that <u>ToM-like ability</u> does not have to be explicitly engineered into AI systems."*
  - **Task & Subject**: *"we administer two versions of the classic <u>false-belief task</u> widely used to test ToM in <u>humans</u> to several <u>language models</u>."*
  - **Data**: *"As GPT-3.5 may have encountered the original task in its training, hypothesis-blind research assistants (RAs) prepared 20 bespoke Unexpected Contents Task tasks."*
  - **Metric**: *"A task was considered solved correctly only if all 3 questions were answered correctly for both original and reversed task."*

NEURAL INFORMATION PROCESSING SYSTEMS

Michal Kosinski. Evaluating large language models in theory of mind tasks [Theory of Mind May Have Spontaneously Emerged in Large Language Models]. PNAS, 2024.

Demarcation

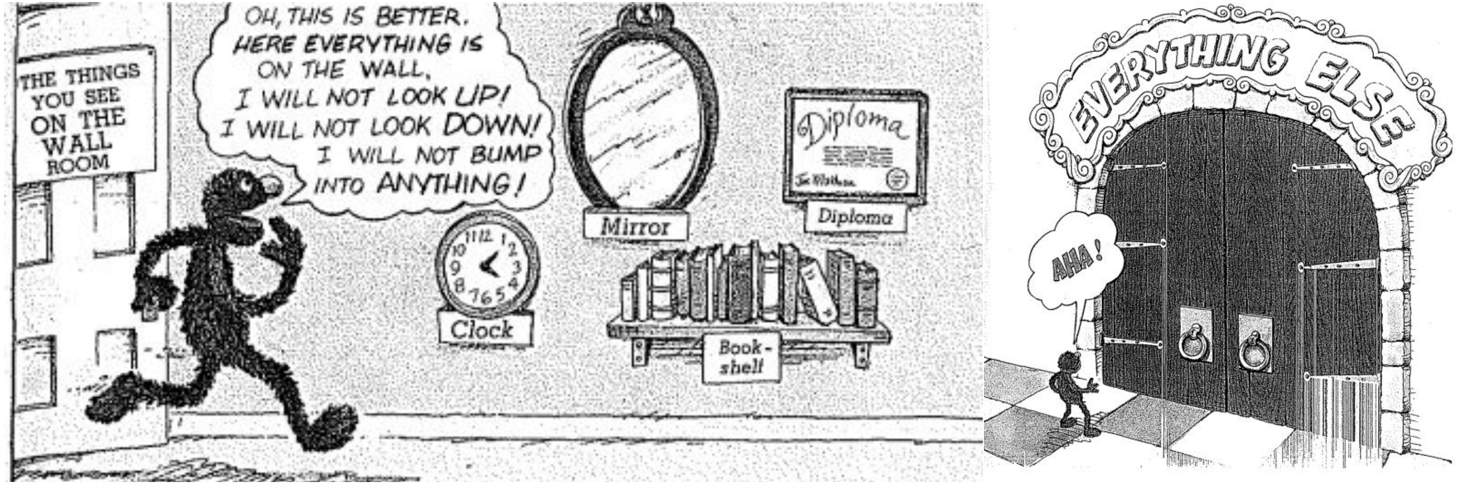
- Roberts et al. (2025), ZeroBench:
  - **Motivation**: *"…there is a pressing need for difficult benchmarks that <u>remain relevant for longer</u>"*
  - **Task & Subject**: *"…a lightweight <u>visual reasoning</u> benchmark that is <u>entirely impossible</u> for contemporary <u>frontier LMMs</u>."*
  - **Data**: *"Our benchmark consists of 100 manually curated questions and 334 less difficult subquestions."*
  - **Metric**: *"We use accuracy as our metric to evaluate the 100 ZeroBench main questions"*

NEURAL INFORMATION PROCESSING SYSTEMS

Jonathan Roberts et al. ZeroBench: An Impossible Visual Benchmark for Contemporary Large Multimodal Models. Preprint, 2025.

Broadening the Definition

- Raji et al. (2021): *"The imagined artifact of the 'general' benchmark does not actually exist."*
- But we are seeing more "general" benchmarks than domain-specific benchmarks nowadays.



Grover and the Everything in the Whole Wide World Museum

Inioluwa Deborah Raji, et al. AI and the Everything in the Whole Wide World Benchmark. NeurIPS Datasets and Benchmarks Track (Round 2), 2021.

Broadening the Definition

- AI benchmarking essentially depends on our philosophy of measurement.
- AI Benchmarking often drifts into operationalism (capability = score) without accepting its epistemic commitments.
  - I.e., "SOTA on SWE-Bench = Best SWE agent"
  - Benchmarks are fallible indicators of a richer construct.

Broadening the Definition

- These "general" AI benchmarks are proto-psychometrics: tests of latent abilities, but commonly with underdeveloped theory and validation.
  - Their job is **construct validation** (Bean, et al., 2025), not **definition**;
  - *"having measures that represent what matters to the phenomenon."*
  - We hope to argue: this task **T** is a useful probe of ((aspect **X** of) capability **C** of) a system **S**, under specific assumptions **A1**, **A2**, …

Andrew M. Bean, et al. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. NeurIPS, 2025.

Broadening the Definition

- What is a benchmark?
  - Defined by what we study (the system) and what we want (the goal).
- What is a **good** benchmark?
  - No single answer, laden with values and methodological assumptions.
- How should we use benchmarks?
  - Follow their intended scope and interpret scores within the benchmark's and your system's assumptions.

NEURAL INFORMATION PROCESSING SYSTEMS

The Target System

- A machine learning algorithm:
  - I.e., classical supervised learning regime;
  - E.g., a new linear transformer;
  - Dataset split: train / validation / test;
  - A benchmark is a dataset with **fixed split (train/val/test)** plus an evaluation metric, used to compare algorithms under controlled data.

The Target System

- (Edge case) continual or test-time adaptable algorithms:
  - E.g., MEMO, TENT, TTT-based algorithms with batch norm (BN);
  - Updates are computed from batches of test data;
  - Changing batch size → changes the gradient and BN statistics;
  - Changing batch order → changes the adaptation trajectory;
  - For continual or test-time adaptable algorithms, a meaningful benchmark specification should include, in addition to the dataset splits and metric, **the test-time batch size and the evaluation order**, since these affect the adaptation dynamics and thus performance.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. ICML, 2020.
Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, Trevor Darrell. Tent: Fully Test-time Adaptation by Entropy Minimization. ICLR, 2021.
Marvin Zhang, Sergey Levine, Chelsea Finn. MEMO: Test Time Robustness via Adaptation and Augmentation. NeurIPS, 2022.

The Target System

- A pretrained weight:
  - I.e., Self-supervised learning and pretraining regimes;
  - E.g., BERT-base, Qwen3-0.6B-Base, DINOv2, MAE-L;
  - Dataset split: dev / test;
  - A benchmark is a test set plus an evaluation metric, **optionally with a dev set for tuning**, used to compare pretrained weights under controlled data conditions.

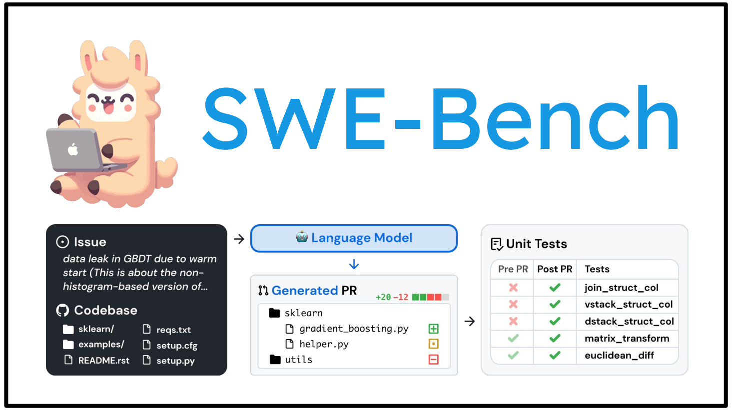
The Target System

- A digital (code/web/terminal/…) agent:
  - E.g., modern LLM agents
  - “I developed the SOTA agent”: A benchmark is a test set plus an evaluation metric, used to compare agents.
    - E.g., OpenHands + GPT-5 scores 71.8 on SWE-Bench (verified).
  - “I developed the SOTA LLM with agentic capability”: A benchmark is a test set, **a fixed agent workflow**, plus an evaluation metric, used to compare pretrained weights under same workflow and workload.
    - E.g., Claude 4.5 Opus scores 74.4 on SWE-Bench (Verified) with a minimal agent (aka, the Bash Only setting).

The Research Goal

- Task-driven.
- To understand our progress on a task or well-defined application.

“we annotate each lidar point from a keyframe in nuScenes with one of 32 possible semantic labels (i.e. lidar semantic segmentation).”



“evaluates LMs in a realistic software engineering setting… featuring [GitHub issues and pull requests] from 12 repositories”

Holger Caesar, et al. nuScenes: A Multimodal Dataset for Autonomous Driving. CVPR, 2020.
Carlos E. Jimenez, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024.

The Research Goal

- Capability-driven.
- To evaluate the level of competence of a cognitive capability.

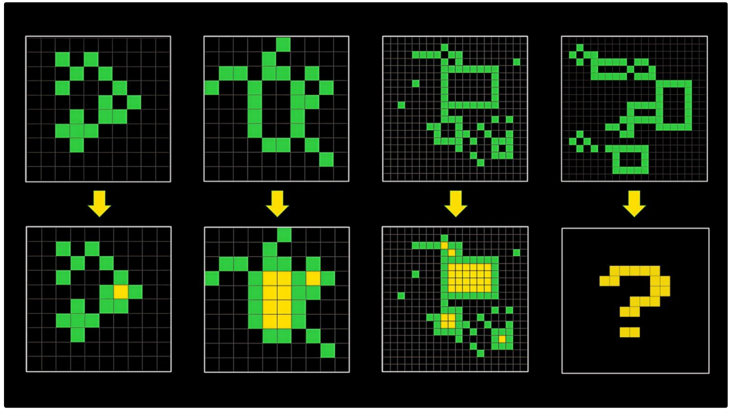
"…a collection of tools for evaluating the performance of models across a diverse set of existing NLU tasks…"

"We propose a novel multimodal video benchmark…to evaluate the perception and reasoning skills of pre-trained multimodal models"

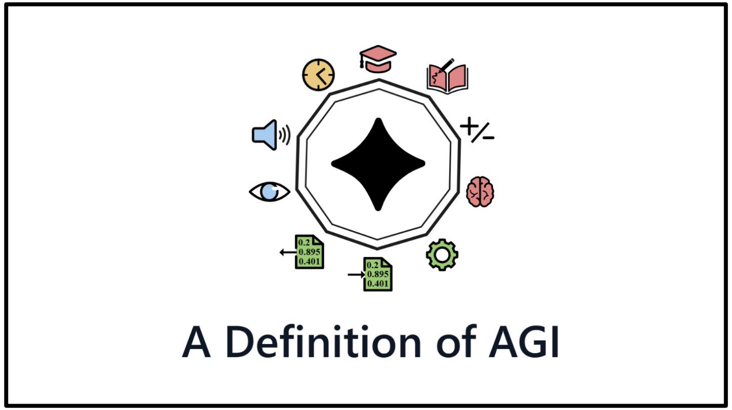
Alex Wang, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. BlackboxNLP, 2018.
Alex Wang, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. NeurIPS, 2019.
Viorica Patraucean, et al. Perception test: A diagnostic benchmark for multimodal video models. NeurIPS, 2023.

The Research Goal

- Model-driven.
- To measure the "intelligence" of a model.



"We argue that ARC can be used to measure a human-like form of general fluid intelligence…"



"The framework dissects general intelligence into ten core cognitive domains…"

François Chollet. On the Measure of Intelligence. Preprint, 2019.
Dan Hendrycks, et al. A Definition of AGI. Preprint, 2025.

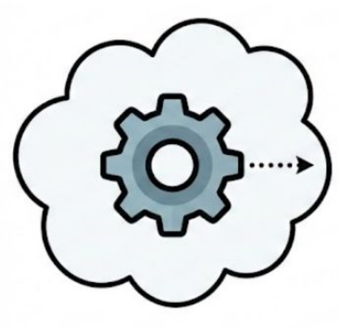
Defining "Quality" of Benchmarks

- What is a good benchmark?
- No single answer. Empirical results are laden with values and theoretical commitments (Boyd and Bogen, 2019).
- If one is designing a task- or capability-driven benchmark, a good benchmark is **a good data sampling function** over the **problem space.**

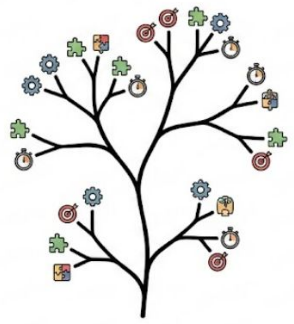
Nora Mills Boyd, and James Bogen. Theory and observation in science. Stanford Encyclopedia of Philosophy. 2009.

Defining "Quality" of Benchmarks

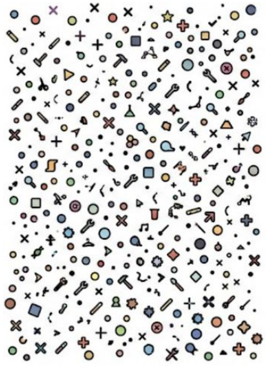
- When designing a new benchmark for a task, you would prefer …



A Proxy Subtask



A Taxonomy of Subtasks

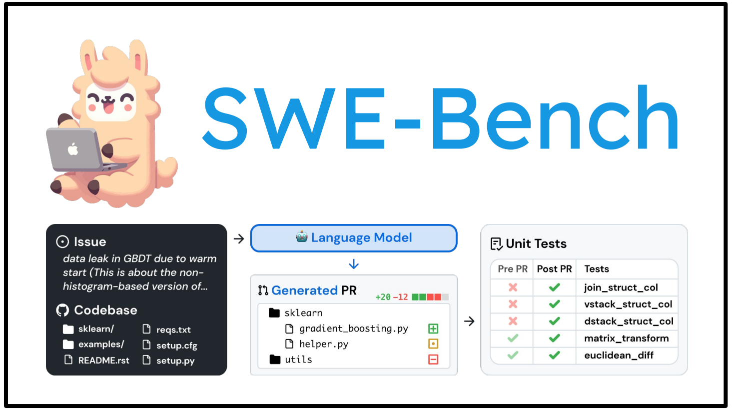


No Subtasks, Just Massive Testcases

Nora Mills Boyd, and James Bogen. Theory and observation in science. Stanford Encyclopedia of Philosophy. 2009.

Coverage in Benchmarks

- A good data sampling function = a representative proxy subtask.

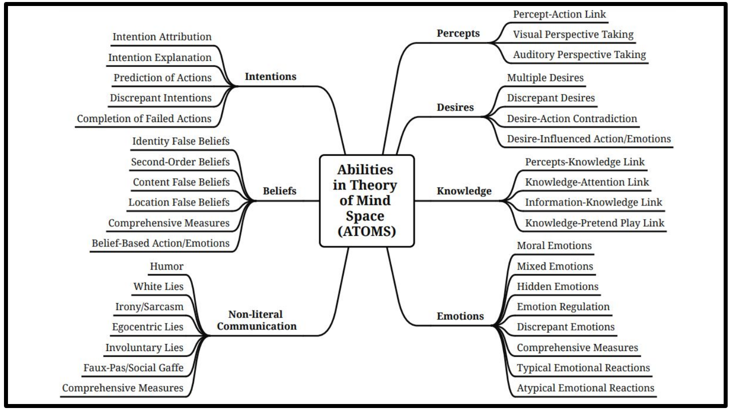


"However, existing coding benchmarks, such as HumanEval, mostly involve self-contained problems that can be solved in ***a few lines of code****. In the real world, software engineering is not as simple … we introduce SWE-bench, a benchmark that evaluates LMs in a* ***realistic software engineering setting****. models are tasked to* ***resolve issues*** *(typically a bug report or a feature request) submitted to popular GitHub repositories"*

Carlos E. Jimenez, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024.

Coverage in Benchmarks

- A good data sampling function = a systematic taxonomy.



"While the exact definition of ToM remains a central debate, the AI community can benefit from looking at what psychologists have viewed as an initial step. In this paper, we ***follow Beaudoin et al. (2020)'s taxonomy*** *of ToM sub-domains, i.e., the Abilities in Theory of Mind Space (ATOMS)."*

Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. EMNLP Findings, 2023.
Zhuang Chen, et al. ToMBench: Benchmarking Theory of Mind in Large Language Models. ACL, 2024.

Coverage vs Difficulty

- A good data sampling function = coverage, if you hope to estimate
  or difficulty, if you hope to differentiate

GLUE
SuperGLUE

*"GLUE is **a collection of nine language understanding tasks built on existing public datasets**, together with private test data, an evaluation server, a single-number target metric, and an accompanying expert constructed diagnostic set."*

*"SuperGLUE retains the two hardest tasks in GLUE. The remaining tasks were identified from those submitted to **an open call for task proposals and were selected based on difficulty** for current NLP approaches"*

Reality Checking on Coverage

- Reality check: how good are benchmarks as data sampling functions?
- Formulation.
  - Let **T** be the true task/domain distribution and **B** the benchmark's sampling distribution.
  - A benchmark is "good as a data sampling function" if **B** is (for the purposes we care about) indistinguishable from **T**.
  - That is, no reasonable test can look at the data alone and reliably tell whether an example came from the real domain or from the benchmark.

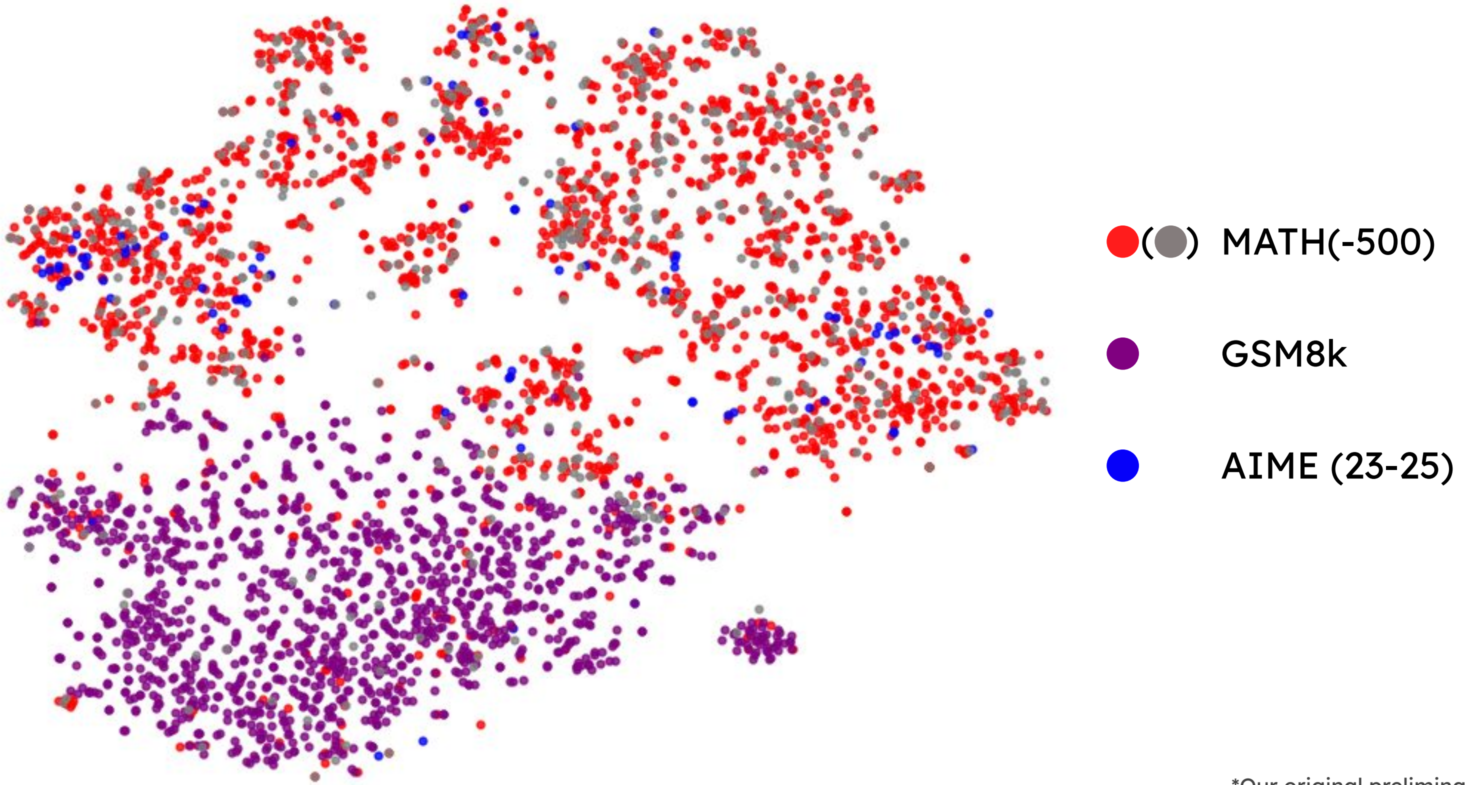
NEURAL INFORMATION PROCESSING SYSTEMS

Reality Checking on Coverage

- Reality check: how good are benchmarks as data sampling functions?
- Operationalized formulation.
  - A perfect benchmark is unattainable in both theory and practice.
  - We can only probe it indirectly via cross-benchmark checks and;
  - (i) Reproduce the data collection with the same described methodology or from real user interaction logs;
  - (ii) Compare to surrogate human data;
  - Use the OpenAI *text-embedding-3-small* model to encode the tasks into embeddings and compare benchmarks and human queries on similar tasks.

Reality Checking on Coverage

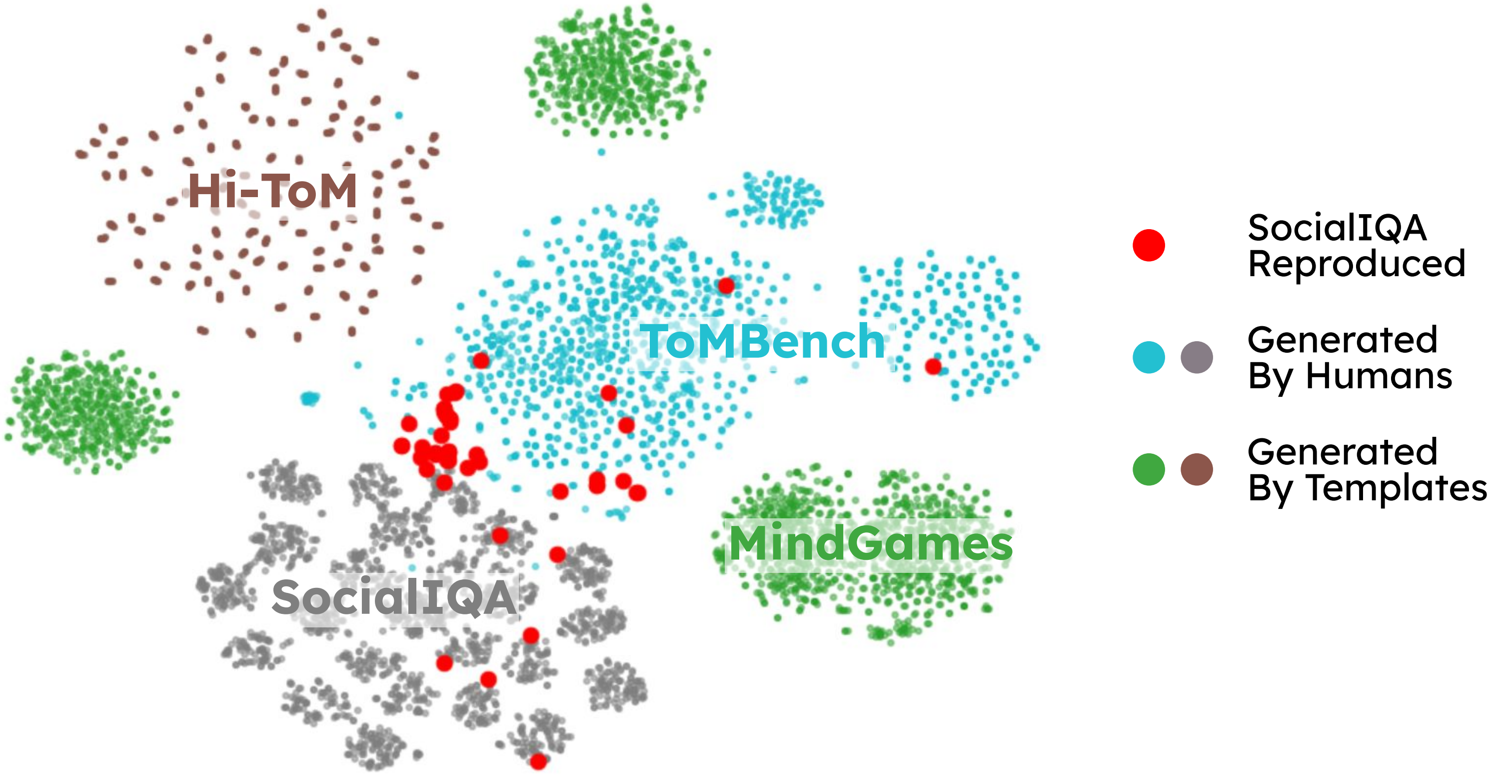
- Mathematics.



*Our original preliminary results.

Reality Checking on Coverage

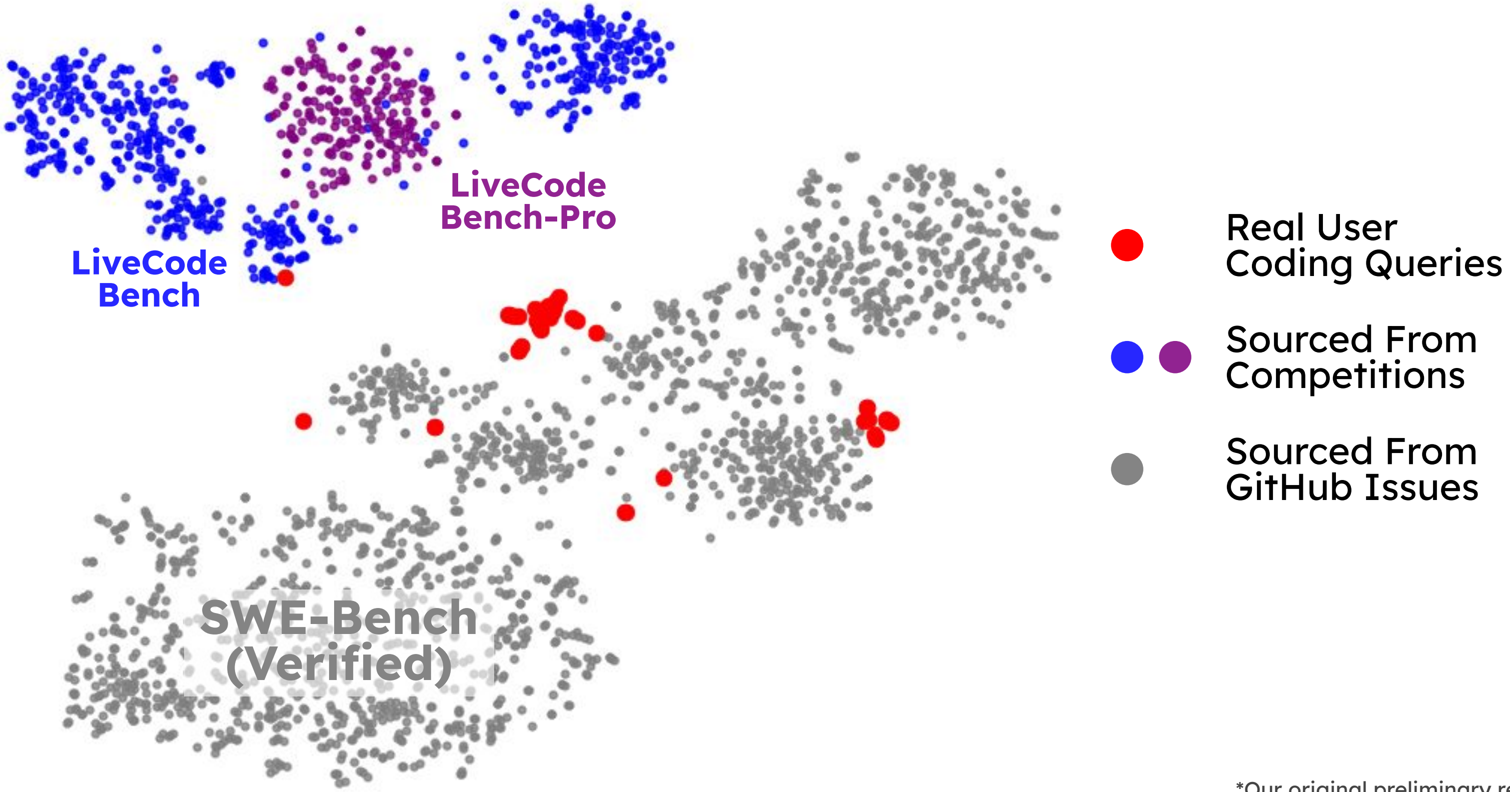
- Social Reasoning and Theory of Mind.



*Our original preliminary results.

Reality Checking on Coverage

- Coding and software engineering.



*Our original preliminary results.

NEURAL INFORMATION PROCESSING SYSTEMS

Reality Checking on Coverage

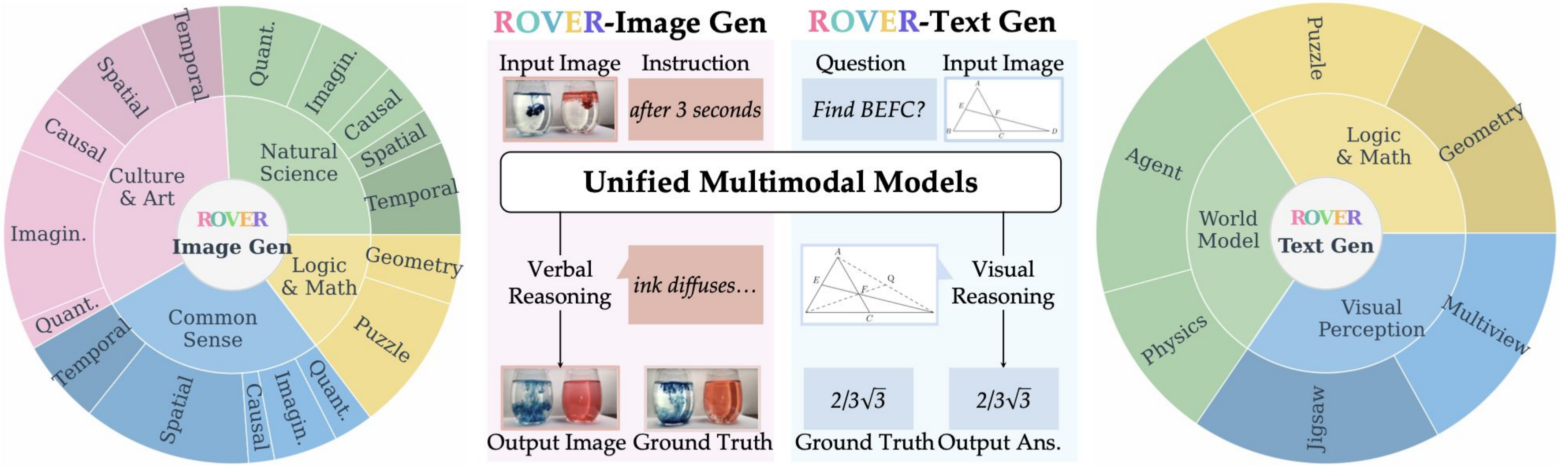
- Reality check: how good are benchmarks as data sampling functions?
- Summary.
  - Templated, proxy- and taxonomy-based datasets can easily fragment into many tiny clusters compared to human-generated ones.
  - The data collection pipelines described in papers can rarely be reproducible in practice.
  - Real user queries in human–LLM logs look very different from benchmark prompts.

Model-Driven Benchmarks

- Raji et al. (2021): benchmark should be task-driven.
- When we try to broaden to model-driven benchmarks, we are…
- Either …
  - Motivated by new capabilities demonstrated by new models;

Benchmarks Motivated by New Capabilities

- Example: omni-modal generation in unified multimodal models.



Yongyuan Liang, Wei Chow, et al. ROVER: Benchmarking Reciprocal Cross-Modal Reasoning for Omnimodal Generation. Preprint, 2025.
Zhiyuan Yan et al. Unified Multimodal Model as Auto-Encoder. Preprint, 2025.

Model-Driven Benchmarks

- Raji et al. (2021): benchmark should be task-driven.
- When we try to broaden to model-driven benchmarks, we are…
- Either …
  - Motivated by new capabilities demonstrated by new models;
- Or …
  - Hoping to challenge capable models with tasks that fail them;
  - Implicitly assuming generality;
  - Sometimes, claiming a measurement of intelligence, or colloquially using these benchmarks as proof that AGI.

Thought Experiment

- Martin’s magic gibberish word: ***flumborix*** (made up by ChatGPT);
- Benchmark question: *What is Martin’s magic gibberish word in his NeurIPS 2025 tutorial?*
- A model with a knowledge cutoff before Dec 1, 2025 must search the internet, find this slide, and extract this word.
- A model with a knowledge cutoff after Dec 1, 2025 can simply recall the word from its training dataset, which happens to include this slide.

Definition of "Intelligence"

- The 1921 Symposium on Intelligence and Its Measurement
  - Intelligence(capacity[knowledge]) = Library(shelf[books])
  - Henmo (1921): *"the capacity for knowledge and knowledge possessed."*
- Journal of AGI Special Issue "On Defining Artificial Intelligence"
  - Intelligence(agent[knowledge]) = Energy(motor[fuel]) *
  - Laird (2020): *"equate[s] with rationality, where an agent uses its available knowledge to select the best action(s) to achieve its goal(s) within an environment."*

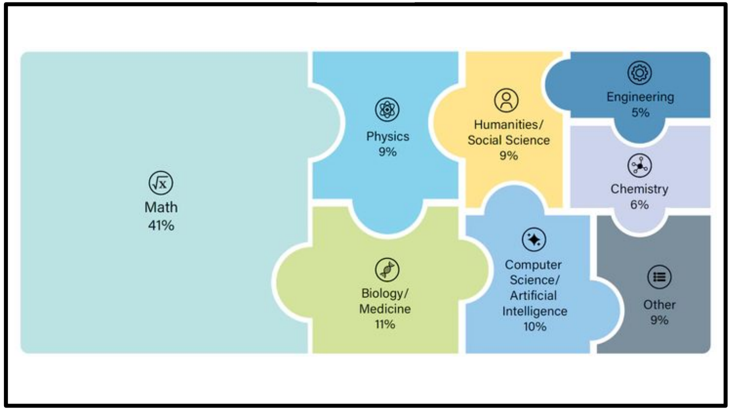
*This analogy is taken from John Laird's EECS 598 (Winter 2022) AGI with permission :)

Henmon, V.A.C. Intelligence and its measurement: A symposium VIII. Journal of Educational Psychology, 1921.
John Laird. Intelligence, knowledge & human-like intelligence. Journal of Artificial General Intelligence, 2020.

Fluid vs Crystallized

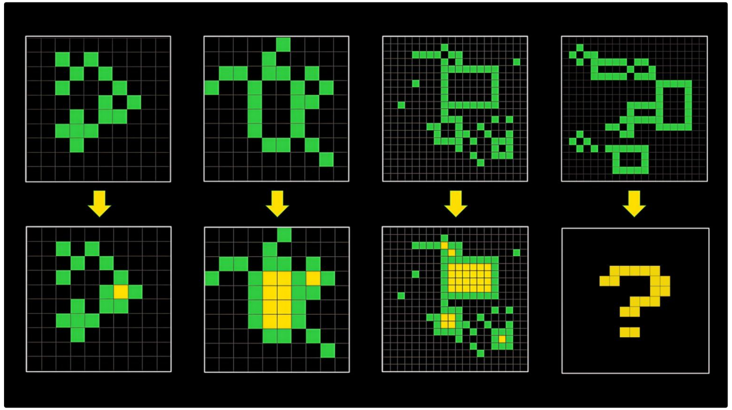
- Fluid intelligence is the ability to reason and solve new problems;
- Crystallized intelligence is the accumulation of knowledge and skills over a lifetime.

HLE



"a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind with broad subject coverage"

ARC-AGI



"We argue that ARC can be used to measure a human-like form of general fluid intelligence…"

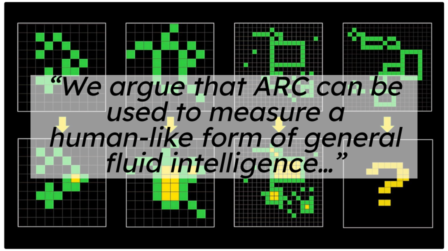
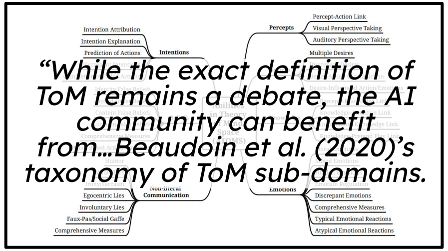
Scale AI. Humanity's Last Exam. Preprint, 2025.
François Chollet. On the Measure of Intelligence. Preprint, 2019.

Summary

- History
- **Definition**
  - Preliminary
  - Subject
  - Objective
- **Axiology**
  - Coverage
  - Fluidity
- Lifecycle
  - Design
  - Retirement
- Outcomes

←what is the primary goal of the evaluation?→

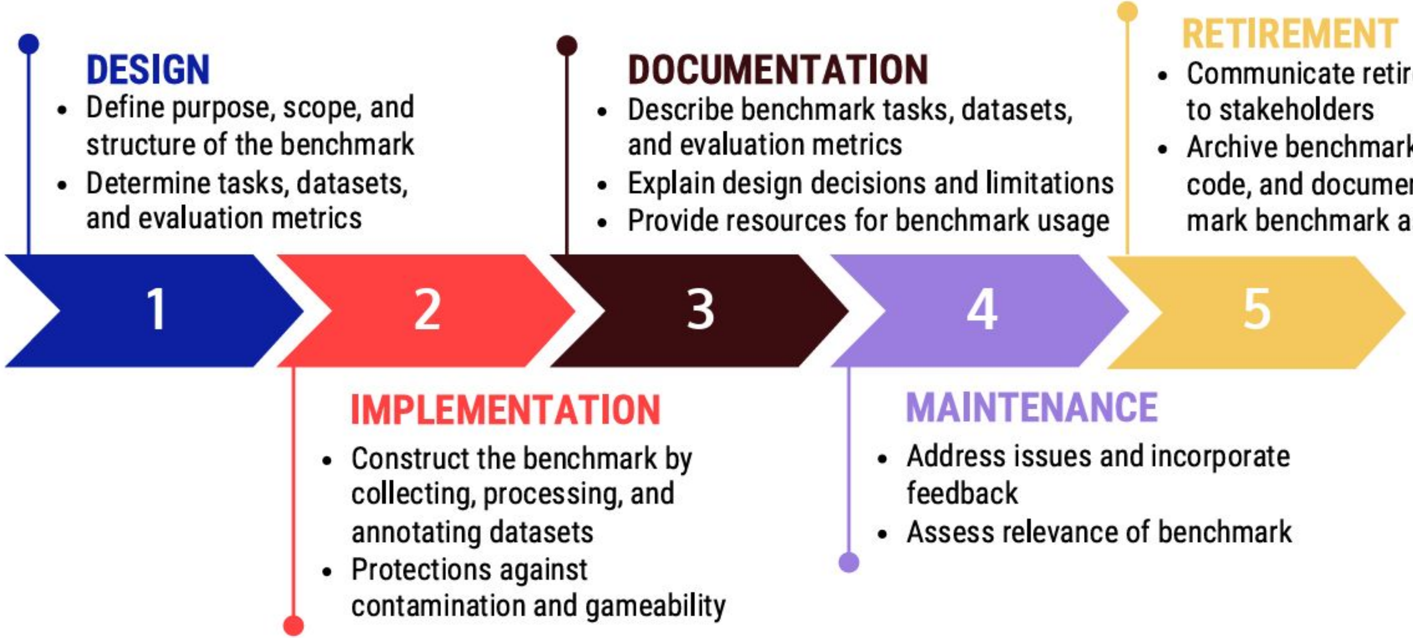
↑what is good data sampling?↓

	Progress of task	capability competence	model's intelligence
proxy is enough	*"evaluates LMs in a realistic software engineering setting… featuring [GitHub issues and pull requests] from 12 repositories"* SWE-Bench	*"We administer classic false-belief tasks, widely used to test ToM in humans, to several language models…"* **Evaluating Theory of Mind in Question Answering** Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, Thomas L. Griffiths **Theory of Mind May Have Spontaneously Emerged in Large Language Models** Michal Kosinski	 *"We argue that ARC can be used to measure a human-like form of general fluid intelligence…"*
taxonomy guided	 NUSCENES Lidarseg Challenge *"we publish the devkit, evaluation code, taxonomy, annotator instructions, and database schema for industry wide standardization."*	 *"While the exact definition of ToM remains a debate, the AI community can benefit from…Beaudoin et al. (2020)'s taxonomy of ToM sub-domains."*	*"The framework dissects general intelligence into ten core cognitive domains…"* **A Definition of AGI**
scale for coverage	*"a diverse set of challenging and realistic benchmark datasets to facilitate scalable, robust, and reproducible graph machine learning research."* OPEN GRAPH BENCHMARK	*"…a collection of tools for evaluating the performance of models across a diverse set of existing NLU tasks…"* GLUE SuperGLUE	*"It is impossible to enumerate the full set of tasks achievable by a sufficiently general intelligence. As such, an AGI benchmark should be a living benchmark"* Google DeepMind *originally published Nov. 2023; updated Jan. 2024* **Levels of AGI: Operationalizing Progress on the Path to AGI**

NEURAL INFORMATION PROCESSING SYSTEMS

Better Practice for Benchmark Creation

- The lifecycle of a benchmark is determined by both the benchmark creators and the community.

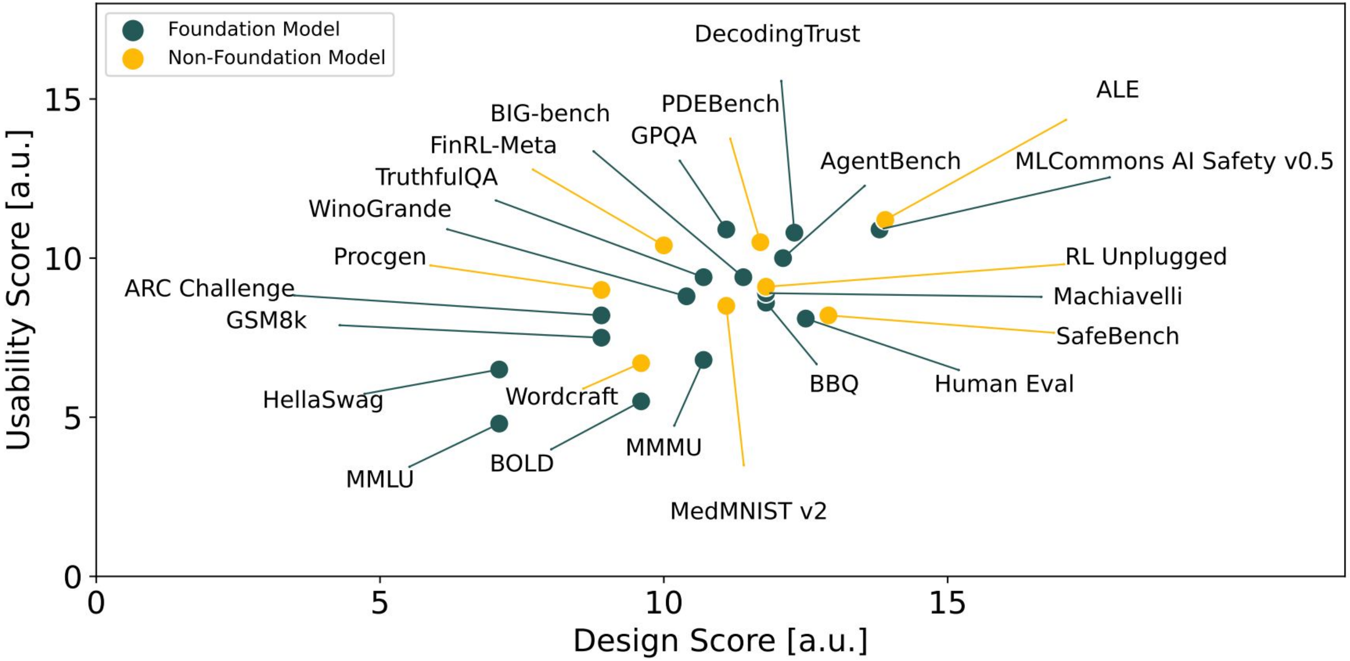


Anka Reuel, et al., BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. NeurIPS, 2024.
Jialun Cao, et al., How Should We Build A Benchmark? Revisiting 274 Code-Related Benchmarks For LLMs. Preprint, 2025.

History
Definition
Preliminary
Subject
Objective
Axiology
Coverage
Fluidity
Lifecycle
Design
Retirement
Outcomes

Better Practice for Benchmark Creation

- MLCommons AI Safety v0.5 (Vidgen et al., 2024) as an example.



Anka Reuel, et al., BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. NeurIPS, 2024.
Bertie Vidgen et al., Introducing v0.5 of the AI Safety Benchmark from MLCommons. Preprint, 2024.

NEURAL INFORMATION PROCESSING SYSTEMS

Benchmark Design

- **Conceptual grounding and domain relevance.**
  - *“We created the taxonomy through an iterative process over 10 weeks…reviewed 25+ existing taxonomies, 50+ AI safety evaluation datasets, 50+ research and policy papers, and 10+ community guidelines from industry Trust and Safety orgs”*
- **Positioning and intended use.**
  - *“The AI Safety Benchmark does not evaluate the safety of AI models ‘in general.’…Instead, the benchmark tests a specific AI system in a specific use case and for a specific set of personas.”*

Bertie Vidgen et al., Introducing v0.5 of the AI Safety Benchmark from MLCommons. Preprint, 2024.

Benchmark Design

- **Metric and scale design.**
  - Multi-stage grading: classifier labels responses → unsafe rates per test → 5-point risk grades → overall grade vs. a reference model.
  - Explicit mapping from unsafe % to risk levels ("Low", "Moderate-Low", etc.), defining practical floors and ceilings.
  - Humans to label a "human eval set" to validate, but no "human score on the benchmark" as a baseline system :(
- **Robustness and evaluation reliability.**
  - Construct 43,090 prompts by combining sentence fragments with 13 interaction types, explicitly stating that they use variation to provide *"holistic coverage of interaction types"* and to test robustness.

Bertie Vidgen et al., Introducing v0.5 of the AI Safety Benchmark from MLCommons. Preprint, 2024.

Benchmark Implementation

- **Accessible evaluation implementation.**
  - Released ModelBench as *"an openly available platform, and downloadable tool … to evaluate the safety of AI systems on the benchmark,"* with links to the GitHub repo; the test-spec schema for prompt generation and setup is also public.
- **Reproducible and statistically sound results.**
  - Provided annotator agreement (Cohen's kappa = 0.79) plus accuracy figures for the evaluator model
  - No CIs or significance tests :(

Benchmark Implementation

- **Benchmark integrity and data protection.**
  - Proposed keeping parts of the dataset hidden or delayed, introduced benchmark deprecation rules, and required publishers not to train directly on the benchmark and to retracted results if contamination is discovered.
- **Safety and release readiness.**
  - Added strong content warnings *"[t]his work involves viewing content that creates a risk of harm and you might find objectionable or offensive,"* provided wellbeing guidance for annotators, anonymize all tested models, and repeatedly stress that v0.5 is a preliminary proof-of-concept that *"should not be used to assess the safety of AI systems,"* with explicit release requirements for responsible use.

Benchmark Documentation

- **Clear software & usage docs.**
  - Open platform and downloadable tool on GitHub and provide a formal *"test specification"* schema;
  - Great inline code comments, and code documentation;
  - A "Trying It Out" for quick start of the code.
- **Transparent benchmark and evaluation design.**
  - The white paper itself is the design doc.
- **Thorough dataset and metadata documentation.**
  - Describe how 43,090 prompts are generated from sentence fragments and templates, gave detailed breakdown tables, and included a brief datasheet.

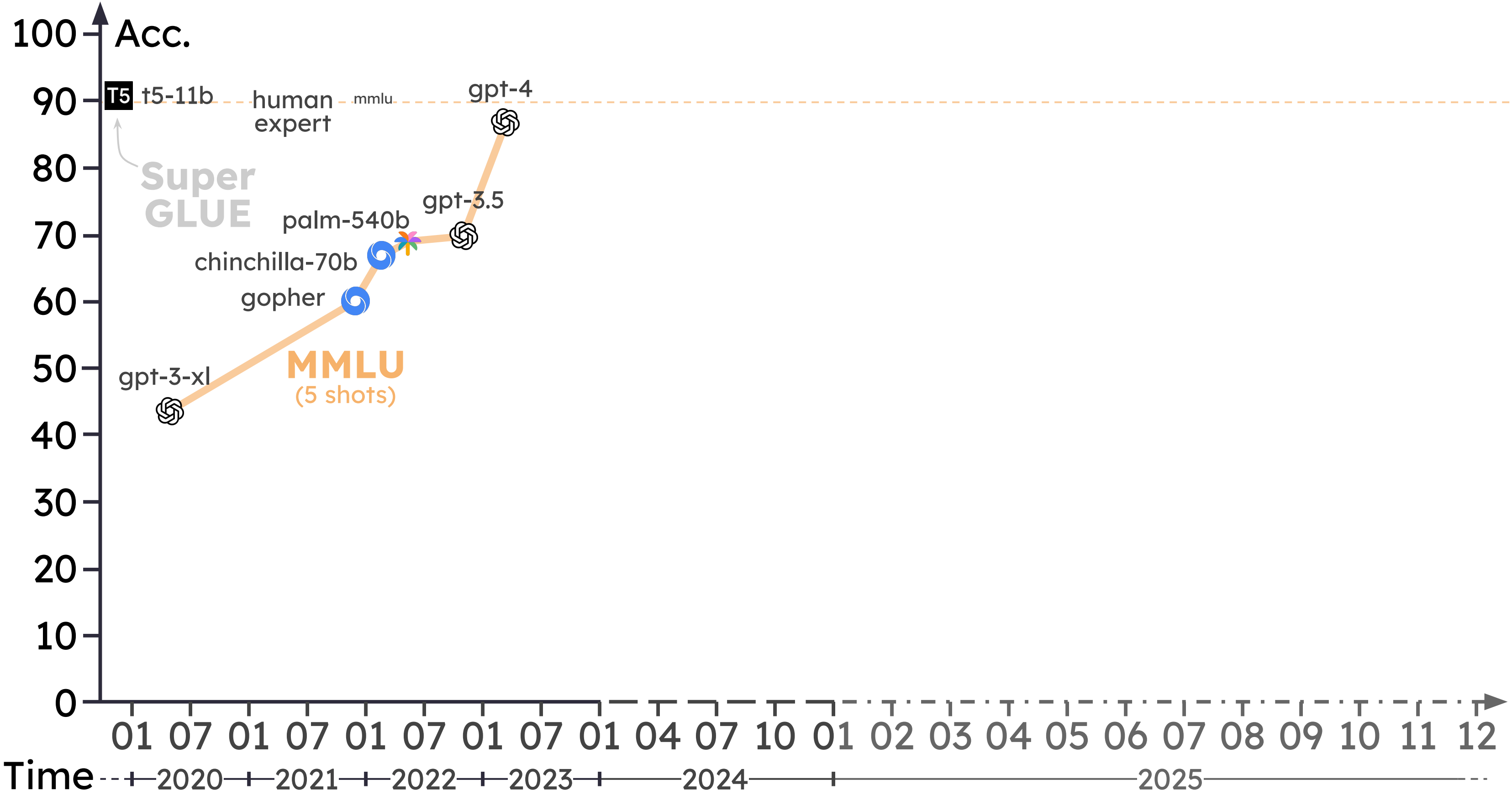
Benchmark Maintenance

- **Code usability was checked within the last year.**
  - The dataset will be *"regularly updated,"* supported by MLCommons and HELM.
- **A feedback channel for users is available.**
  - Explicitly invited community feedback, linked the AI Safety WG page, and stated that MLCommons can be contacted via the website; anyone can also join the WG
- **Point of contact.**
  - *"Reach out to Bertie if you have questions."*

Benchmark Retirement

NEURAL INFORMATION PROCESSING SYSTEMS

Progress$^{(?)}$ on Benchmark Leaderboards



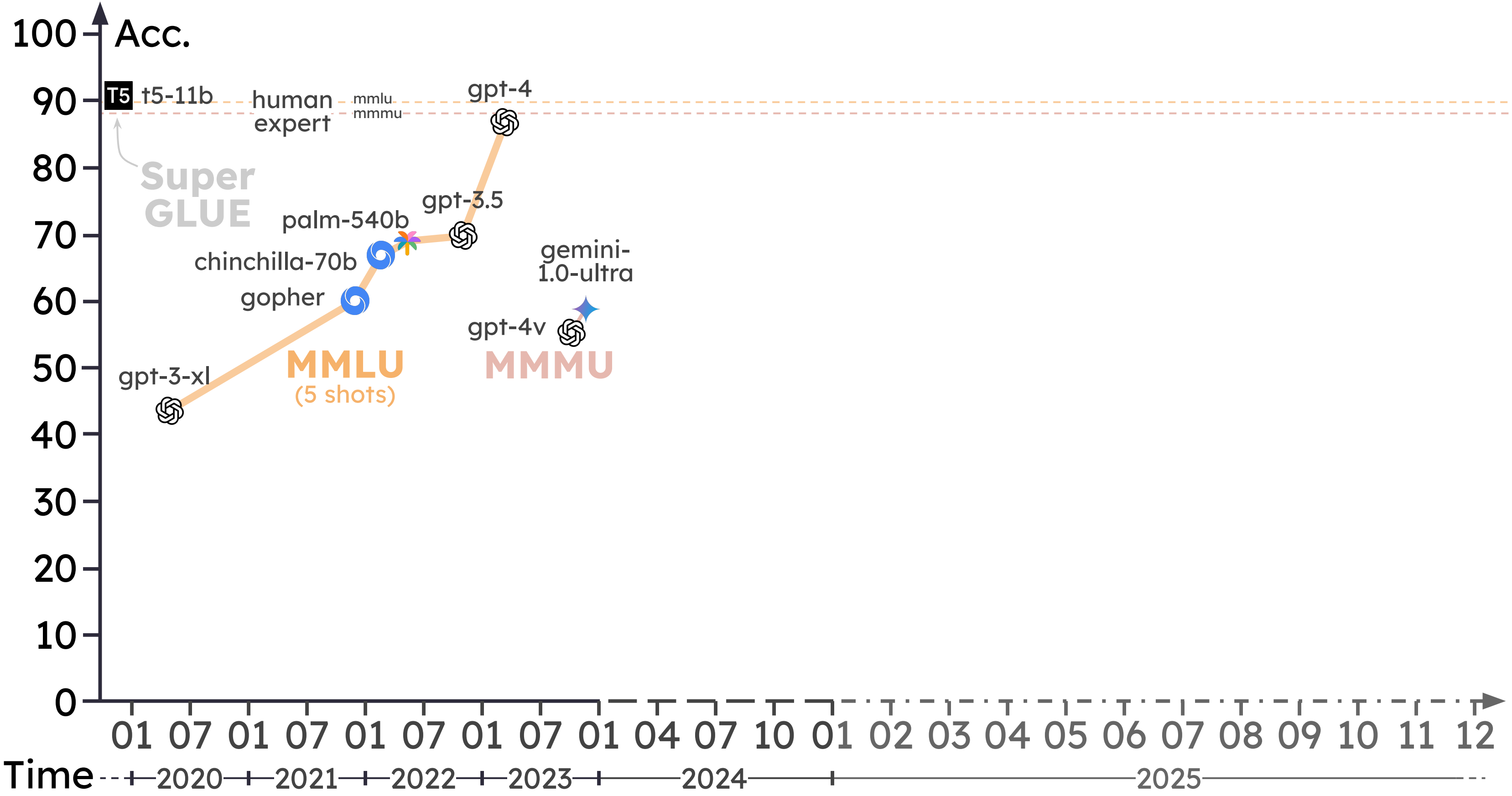
History
Definition
Preliminary
Subject
Objective
Axiology
Coverage
Fluidity
Lifecycle
Design
Retirement
Outcomes

NEURAL INFORMATION PROCESSING SYSTEMS

What's Measured? What's Missed? What's Next?

Progress(?) on Benchmark Leaderboards

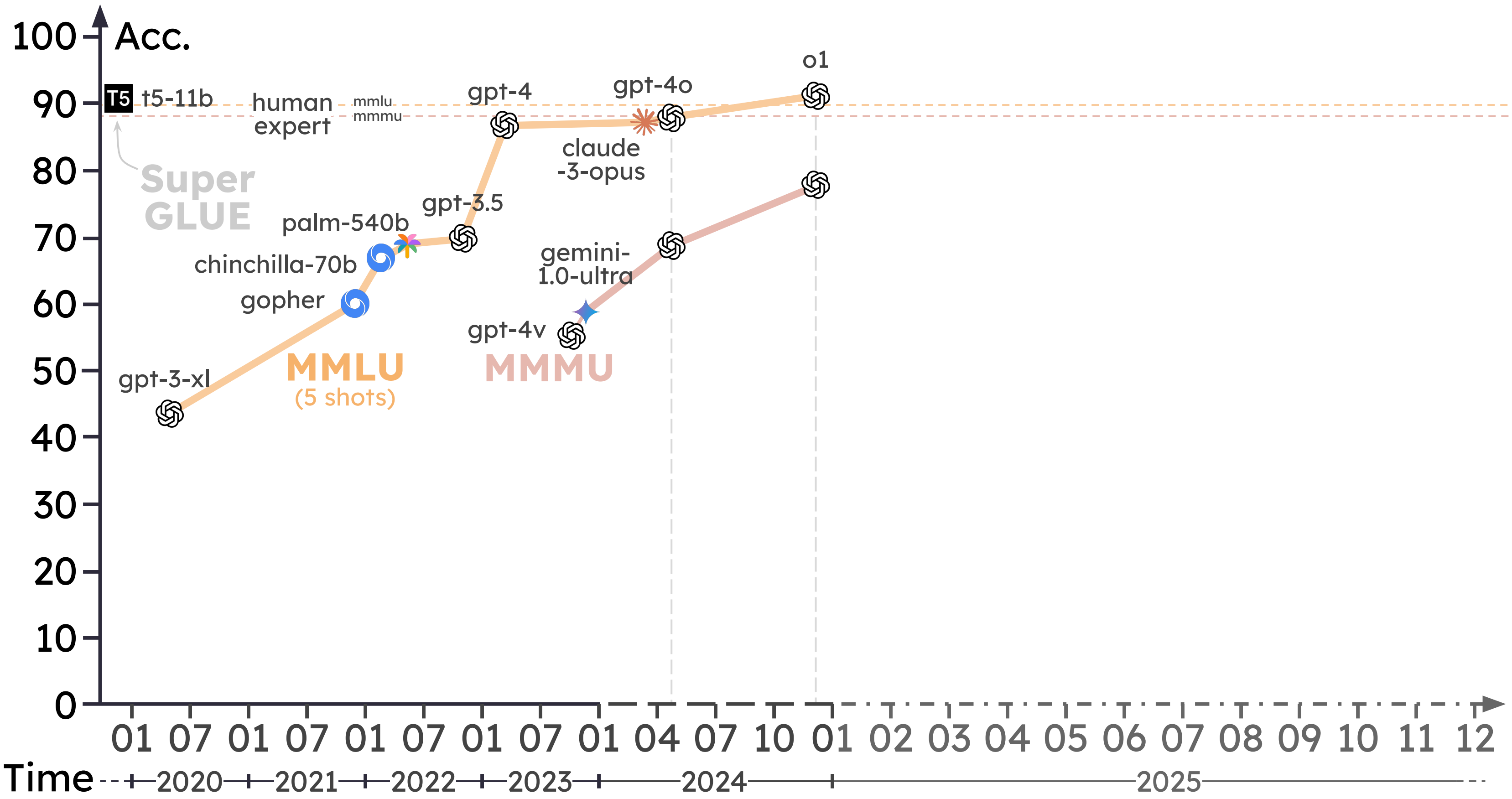
History
Definition
Preliminary
Subject
Objective
Axiology
Coverage
Fluidity
Lifecycle
Design
Retirement
Outcomes



NEURAL INFORMATION PROCESSING SYSTEMS

What's Measured? What's Missed? What's Next?

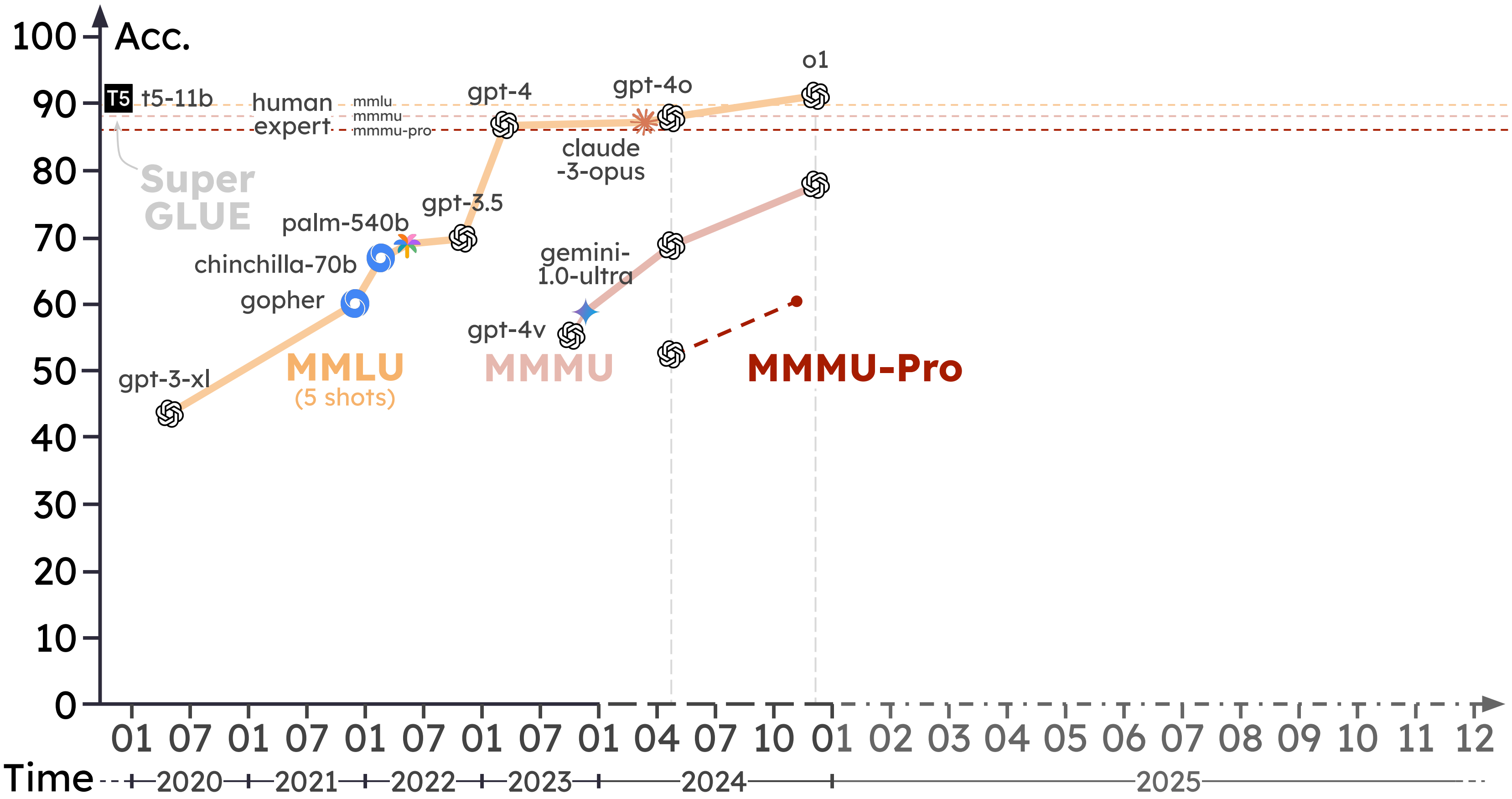
Progress$^{(?)}$ on Benchmark Leaderboards



NEURAL INFORMATION PROCESSING SYSTEMS

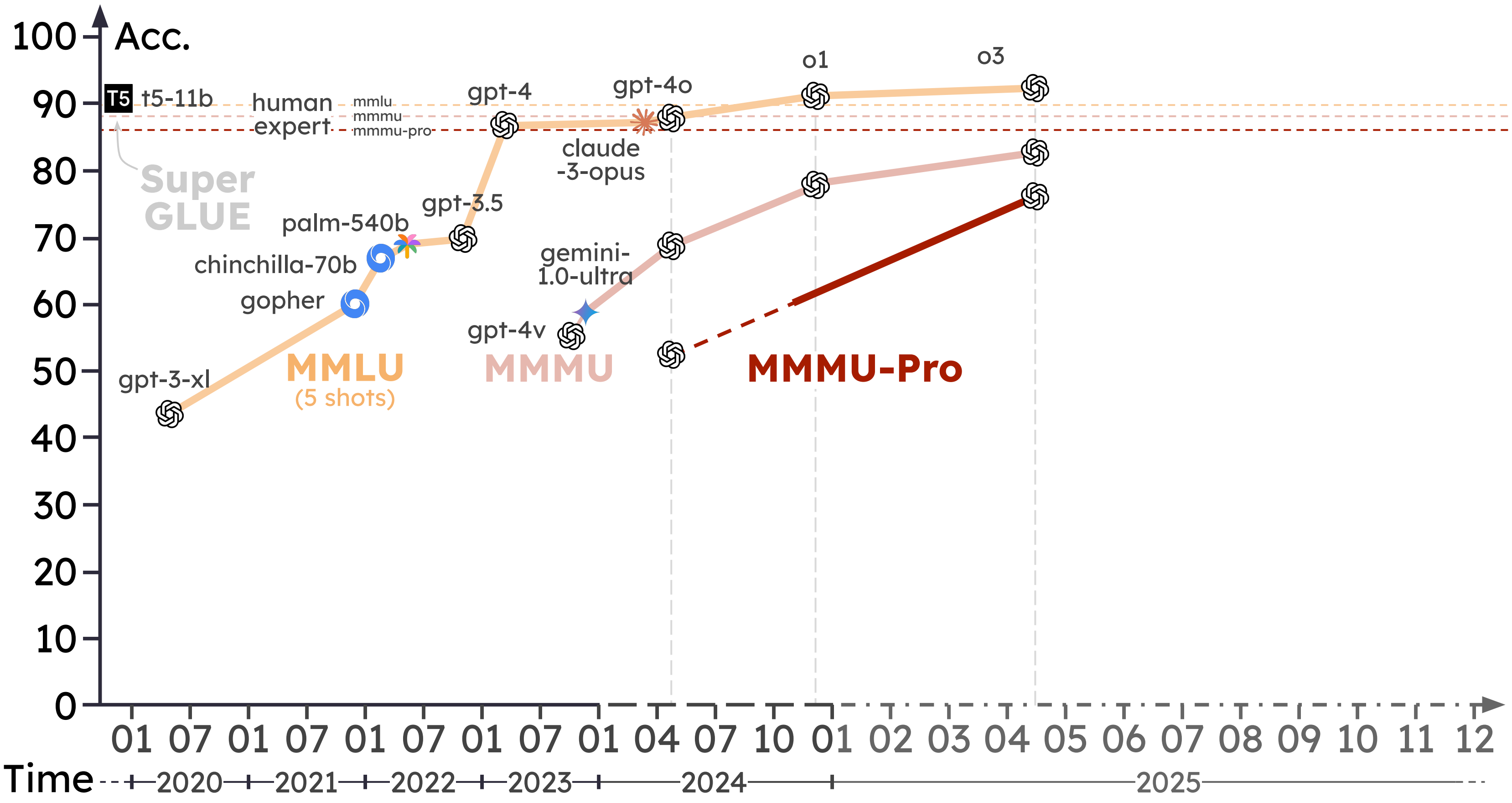
What's Measured? What's Missed? What's Next?

Progress(?) on Benchmark Leaderboards



NEURAL INFORMATION PROCESSING SYSTEMS

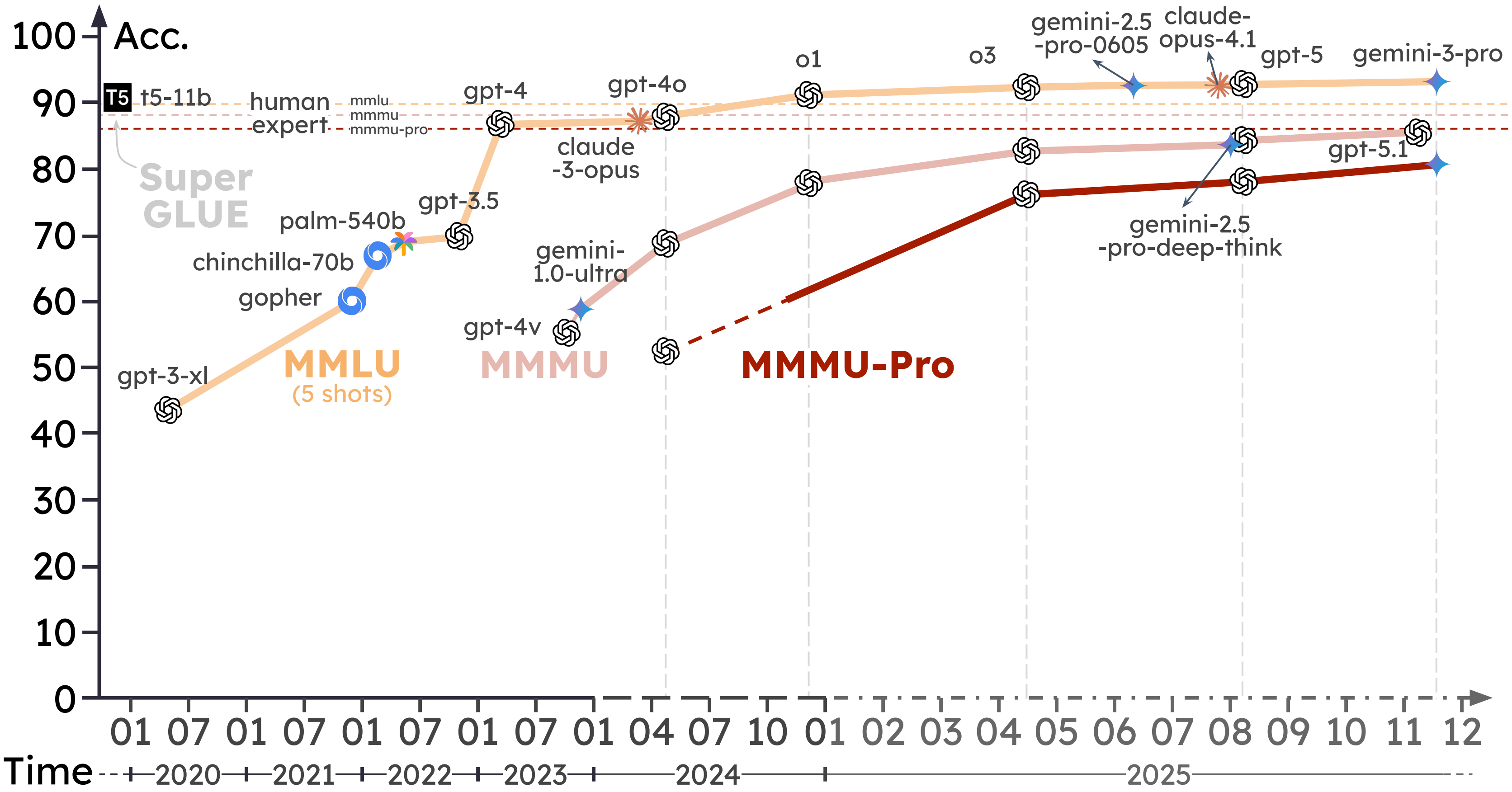
Progress(?) on Benchmark Leaderboards



NEURAL INFORMATION PROCESSING SYSTEMS

Progress^(?) on Benchmark Leaderboards

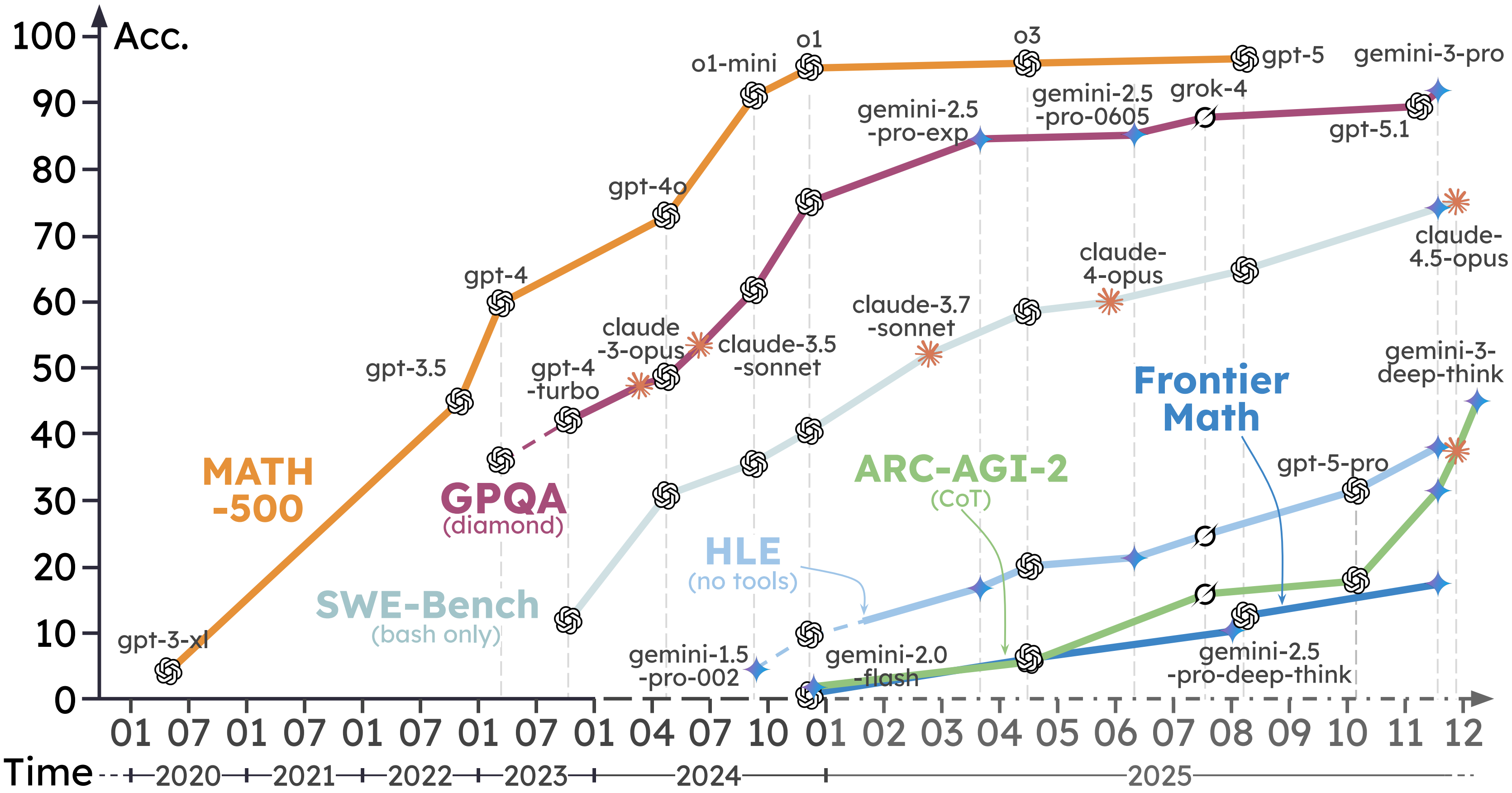
History
Definition
Preliminary
Subject
Objective
Axiology
Coverage
Fluidity
Lifecycle
Design
Retirement
Outcomes



NEURAL INFORMATION PROCESSING SYSTEMS

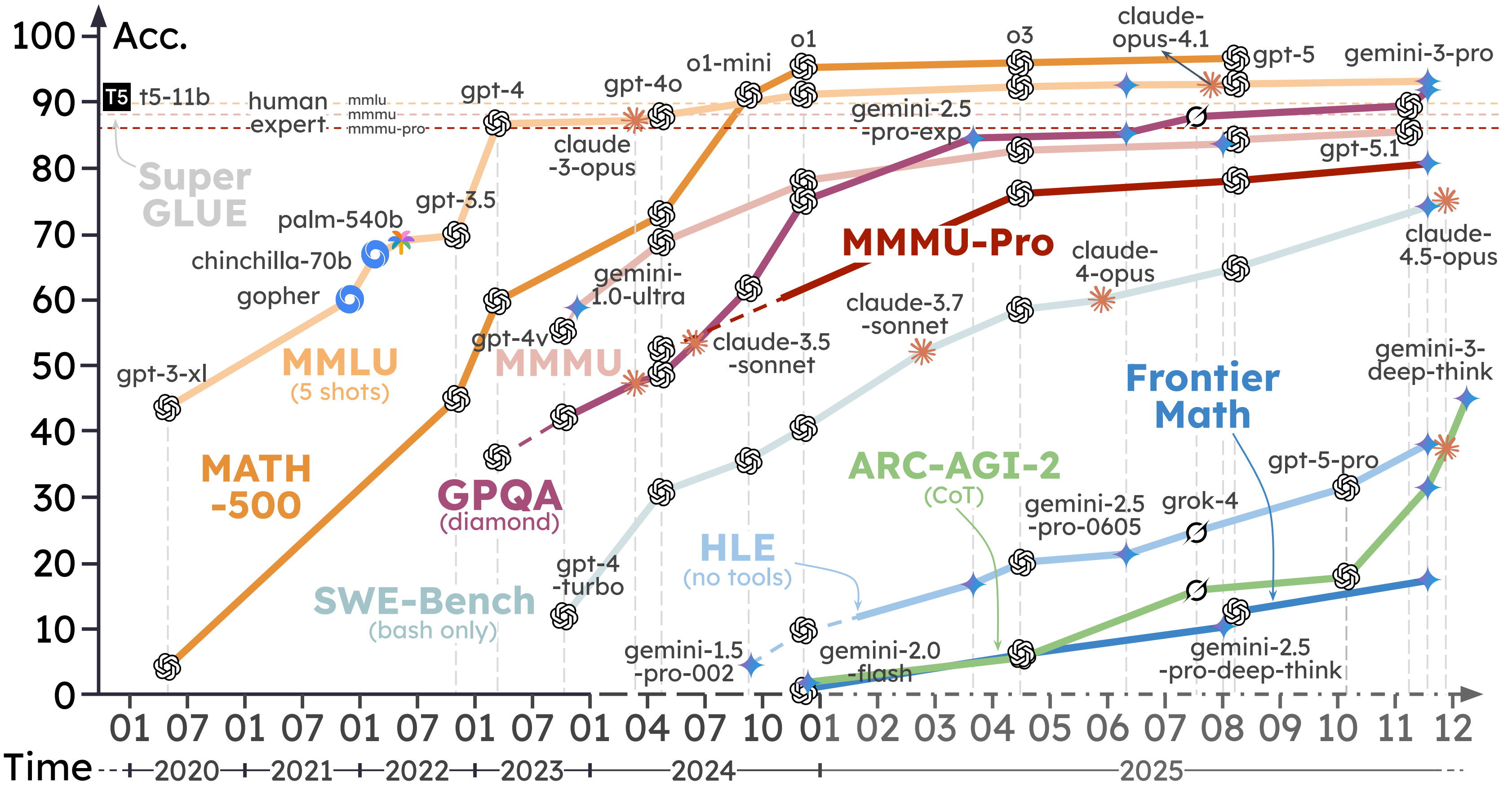
What's Measured? What's Missed? What's Next?

Progress(?) on Benchmark Leaderboards



NEURAL INFORMATION PROCESSING SYSTEMS

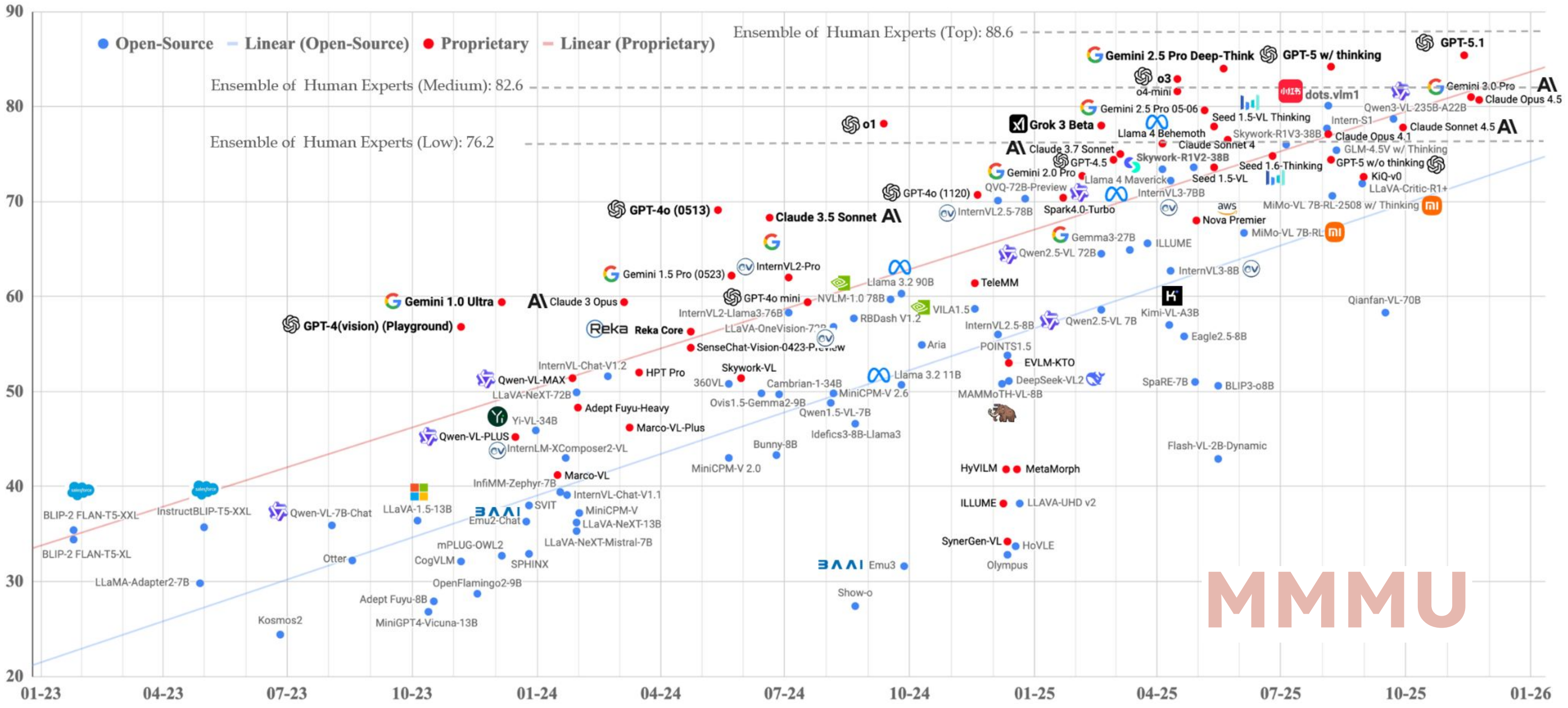
Progress$^{(?)}$ on Benchmark Leaderboards



NEURAL INFORMATION PROCESSING SYSTEMS

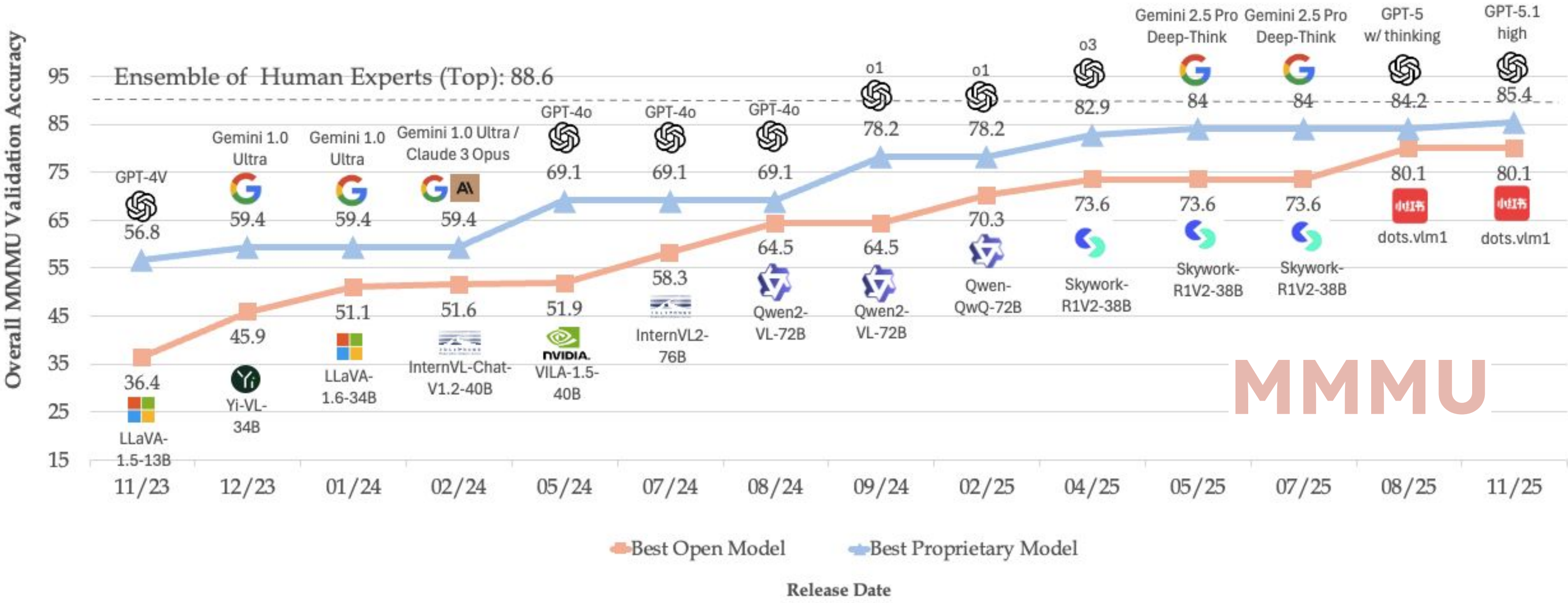
Benchmark Retirement

- Benchmarks that have retired from frontier lab competitions are still relevant to open-source communities.



Benchmark Retirement

- Benchmarks that have retired from frontier lab competitions are still relevant to open-source communities.



NEURAL INFORMATION PROCESSING SYSTEMS

Interpreting Benchmark Outcomes

- Is this performance improvement genuine?
  - Case: classification.
  - *"Through extensive experiments, we show that our method B outperform previous SOTA by a large margin."*

Model	**Accuracy (↑)**
A (Prev SOTA)	0.799
B (Ours)	**0.862**

NEURAL INFORMATION PROCESSING SYSTEMS

Interpreting Benchmark Outcomes

- Is this performance improvement genuine?
  - Case: classification.
  - *"Through extensive experiments, we show that our method B outperform previous SOTA by a large margin."*
  - …Probably not.

Model	Accuracy (↑)	Std. Dev.
A (Prev SOTA)	0.799	± 0.233
B (Ours)	**0.862**	± 0.666

Interpreting Benchmark Outcomes

- Is this performance improvement genuine?
  - Record per-seed per-example correctness;
  - Per-example: McNemar / sign test on disagreements:
    - McNemar test (per-example correctness): $p < 1e{-8}$ (***)
    - 95% CI for Δ (accuracy, across seeds): [0.048, 0.078]
  - Across seeds: paired t-test or Wilcoxon test:
    - Paired t-test across 5 seeds: mean $p = 5e{-4}$ (***)

Model	Accuracy (↑)	Std. Dev.
A (Prev SOTA)	0.799	± 0.004
B (Ours)	**0.862 (***)**	± 0.012

Interpreting Benchmark Outcomes

- Is this performance improvement genuine?
  - Case: novel view synthesis (NVS).
  - *"Through extensive experiments, we show that our method B outperform previous SOTA by a large margin."*

Model
A (Prev SOTA)
B (Ours)

PSNR (↑)
23.02
24.10

Interpreting Benchmark Outcomes

- Is this performance improvement genuine?
  - Case: novel view synthesis (NVS).
  - *“Through extensive experiments, we show that our method B outperform previous SOTA by a large margin.”*
  - …Probably not.

Model	**Resolution**	**PSNR (↑)**
A (Prev SOTA)	256 × 256	23.02
B (Ours)	512 × 512	**24.10**

Interpreting Benchmark Outcomes

- Is this performance improvement genuine?
  - –log(average error) ≤ average of [–log(error per region)]
  - Intuition: Higher resolution → more pixels → PSNR goes up when the underlying reconstruction quality has not meaningfully changed.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$$

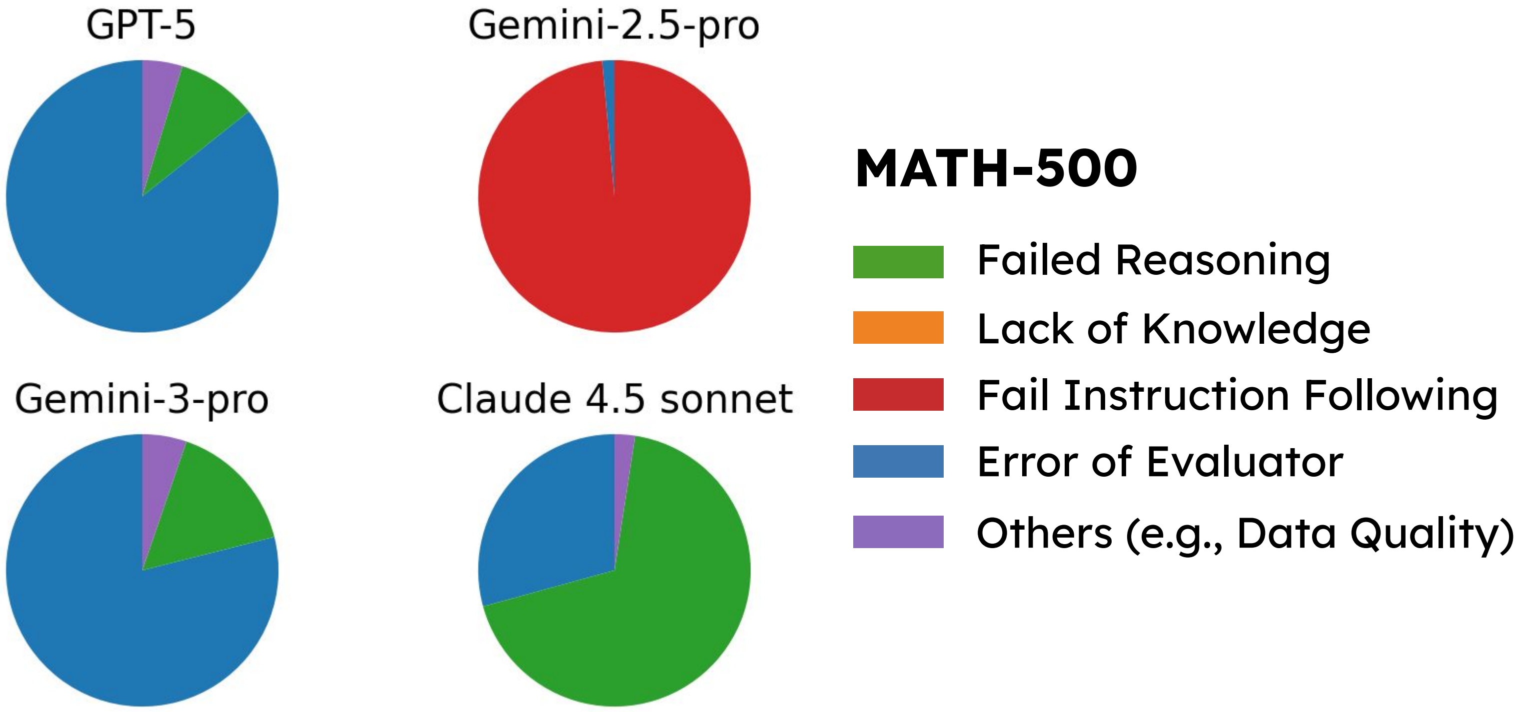
MAX: the maximum possible pixel value of the image given its encoding

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} e_i^2$$

NEURAL INFORMATION PROCESSING SYSTEMS

Reality Check on Benchmark Outcomes

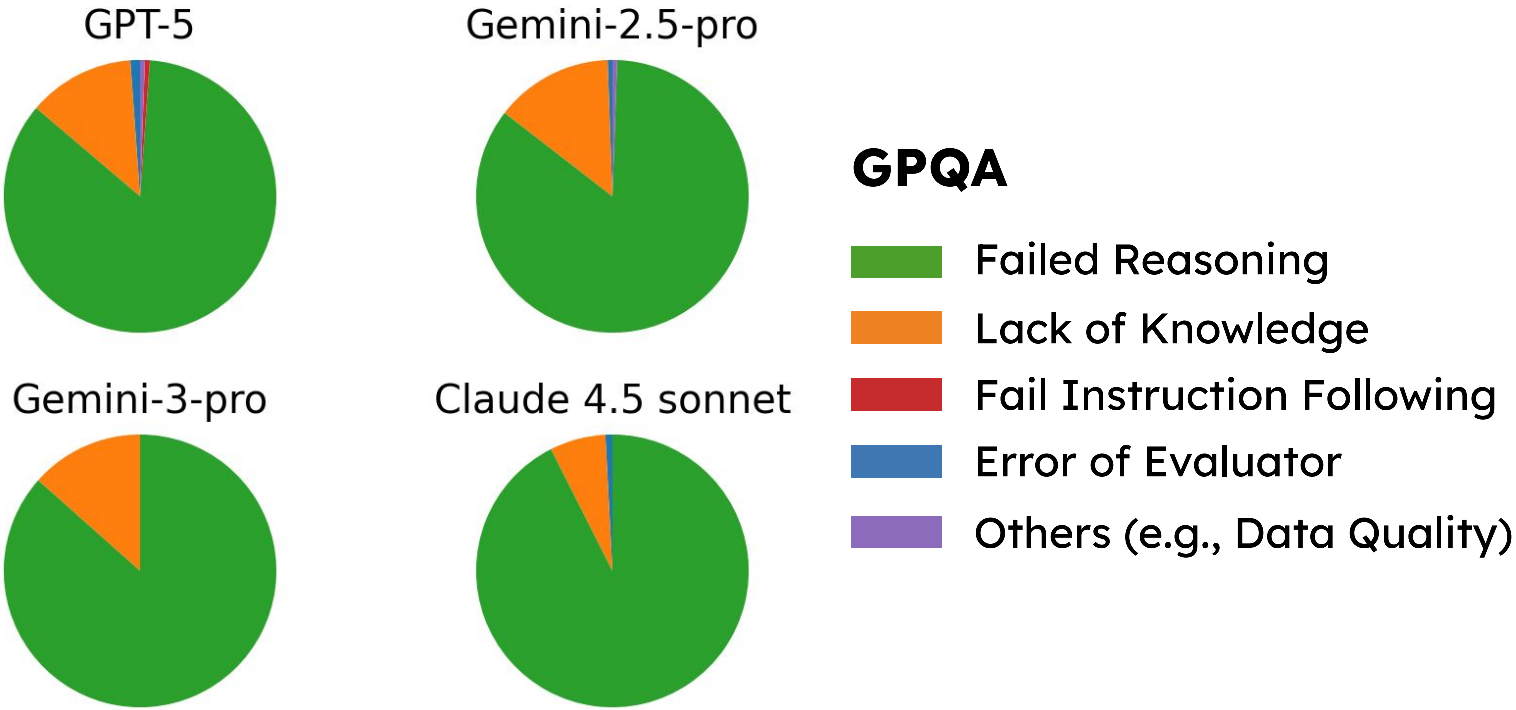
- What happened in the unsolved portion?



*Our original preliminary results.

Reality Check on Benchmark Outcomes

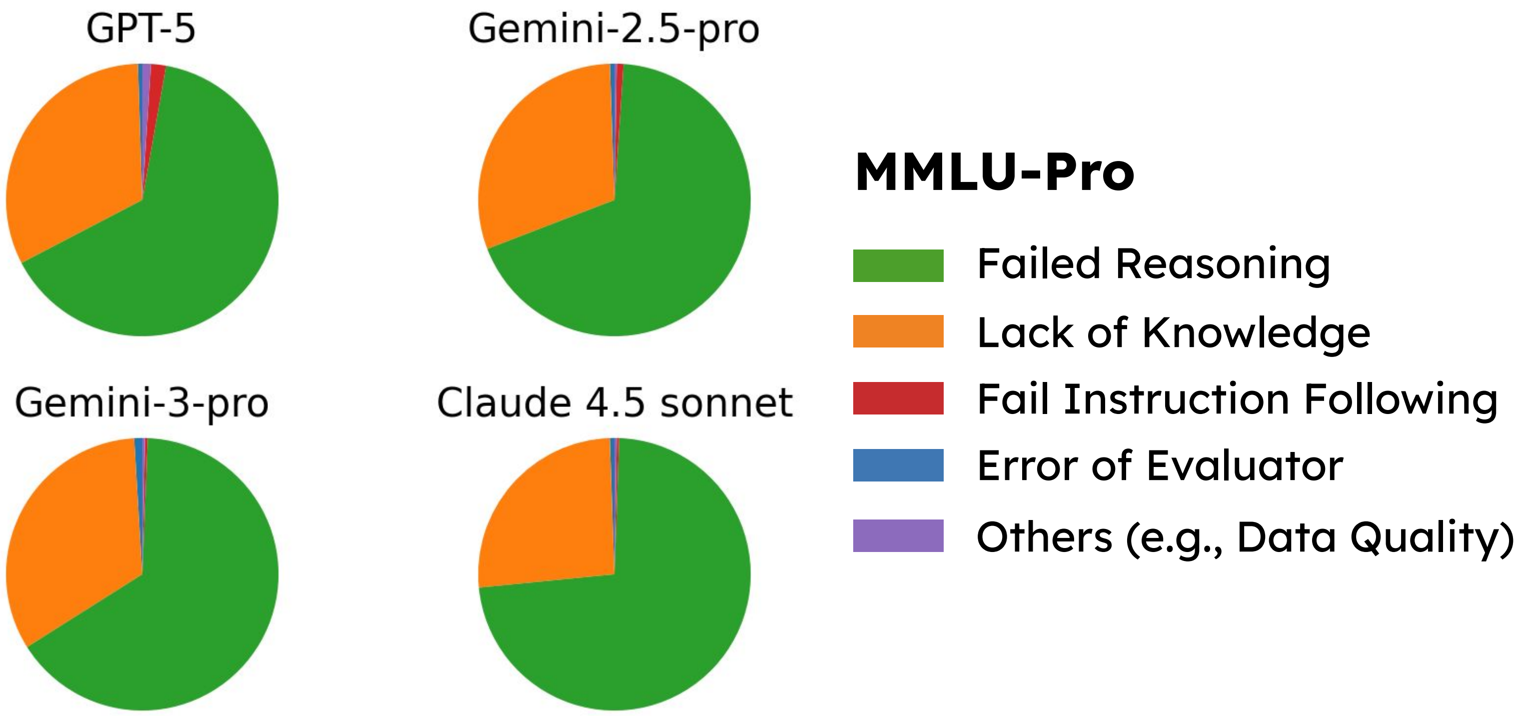
- What happened in the unsolved portion?



*Our original preliminary results.

Reality Check on Benchmark Outcomes

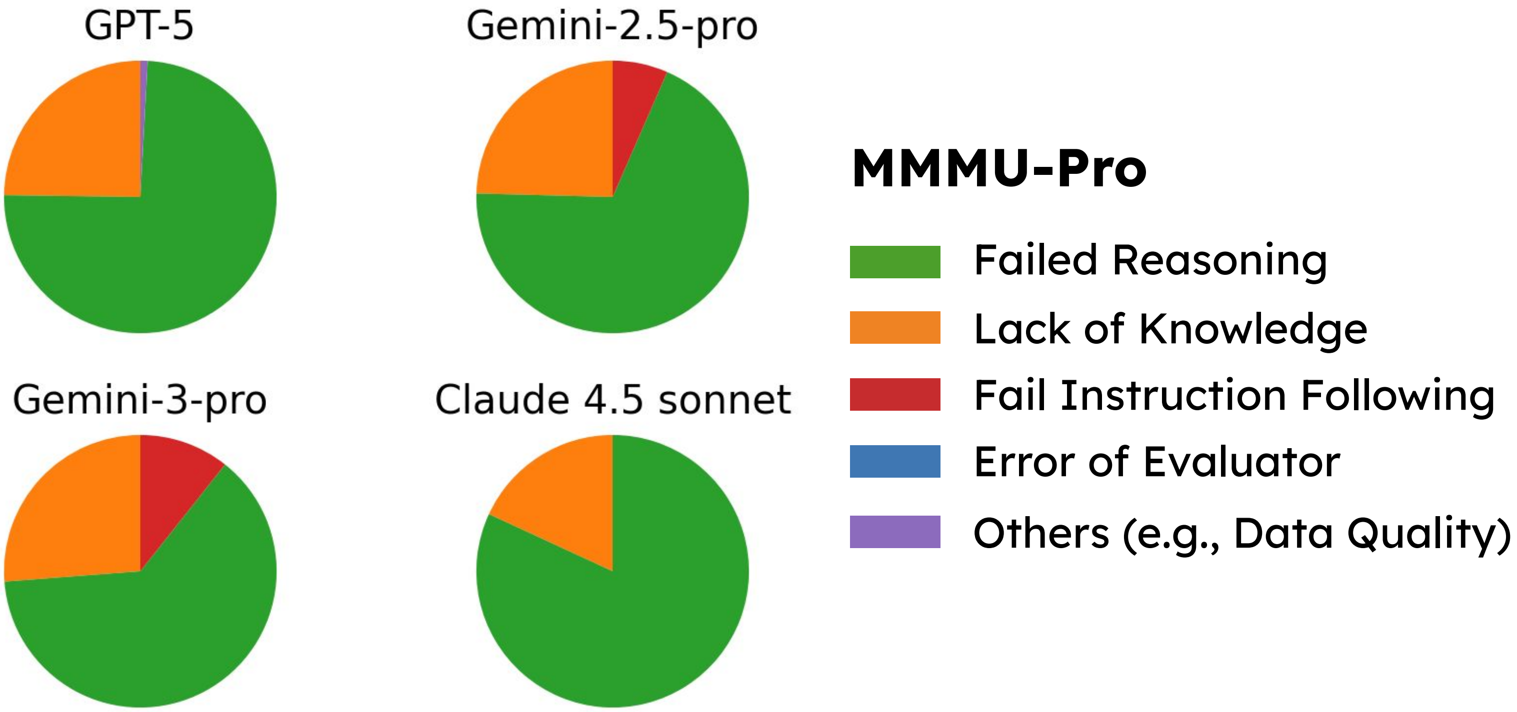
- What happened in the unsolved portion?



*Our original preliminary results.

Reality Check on Benchmark Outcomes

- What happened in the unsolved portion?



*Our original preliminary results.

NEURAL INFORMATION PROCESSING SYSTEMS

What's Missed?

Michael Saxon
University of Washington

https://benchmarking.science/slides.pdf

Agenda

- What’s Measured? (1:30PM - 2:10PM)
- What’s Missed? (2:10PM - 2:40PM)
  - Practical issues: data, integrity, measurement problems
  - Deeper issues: Systemic and epistemic problems
- What’s Next? (2:40PM - 3:15PM)
- Panel Discussion (3:20PM-4:00PM)

NEURAL INFORMATION PROCESSING SYSTEMS

What's missed?

- There are many criticisms of modern benchmarking practices
- **They are easy to miss when designing a new one**

- Here we borrow from the following works at a high level:

- *Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation*, Eriksson et al. (2025)

- *Benchmarks as Microscopes: Toward a Science of Model Metrology*, Saxon et al. (2024)

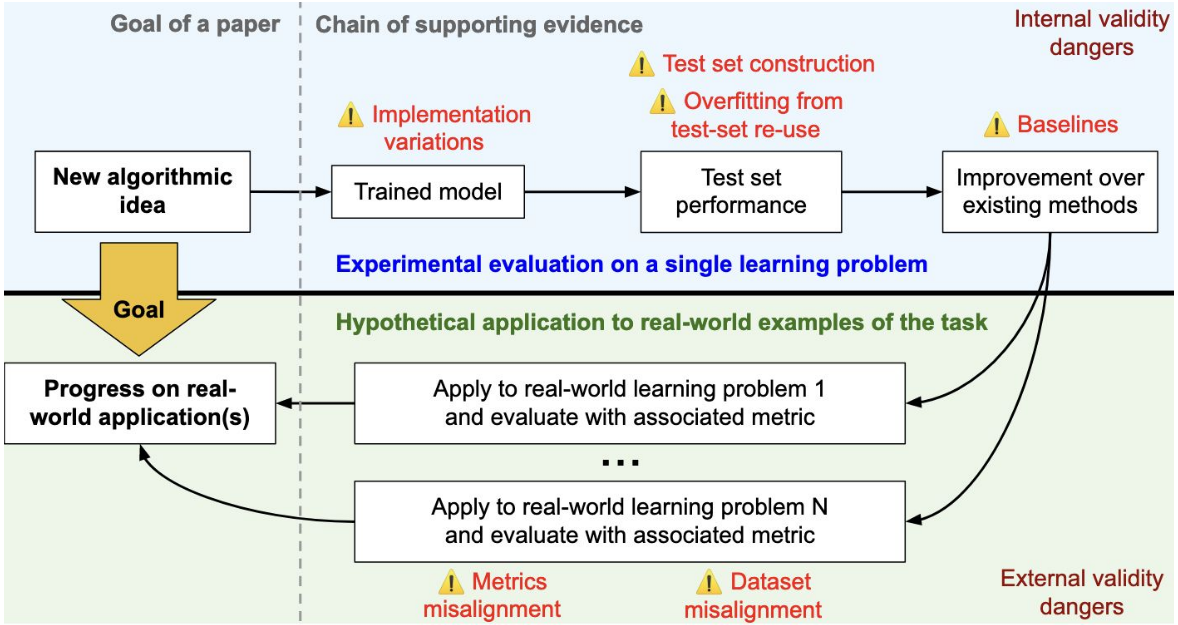
- *AI and the Everything in the Whole Wide World Benchmark*, Raji et al. (2022)

What's missed?

- Many fundamental challenges and critiques of benchmarking have been levied, but are often **missed** by practitioners. Including:
- **Data issues**
- **Measurement issues**
- **Systemic issues**
- **Epistemic issues**
- We will discuss both **the critiques** themselves and **some solutions** that others may want to apply.

Validity issues

- **Internal validity issues**: arise for a single benchmark/resource
- **External validity issues:** apply to the world of the task



NEURAL INFORMATION PROCESSING SYSTEMS

Construct validity

- **Construct validity** may be the most important property of a good evaluation resource (Raji et. al, 2021)
- Definition:

  The central question of external validity: ***how well does a resource measure what it purports to measure?***

- Unfortunately, this is more of an abstract **goal** than a concrete requirement
- Many of the issues with benchmarks can be traced back to construct validity!

Data issues

- **Data issues** are challenges to the validity of benchmarks that arise at creation time.
- Lifecycle problems
  - "Where" and "when" data is created
- Noise & spurious correlations
  - When bad examples harm the ability of the benchmark to discriminate
- Resource creator positionality & construct validity
  - "Who" created data for "what" task?
- Contamination
  - What happens when examples from the benchmark are trained on

Static datasets vs. dynamic realities

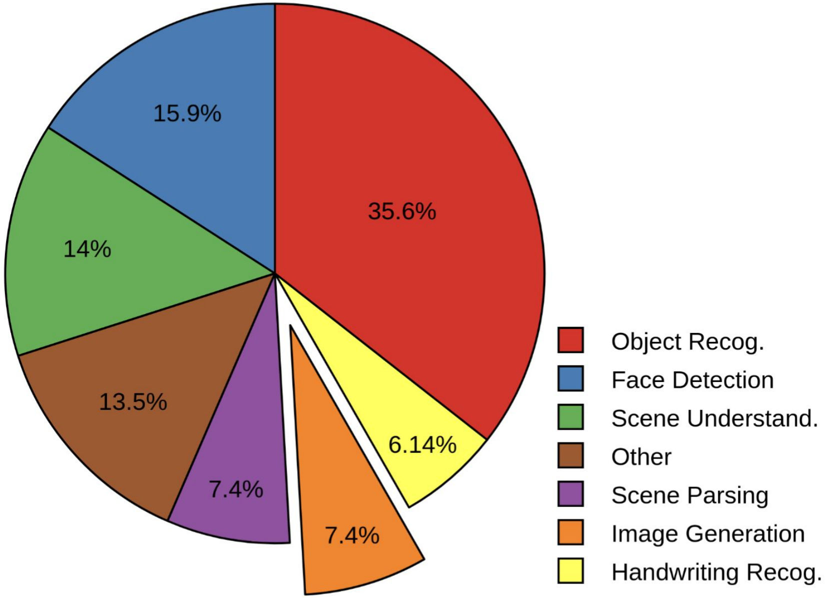
- Most dominant benchmarks are over **static** sets of exemplars collected once, held as test set forever.
- Static dataset lifecycle ends in a way at **release**
- However, reality is **dynamic**
- Static multiple choice question (MCQ) datasets “fail to reflect the evolving nature of human-AI interactions” (McIntosh et al.)
- Static datasets have no way to account for future improvements in model capabilities (Saxon et al, 2024)

Lifecycle challenges

- Producing *high-quality* data is costly and time-consuming

- Consequently we often rely on "reduced, reused, recycled" data

- 70% of computer vision datasets reuse data from other **domains** (Koch et. al 2021)



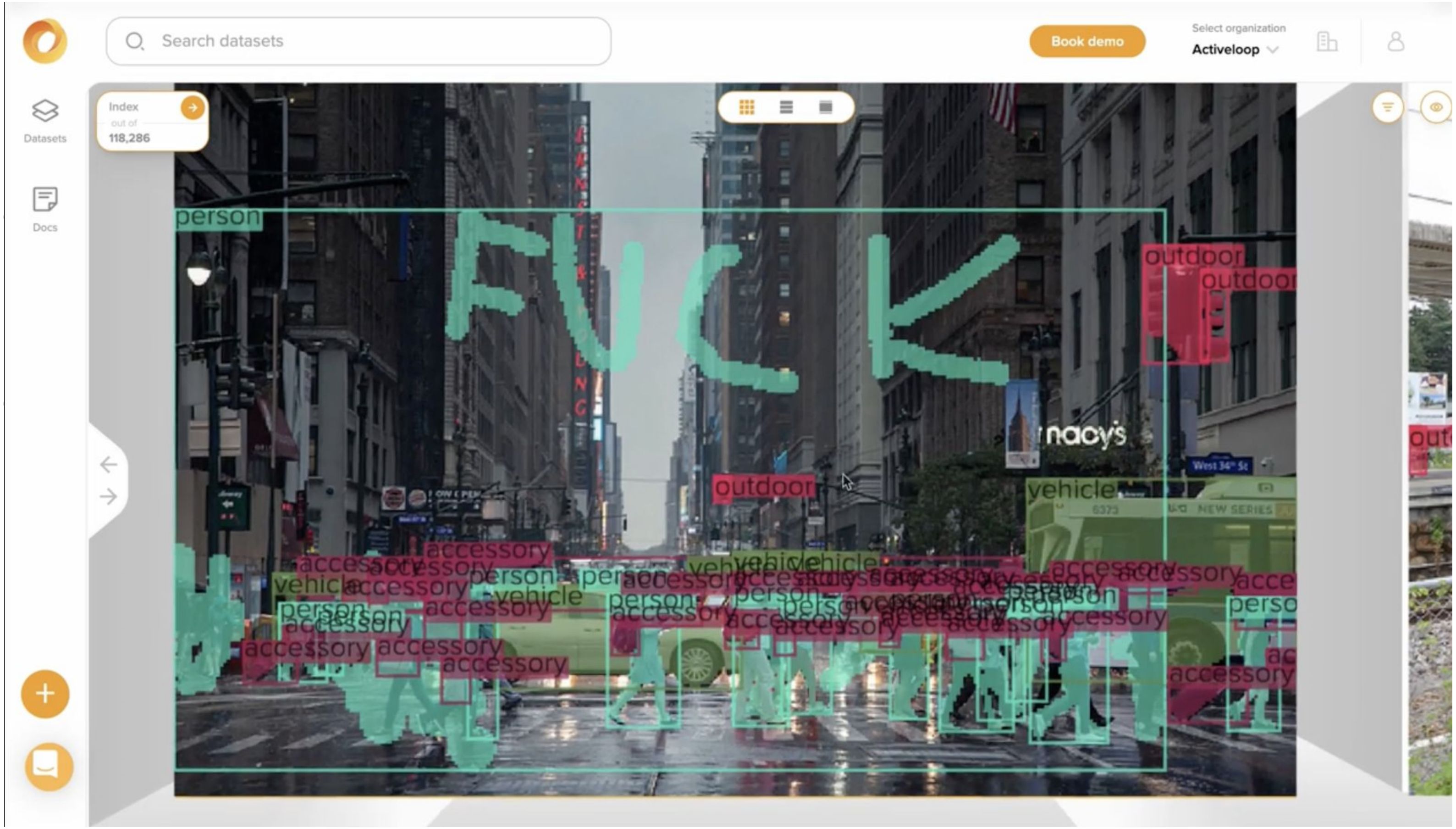
Koch et. al (2021): Original data source for *image generation* datasets, by task (only 7.4% were built for the purpose of image generation)

Reduced, reused, recycled data

- Issues in recycled data propagate
- Egregious examples of bad data have survived even in huge datasets
- We have improperly labeled images, such as this one from Imagenet
- But crowdsourced datasets like MS COCO have many more egregious labels and masks
- For example…

Label: car

A real example from MS COCO



NEURAL INFORMATION PROCESSING SYSTEMS

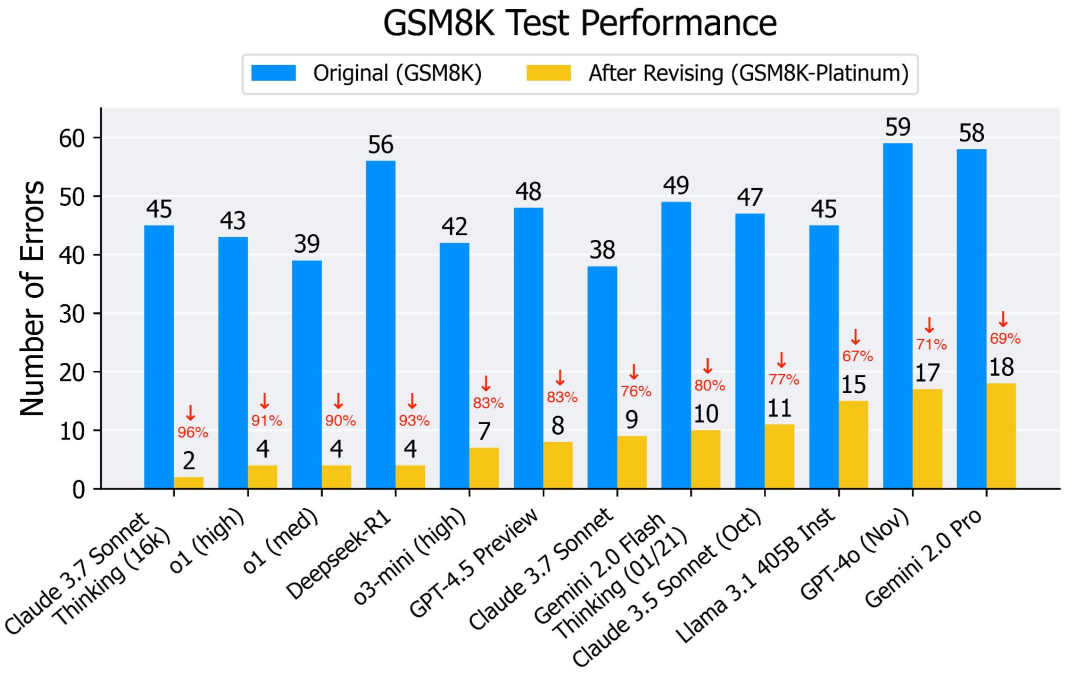
Noise & spurious correlations

- Bad labels are present all over.
- So what?
- Sometimes these present **artifacts**: systematic spurious correlations which models can use to cheat on tasks such as:
    - Medical scan classification (Oakden-Rainer et. al, 2019)
    - Entailment recognition (McCoy et. al, 2019)
- Bad labels also produce **noise**, where an LM's performance is effectively random

NEURAL INFORMATION PROCESSING SYSTEMS

Label noise

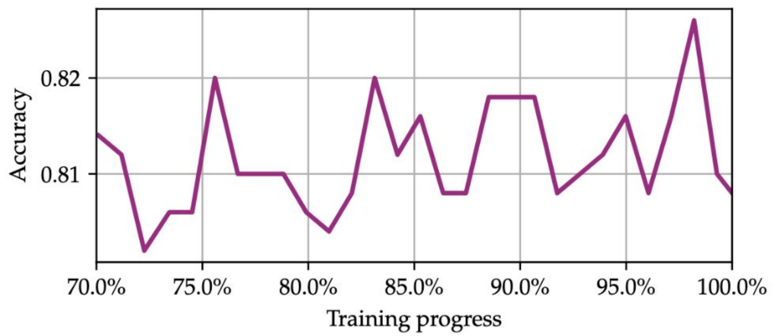
- Vendrow et. al (2025) remove bad examples from GSM8K
- After removing the noise ceiling (making 100% acc possible), **the ranking swaps**
- Noise isn't *just* a ceiling—it also sets the **resolution**
- **How can we deal with it?**



NEURAL INFORMATION PROCESSING SYSTEMS

One way to remove noise: Fluid benchmarking

- **Item response theory (IRT)**
- Capability as a **latent property** which we infer by not just # correct answers but **which difficulty level** of questions a test-taker answers right.
- Each question has a “difficulty” and “discriminability”
- Hoffman et al (2025) train an IRT model on benchmarks, use it to **select** the next test question given the **current capability estimate**
- Noise samples naturally have low discriminability and are ignored
- **Datasets become higher resolution** (monotonic with train)

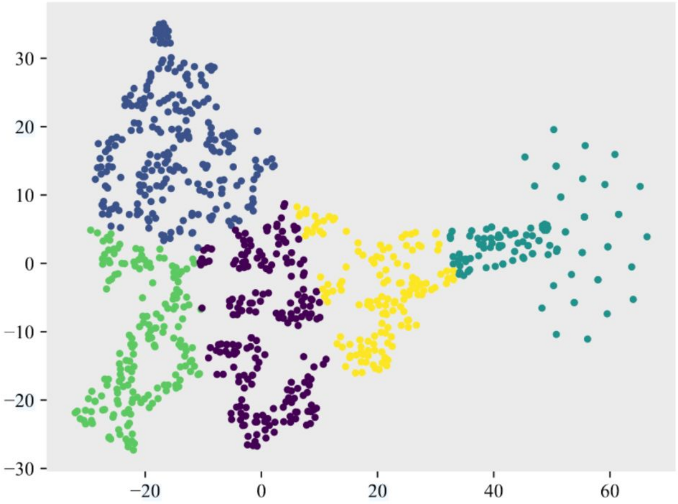


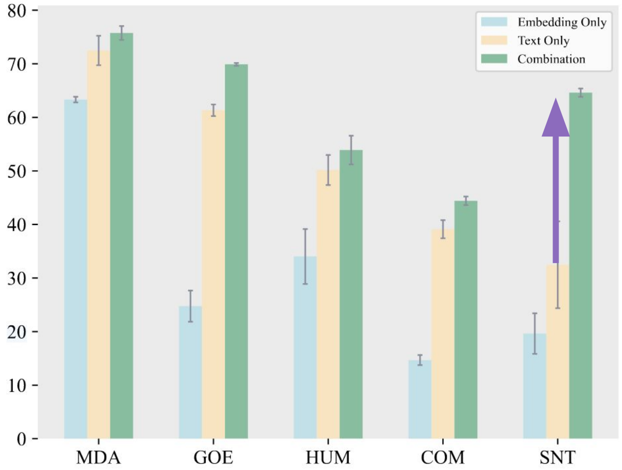
Positionality: subjectivity in annotation

- Many tasks we want to evaluate are **inherently subjective**
  - Humor
  - Hate speech
  - Safety (in text generation)
- Others may have **subjective attributes** at times
  - Acceptability of phrases (is this grammatical?)
    - Not all native speakers will agree on the grammaticality of all phrases
  - Entailment of statements (NLI)
    - Annotators may disagree on how well a sentence entails another
- Typically, the way we deal with this subjectivity is **average or majority vote**
- **Is this ideal?**

Dealing with subjectivity: annotator embeddings

- Deng et al (2023) introduce *annotator embeddings* to handle **inherent annotator disagreement** across humor detection, hate speech detection, and NLI
- They **keep the disparate annotations** and frame the learning task as:
- Predict label given input *and latent annotator descriptor*
- "annotation embeddings" give up **2x boost** to CLS performance over text alone





Positionality: definitions matter!

Data issues

Lifecycle

Noise

Positionality

Contamination

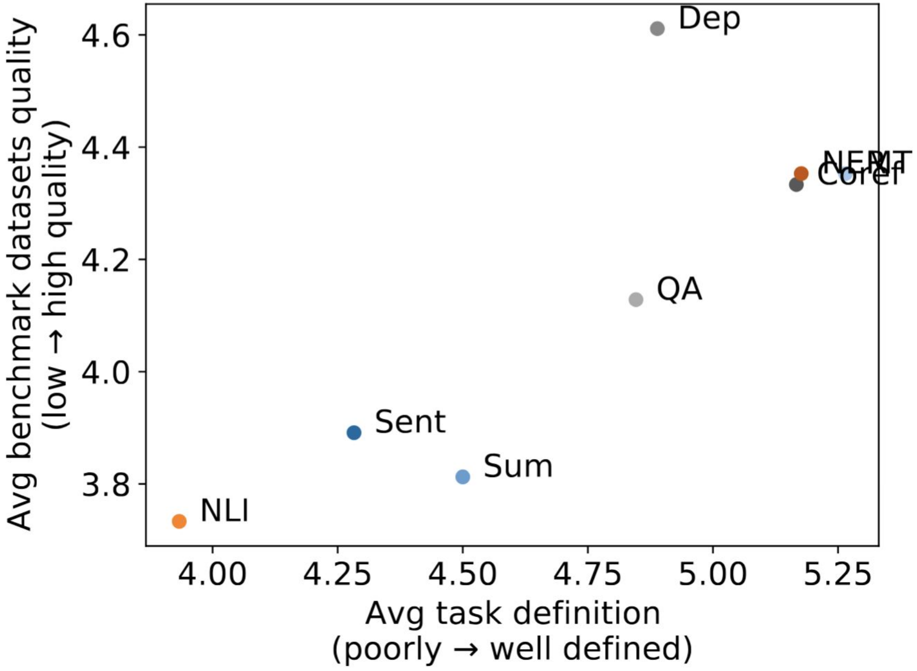
Measurement issues

Systemic issues

Epistemic issues

Concluding

- Positionality issues also include **how developers *understand* a task**
- From Subramonian et al. (2023):
- Same "task" often operationalized differently in different works (usually poorly defined too)
- Direct relationship between poor task definition and low dataset quality.
- **Poor conceptualization leads to poor data collection**



Positionality: values embedded in data source

Data issues

Lifecycle

Noise

Positionality

Contamination

Measurement issues

Systemic issues

Epistemic issues

Concluding

- Many influential LM datasets source test questions from web forums
- Examples: HellaSwag (Zellers et al, 2019), ETHICS (Hendrycks et al, 2021)
- These in turn get “recycled” into bigger resources like MMLU
- Many Reddit AITA dilemmas cover basic interpersonal drama
- Are these the bounds by which ethics in AI systems should be assessed?

NEURAL INFORMATION PROCESSING SYSTEMS

Positionality: values embedded in culture

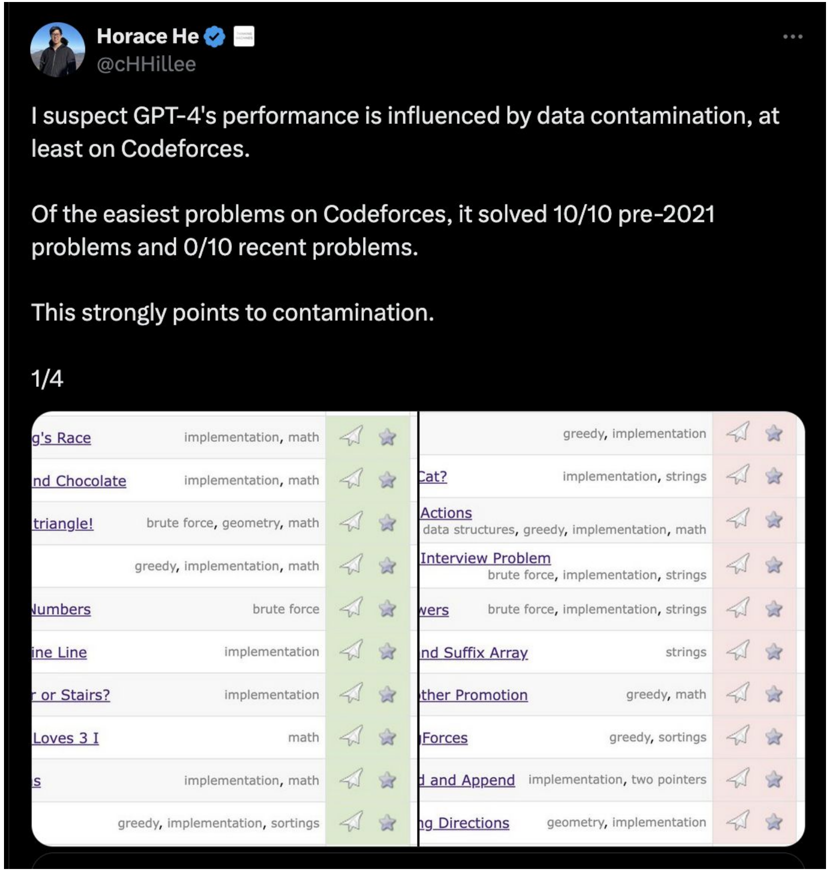
- **Conceptualization itself is culturally embedded!**
- Concepts like “stereotypes” and “offensive language” are contested and culturally contingent (Blodgett et al. 2021)
- Even **annotations themselves** can be culturally contingent (Oh et al. 2025)
- For example, in one study, East Asian annotators consistently preferred lower-valence (mid) answers on a Likert scale to Americans (Lee et al. 2002)
- These influences shape dataset curation!

Positionality: values embedded in community

- Benchmarks are **normative instruments** within the AI community.
  - Eg, investment be in safety benchmarks has successfully motivated work on alignment and safety from EA/x-risk perspectives
- Even as we look to build systems that can do everything, most of us (AI researchers) are **not experts** in most things
- Consequently, many domain-specific eval resources produced by computer scientists are not useful for experts in those domains
- Blagec et al. (2023) surveyed medical practitioners about clinical LM benchmarks: **they fail to capture how LMs meet doctor's needs**
- Benchmarks often **abstract tasks out of their social context** (Selbst 2019)
- Data work is undervalued relative to "model work" (Raji et al., 2021)

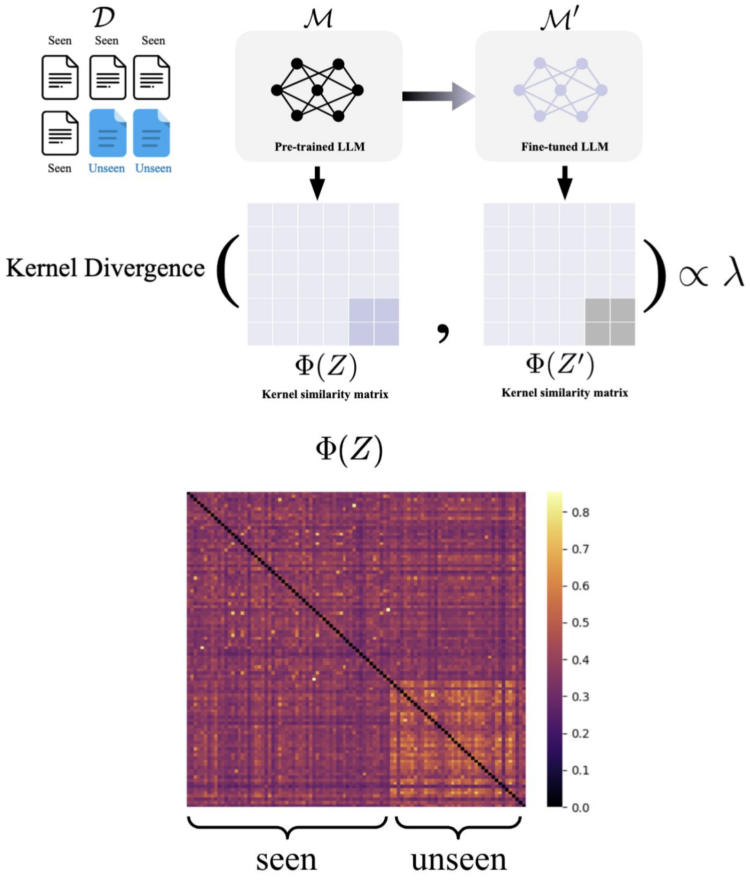
Contamination

- Benchmarks **contaminate** a model when their examples are present in a model's training data
- Fundamental challenge to the utility of an eval
- leads to poor **predictiveness** and **validity**
- GPT-4 performed perfect on pre-knowledge cutoff codeforces easy problems, and 0/10 on post-cutoff
- How can we address it?



"How contaminated is your benchmark?"

- Choi et al (2025) introduce *kernel divergence score* as a proposed white-box method to test for contamination
- Check the divergence of the RBF kernel of the embeddings of some test documents before and after fine-tuning the model on them
- If the embeddings don't diverge, it likely has already seen them



What's in my big data? (WIMBD)

- Directly search for observed model outputs in the training data (Elazar et. al 2023)
- It is possible that many emergent capabilities in LLMs are picked up in the training data
- "Let's think step-by-step" site shows up 10s of thousands of times in C4
- **This is a broader philosophical challenge for understanding generalization in the LLM era**

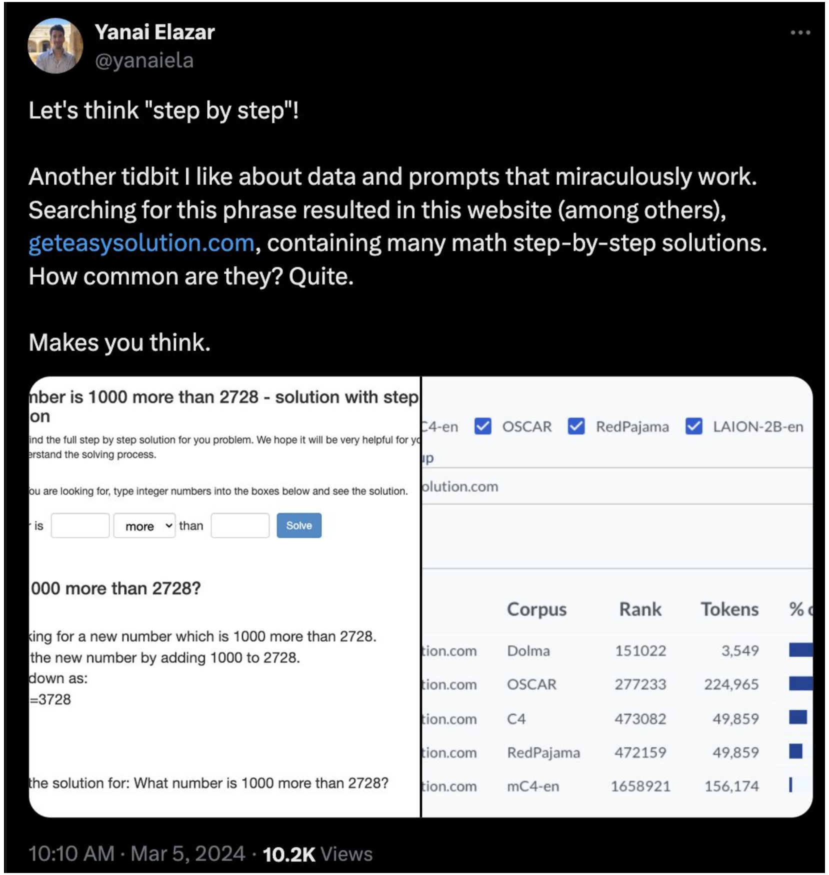


Image Contamination

Train Sets

Train Sets	RefCOCO	MSCOCO		Flickr		VQAv2	MMMU		MMMU-pro	Charxiv		MathVista	Geom-3K		Pixmo-Docs
	test	val	test	val	test	val	val	test	test	val	test	test	val	test	val
MSCOCO	100.0	2.7	0.1	0.2	0.5	87.7	0.0	0.0	0.0	0.0	0.0	2.6	0.0	0.0	0.0
LAION-2B	-	2.4	2.3	1.7	1.7	2.4	10.2	7.6	11.7	0.2	0.4	7.2	0.0	0.0	0.0
Flickr-30K	-	0.1	0.1	0.5	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pixmo-Docs	-	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.2	0.2	0.1	0.0	0.0	0.3
Geom-3K	-	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	3.3	49.0	51.6	0.0
ArXivQA	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	0.4	0.0	0.0	0.0	0.0

*Our original preliminary results.

Image Contamination

Data issues
Lifecycle
Noise
Positionality
Contamination
Measurement issues
Systemic issues
Epistemic issues
Concluding

- Risk level: Same.

MMUU (test)

Q: Which is not a negative outcome of <image 1>?

LAION-2B

Caption: Dredging and Capping | Onondaga Lake Cleanup

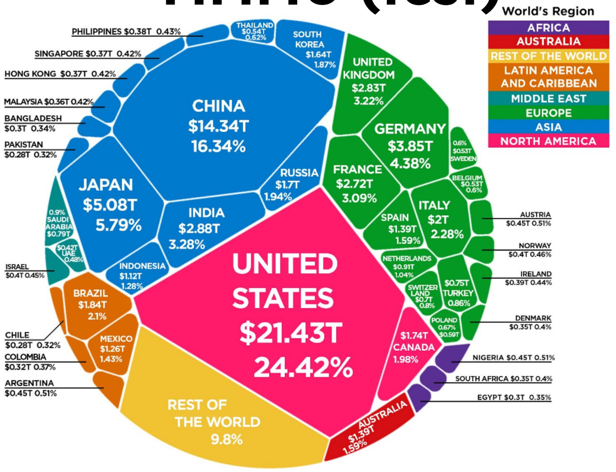
NEURAL INFORMATION PROCESSING SYSTEMS

*Our original preliminary results.

Image Contamination

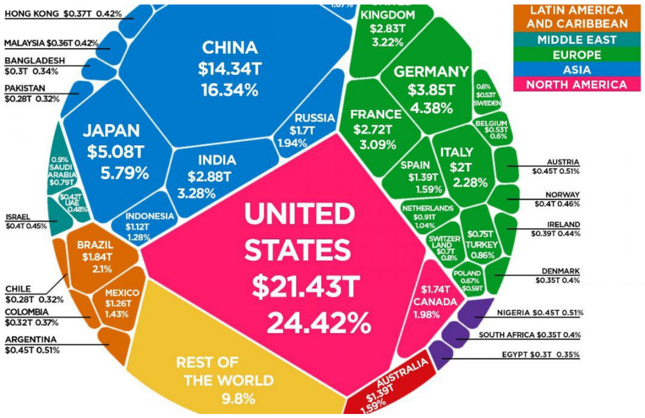
- Risk level: High.

MMMU (test)



Q: If a sociologist says that nations evolve toward more advanced technology and more complex industry as their citizens learn cultural values that celebrate hard work and success, she is using _ theory to study the <image 1>.

LAION-2B



Caption: The $88 Trillion World Economy in One Chart

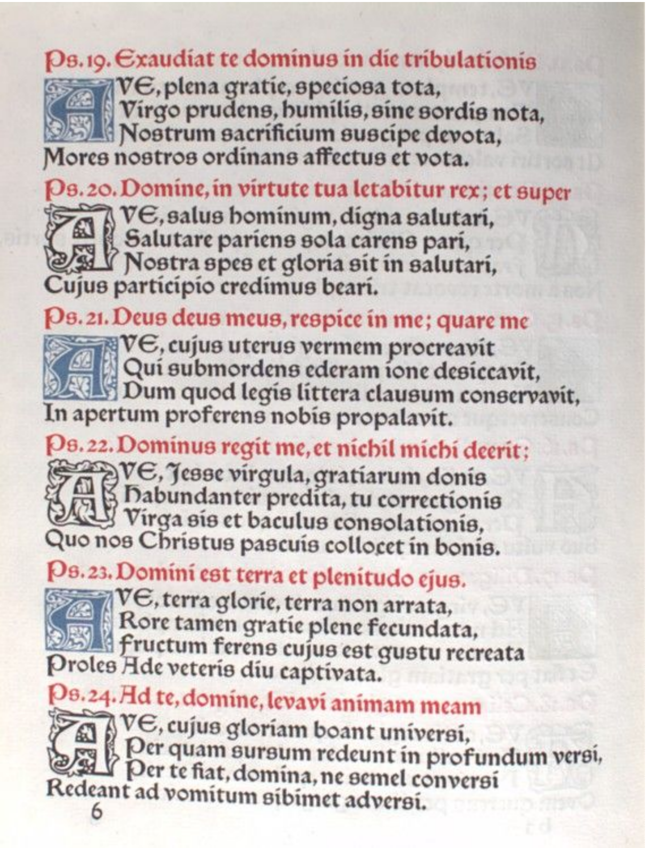
*Our original preliminary results.

Measurement issues

- **Measurement issues** refer to challenges in benchmarking which arise from:
- Poorly-developed **rubrics** for annotation
- Challenges with **metrics** used to grade or score the outputs of a model under test
  - **LM judges**
  - **Learned metrics**
  - **Algorithmic metrics**

Measurement issues: rubrics

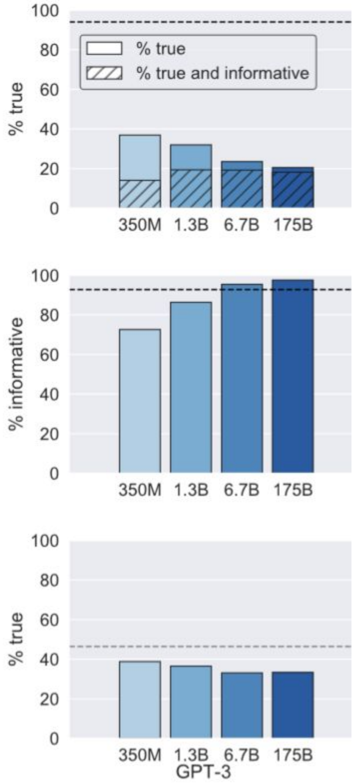
- **Rubrics**, from the Latin name for red ink writing, refer to scoring guidelines for writing assignments in US education (Popham, 1997)
- The rubric pattern has grown popular in AI evaluation, both as for:
  - Guiding human annotators (usually non-expert crowdworkers) in scoring outputs
  - Guiding **LM judges** in scoring outputs
- How do we know when we're writing good rubrics? Does rubric quality matter?


Ps.19. Exaudiat te dominus in die tribulationis
AVE, plena gratie, speciosa tota,
Virgo prudens, humilis, sine sordis nota,
Nostrum sacrificium suscipe devota,
Mores nostros ordinans affectus et vota.
Ps.20. Domine, in virtute tua letabitur rex; et super
AVE, salus hominum, digna salutari,
Salutare pariens sola carens pari,
Nostra spes et gloria sit in salutari,
Cujus participio credimus beari.
Ps.21. Deus deus meus, respice in me; quare me
AVE, cujus uterus vermem procreavit
Qui submordens ederam ione desiccavit,
Dum quod legis littera clausum conservavit,
In apertum proferens nobis propalavit.
Ps.22. Dominus regit me, et nichil michi deerit;
AVE, Jesse virgula, gratiarum donis
Habundanter predita, tu correctionis
Virga sis et baculus consolationis,
Quo nos Christus pascuis collocet in bonis.
Ps.23. Domini est terra et plenitudo ejus.
AVE, terra glorie, terra non arrata,
Rore tamen gratie plene fecundata,
Fructum ferens cujus est gustu recreata
Proles Ade veteris diu captivata.
Ps.24. Ad te, domine, levavi animam meam
AVE, cujus gloriam boant universi,
Per quam sursum redeunt in profundum versi,
Per te fiat, domina, ne semel conversi
Redeant ad vomitum sibimet adversi.
6

NEURAL INFORMATION PROCESSING SYSTEMS

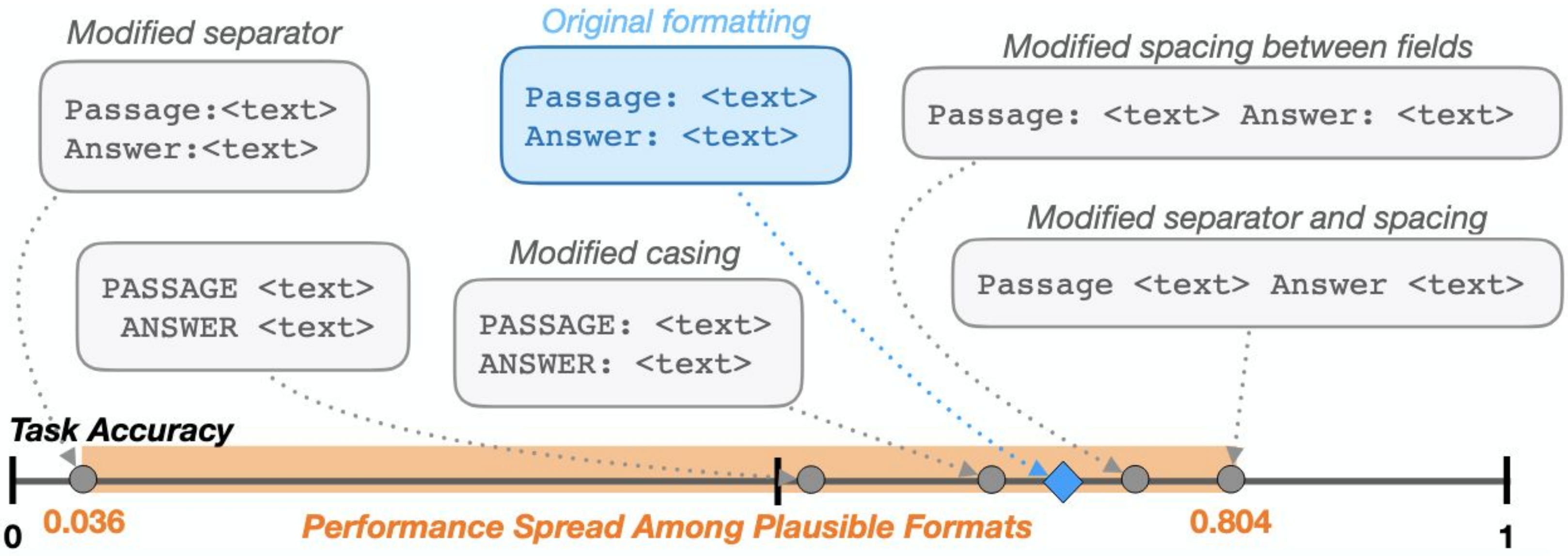
Measurement issues: Judges

- **LM-as-a-judge** is a technique for evaluating natural text since at least TruthfulQA (Lin et. al, 2021)
- An LM is prompted to provide a numerical score given a set of requirements (eg. truth, informative, etc)
- Advantages:
  - you can specify your desiderata in natural language
  - easily handles open-ended generations
  - (can be) straightforward to implement
- Disadvantages:
  - Brittle to prompt variations
  - Rubric implementation complications
  - Self-bias



Writing good rubrics

- **Rubric quality matters a lot**
- LM judges are extremely sensitive to minor variations in prompt phrasing (Sclar et al, 2023)
- How can we improve them?



NEURAL INFORMATION PROCESSING SYSTEMS

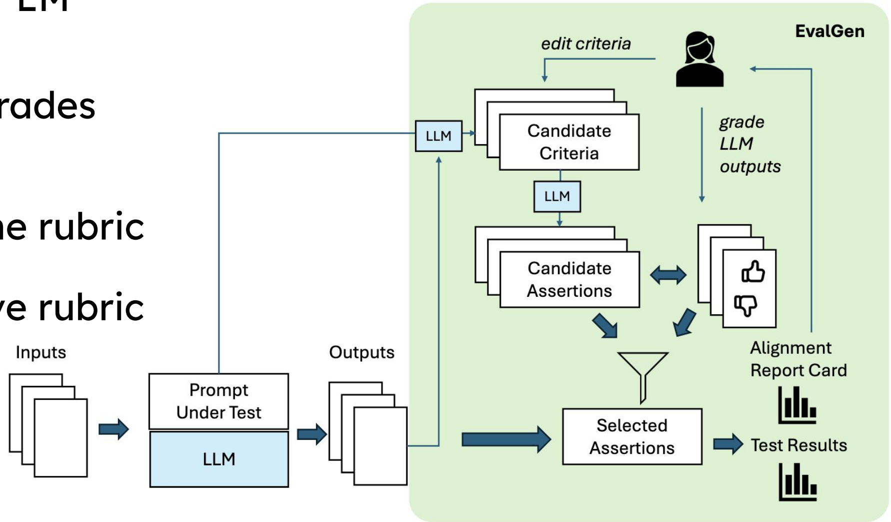
Writing good rubrics: EvalGen

- There's a catch-22 in rubric writing:
  - to grade some outputs, you need criteria
  - to know the criteria, you have to understand what errors occur
- This is why educators update rubrics as they grade (Shankar et. al 2024)

- A similar process can be adopted for LM judges, given seed rubric:

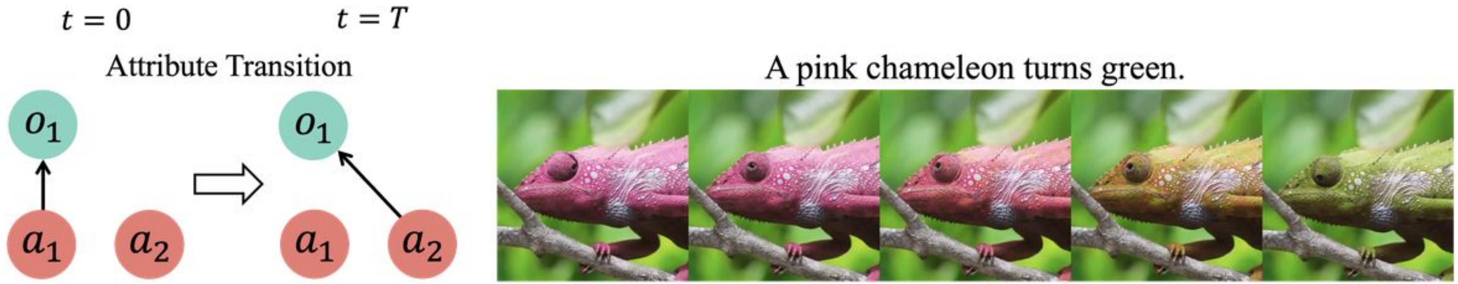
1. Human reviews set of outputs and grades them
2. LM grades according to rubric
3. Human checks which assertions in the rubric are agreed and which aren't
4. LM suggests edit locations to improve rubric

Should be adopted for benchmark dev



Subjectivity and criteria drift

- This problem of **criteria drift** is widespread (Shankar et. al, 2024)
- Even **static rubrics** may be problematic:
  - As models evolve and improve, the types of errors they make will change
  - Rubrics well-suited to poorly performing models may not apply as performant models have more nuanced weaknesses
- Example: 2023-2024 video models struggled with temporal consistency (Feng et. al, 2024) but new video gen models like Veo don't
- Are videogen benchmarks meant to test this evergreen?



NEURAL INFORMATION PROCESSING SYSTEMS

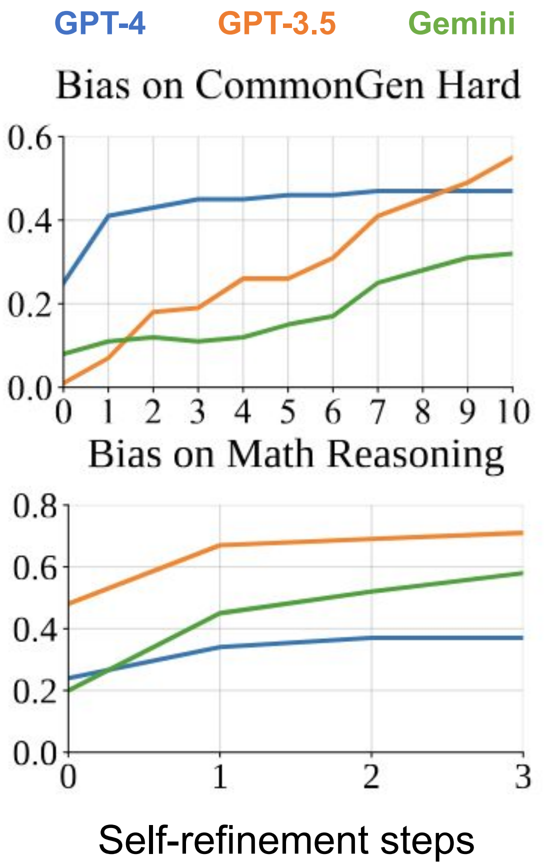
Self-bias in LM judges

- **Self-bias** is the phenomenon of LM judges implicitly preferring text generated by their own base model

- Xu et al (2024) demonstrate that across multiple models, multiple tasks LM judges self-bias (distance from scores assigned by external annotator) grows as more self-refinement steps are made by the model

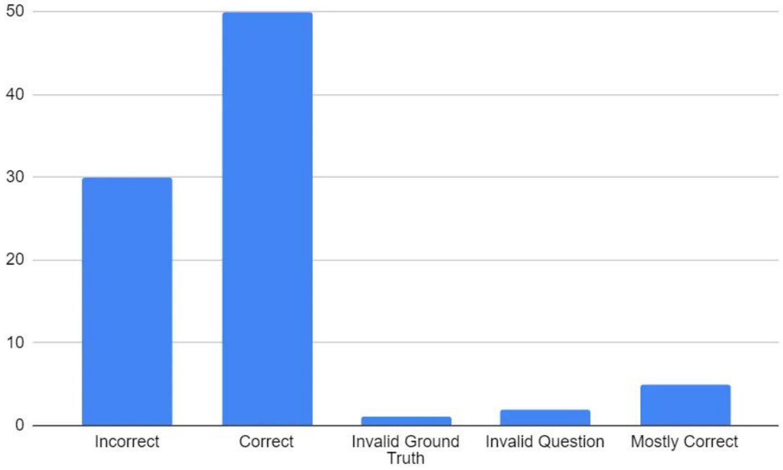
- **LM judges need to be carefully checked**

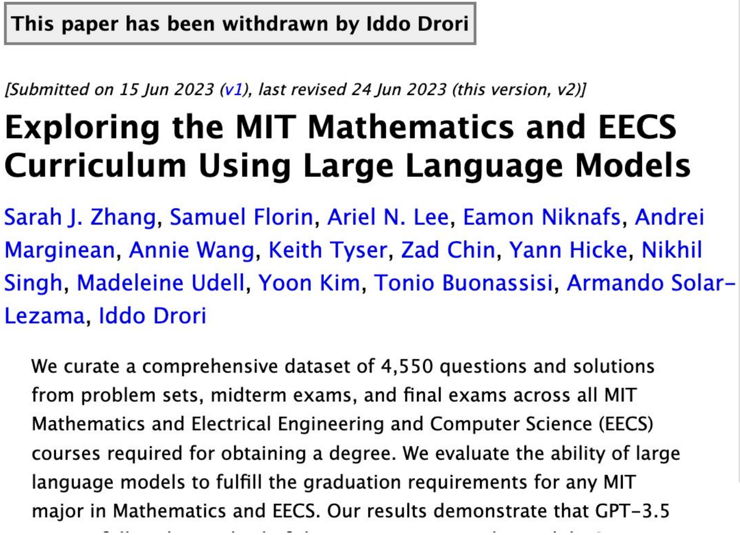
- Failure example: “Exploring the MIT Mathematics and EECS Curriculum” paper



Exploring the MIT Math and EECS with GPT-4

- Claim: *GPT-4 can get 100% on the core MIT EECS exams* (Zhang et. al, 2023)
- Among the problems: incorrect grades assigned by GPT-4 as a judge! (Chowdhuri et. al, 2024)
- **Paper was withdrawn from arXiv**





This paper has been withdrawn by Iddo Drori

[Submitted on 15 Jun 2023 (v1), last revised 24 Jun 2023 (this version, v2)]

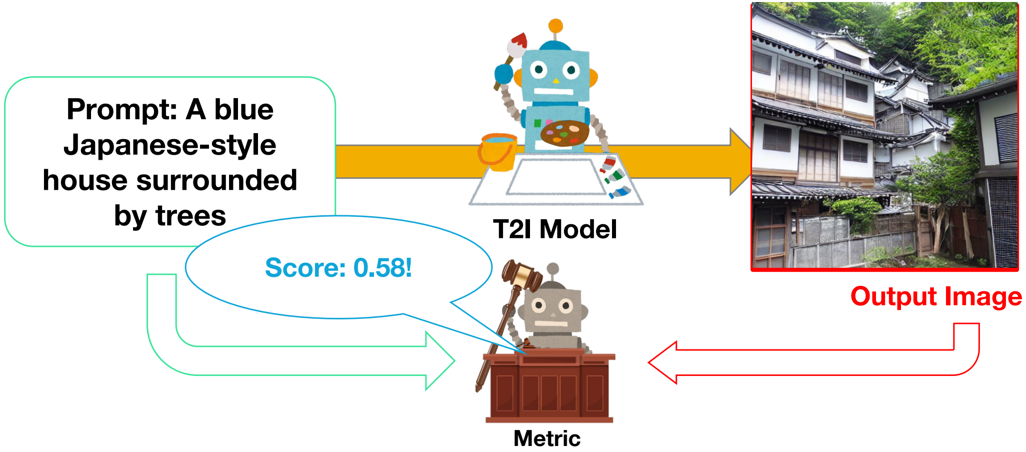
Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models

Sarah J. Zhang, Samuel Florin, Ariel N. Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, Madeleine Udell, Yoon Kim, Tonio Buonassisi, Armando Solar-Lezama, Iddo Drori

We curate a comprehensive dataset of 4,550 questions and solutions from problem sets, midterm exams, and final exams across all MIT Mathematics and Electrical Engineering and Computer Science (EECS) courses required for obtaining a degree. We evaluate the ability of large language models to fulfill the graduation requirements for any MIT major in Mathematics and EECS. Our results demonstrate that GPT-3.5

Measurement issues: Metrics

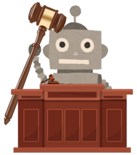
- Many other metrics which aren't strictly LM judge + rubric have been proposed
- For example, in text-to-image evaluation **prompt consistency** is the evaluation task of assigning a numerical score to the alignment of a generated image to its input prompt
- Multiple classes of prompt consistency metrics



Classes of text-to-image faithfulness metrics

- Two predominant classes of prompt-consistency metrics are:
- **Embedding-correlation** metrics and **VLM-VQA metrics**

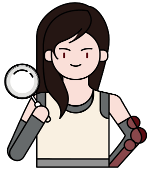
Embedding Correlation



CLIPScore

- Example: CLIPScore (Hessel et. al, 2021)
- Embed image & prompt with CLIP, return cosine similarity of embeddings
- Cheap, fast
- Scores aren't attributable
- Considered less performant

VLM question-answering



TIFA

- Example: TIFA (Hu et. al, 2023)
- **Use LM to generate list of requirements from prompt**
- **Check if each requirement is met in image with vision LM**
- Expensive
- Scores are attributable to natural language requirements
- Considered more performant

NEURAL INFORMATION PROCESSING SYSTEMS

Flawed meta-evaluation establishes superiority

- Previously, superiority of VLM-VQA over correlation based on correlation to human judges **over unrelated images**
- Relative quality of unrelated images is *inherently subjective!*

"A blue house with a tree"

2.7/5

"A black cat reading a book"

3.4/5

Higher score

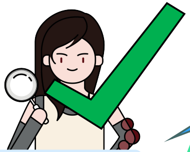
Human Annotators

0.81

0.78

CLIPScore

0.5



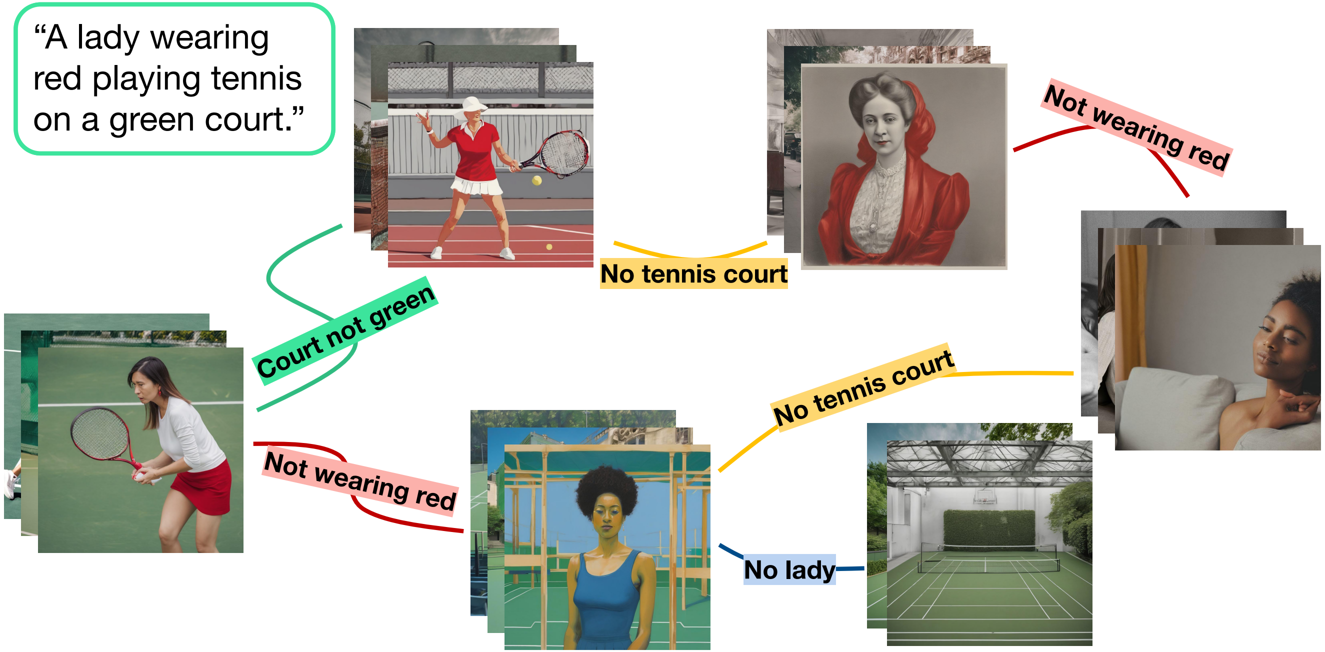
0.663

TIFA

Aligned to *average* of human preferences

Testing which is better with T2IScoreScore

- Saxon et. al (2024) instead analyzed T2I metrics with **semantic error graphs**
- Each graph contains populations of **related images** organized by counterfactual **error count** relative to the prompt
- Performance is judged by how well LMs *reconstruct* the error graph by ordering

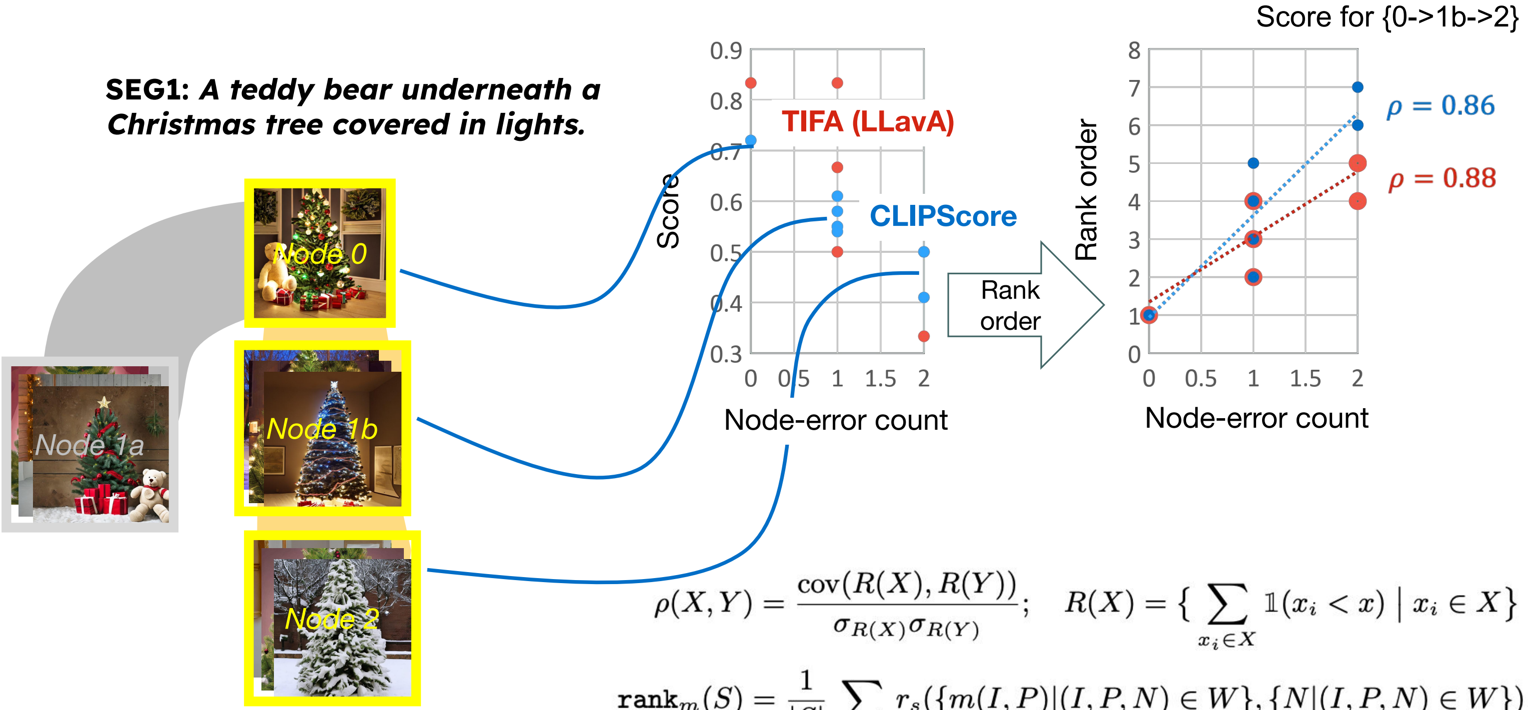


NEURAL INFORMATION PROCESSING SYSTEMS

Testing which is better with T2IScoreScore

Ordering Meta-Metric: Spearman Correlation

SEG1: *A teddy bear underneath a Christmas tree covered in lights.*

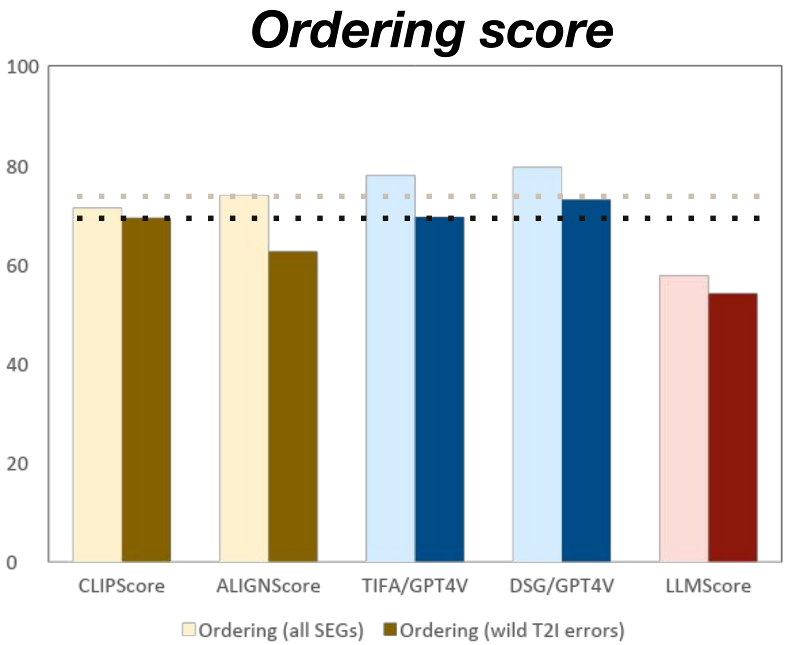


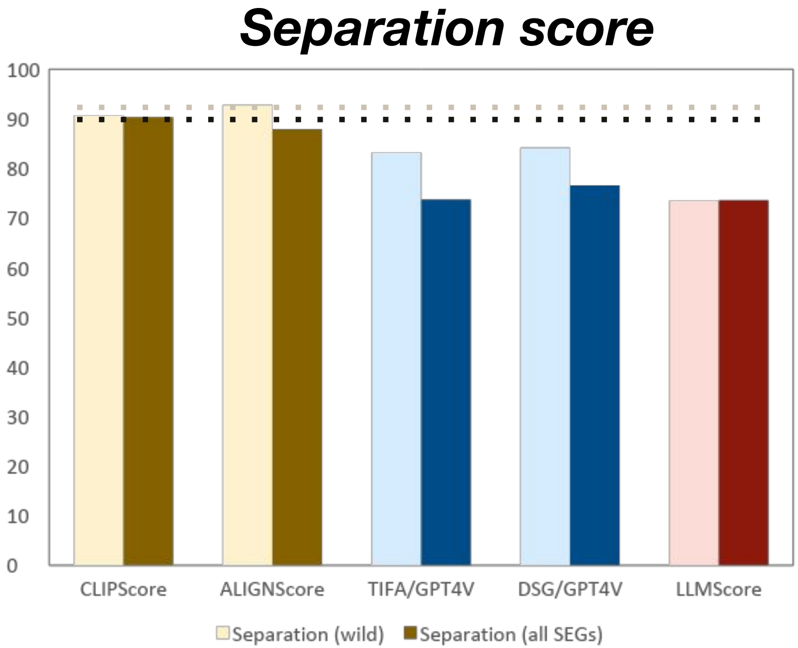
$$\rho(X,Y) = \frac{\operatorname{cov}(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}; \quad R(X) = \{ \sum_{x_i \in X} \mathbb{1}(x_i < x) \mid x_i \in X\}$$

$$\texttt{rank}_m(S) = \frac{1}{|S|} \sum_{W \in S} r_s(\{m(I,P) | (I,P,N) \in W\}, \{N | (I,P,N) \in W\})$$

Measurement issues: Metrics

Data issues
Measurement issues
Rubrics
Judges
Metrics
Systemic issues
Epistemic issues
Concluding





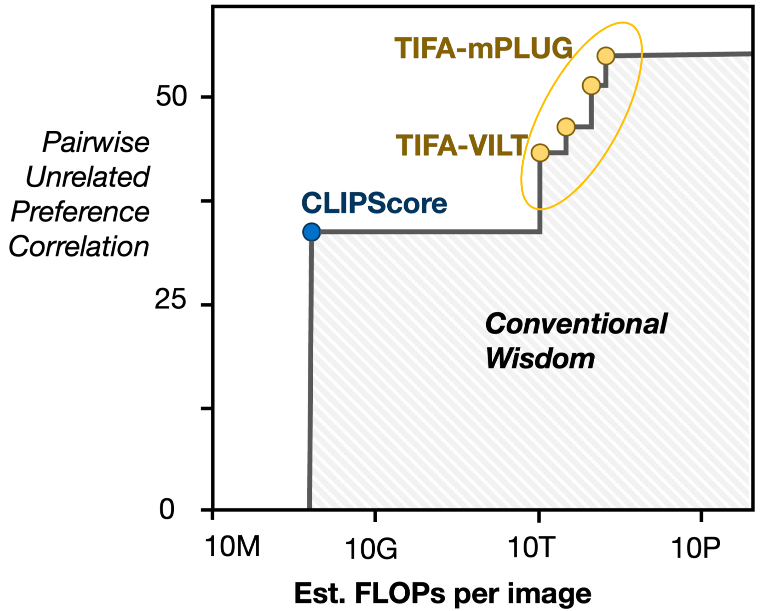
Embedding Correlation

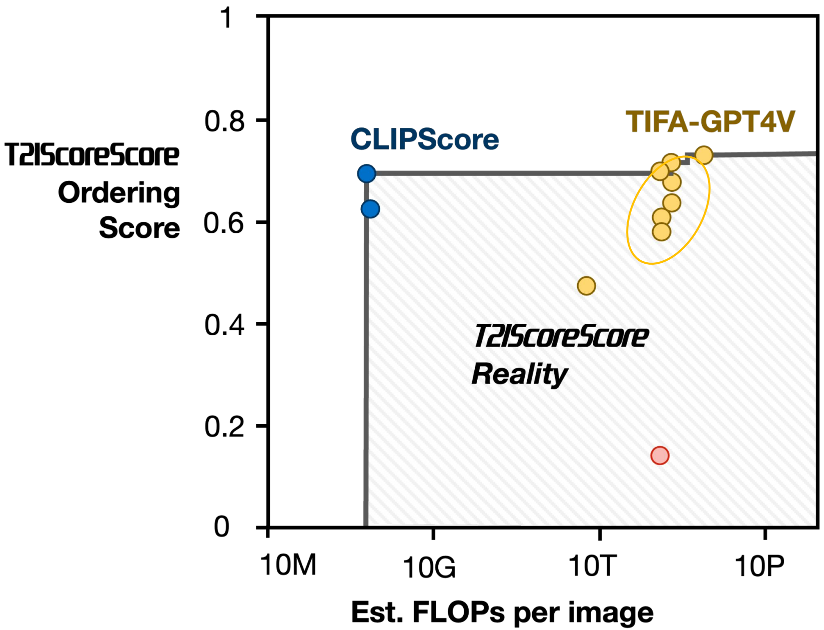
VLM question-answering

Others

Measurement issues: Metrics

- We find that the seemingly dominant versions of TIFA fall **below the pareto frontier** against CLIPScore for cost-performance
- Much more expensive VLMs like GPT4V needed to perform
- Important to meta-evaluate in an **ecologically valid manner!**





Systemic issues

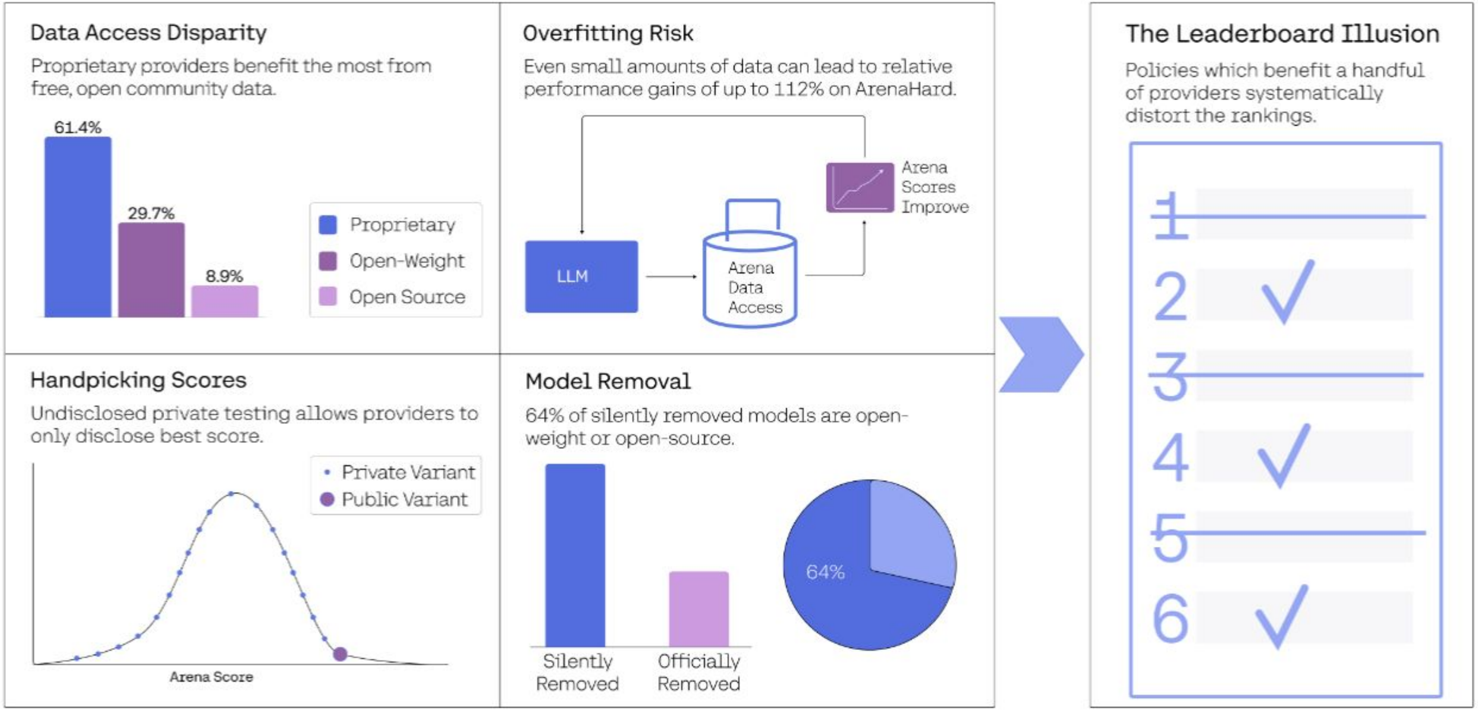
- **Systemic issues** are the cases where the **manner in which an evaluation is conducted or made** are problematic.
- The **leaderboard illusion**: how multiple submissions can game leaderboard style benchmarks
- Illusory improvements in performance from evaluating against **weak baselines**
- (Mis)-**reporting variance**: how choosing a best-of-n result may skew performance

Beware: Goodhart's Law

Once a measure becomes a target it ceases to be a good measure.

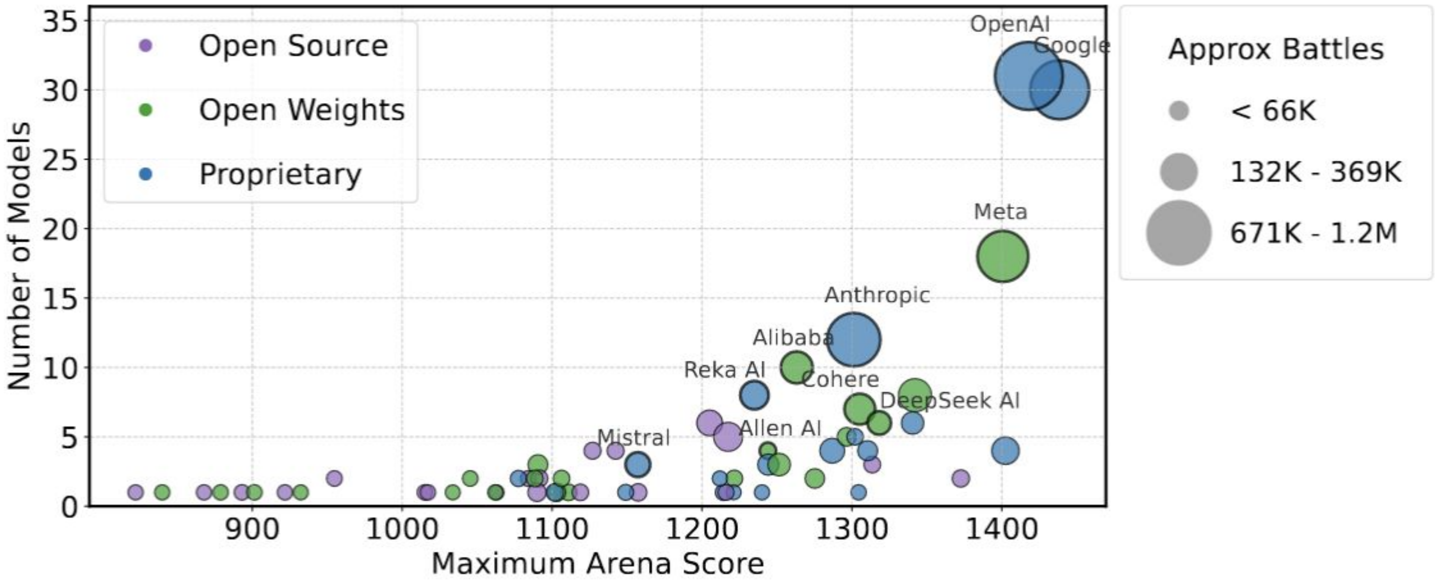
Leaderboard illusion (Singh et. al, 2025)

- Claim: Rankings on leaderboards (eg. LMSys) are very distorted!
- Goodhart's law applies as firms use arena scores in marketing (Saxon et al, 2024)



Leaderboard illusion (Singh et. al, 2025)

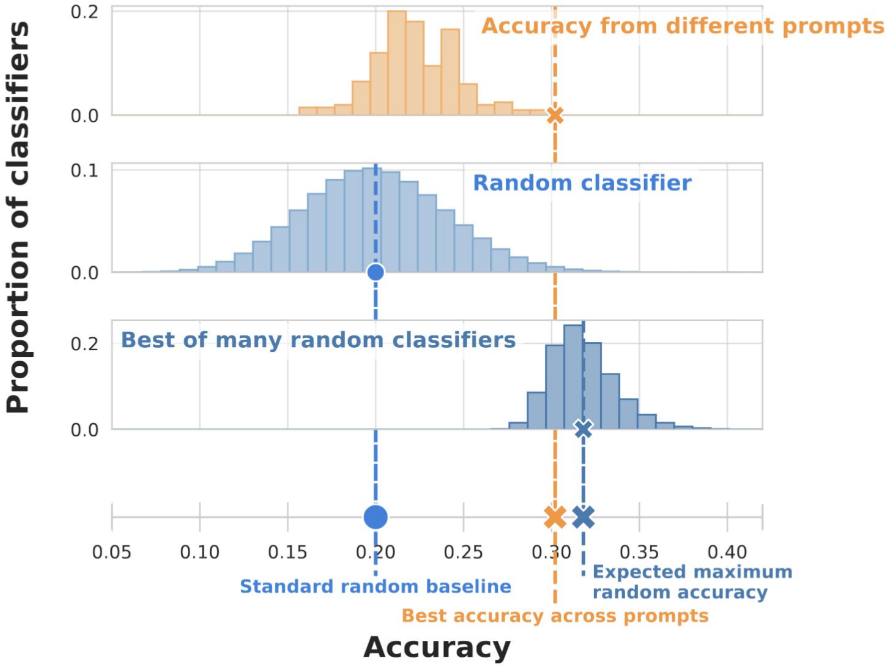
- Support: clear correlation between number of models a provider submits and score
- This gives the opportunity for providers to learn, gaming?



NEURAL INFORMATION PROCESSING SYSTEMS

What is wrong with multiple attempts?

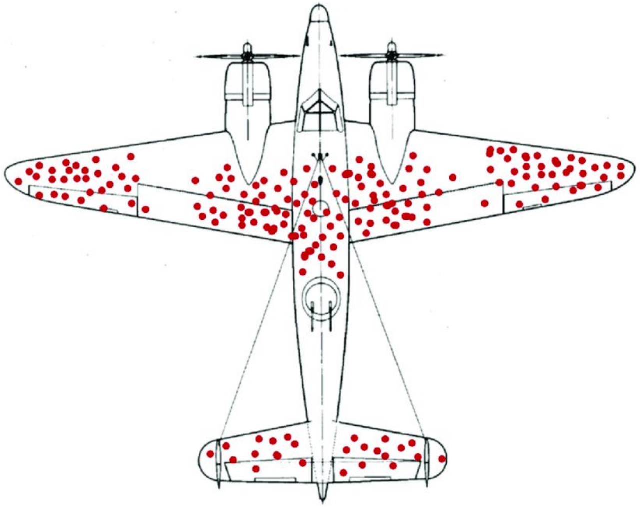
- Reporting asymmetrical best-of-n is also a statistical problem!
- Yauney and Mimno (2024): many methods fail to even beat a **random baseline** when its sampled with the same best-of-n process that models often are
- Leaderboard illusion: doing this not to random baselines, **but competing models**



NEURAL INFORMATION PROCESSING SYSTEMS

Intra-model variance (Reuel et al, 2024)

- The root of many of these statistical reporting issues is **intra-model variance**
- Reporting a single number for a new method or model, rather than the results according to multiple seeds, temperatures, etc is crucial to distinguish **signal from noise**
- The wider the intra-model variance, the lower resolution the benchmark is (as there is a significant **noise band** around any given score)
- Rankings are unreliable in this case
- Unfortunately, it is both expensive to do full extra training runs for this purpose
- But if we can’t, what are we doing here?

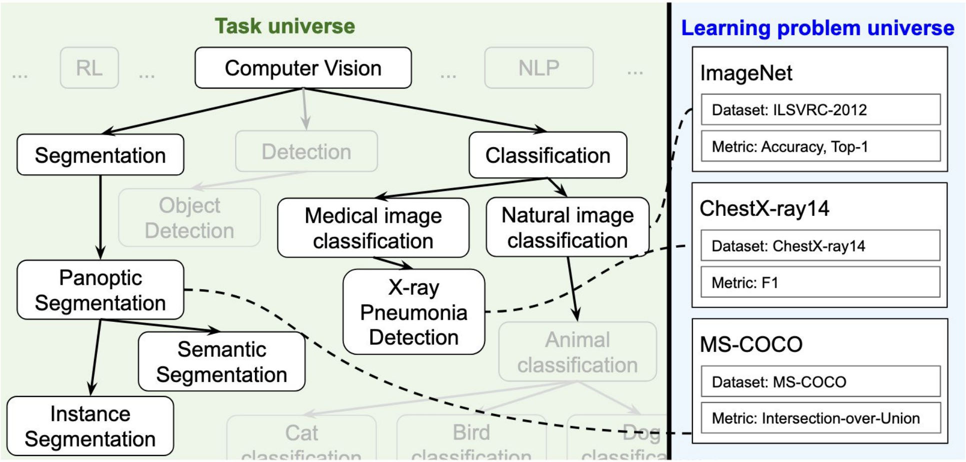


Epistemic issues

- Finally, we go most abstract:
- **Epistemic issues** are the core failures of conceptualization and operationalization that cut across all the more concrete issues below.
- Here we discuss:
- The **distinction between "tasks"** and "learning problems", and how the conflation of the two within a **task universe** causes problems
- Confusions of **map and territory** when thinking about AI
- The pitfalls of treating benchmarks like human **psychometrics**

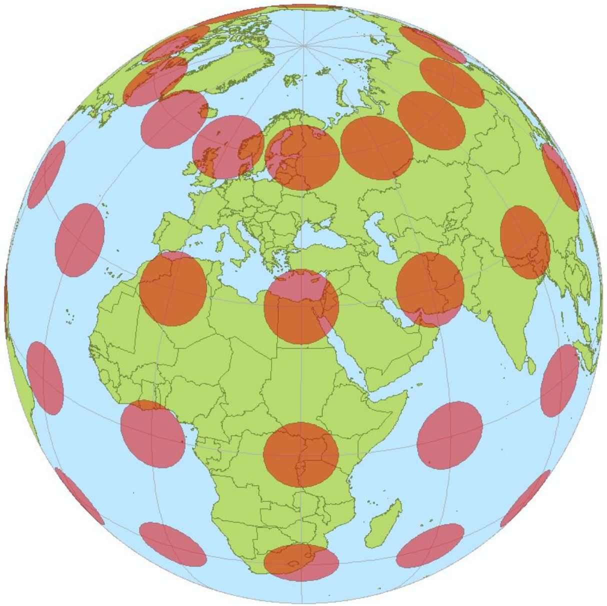
Tasks vs learning problems

- “Tasks” are often conceptualized at multiple scales (within a “task universe”)
- Foundation models represent attempts to create systems for higher and higher level tasks
- **“Learning problems” are distinct from tasks**. Imagenet is a learning problem that purports to capture the task of natural image classification. (Liao et. al 2024)
- Does it?



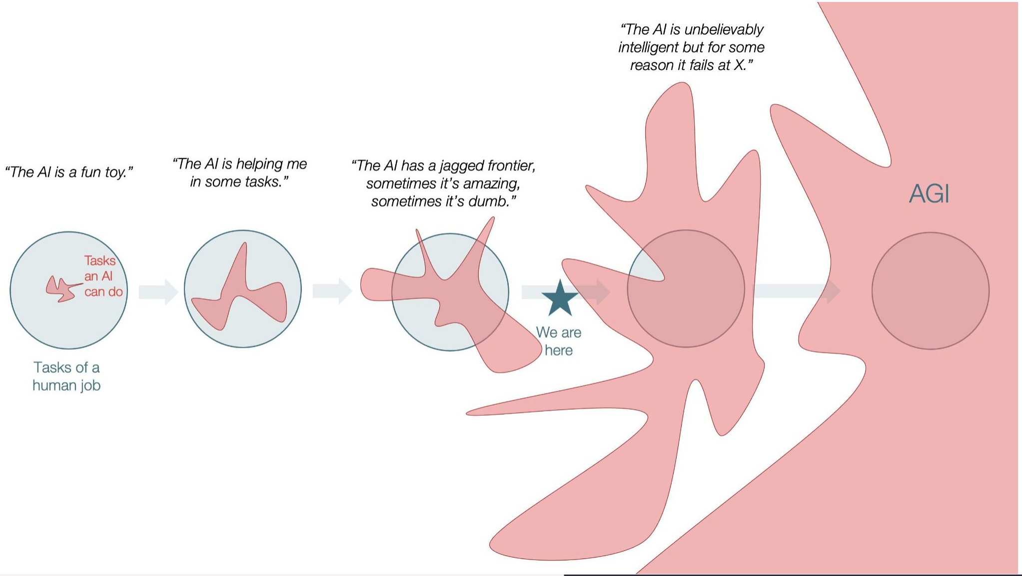
Map-territory and "wishful mnemonics"

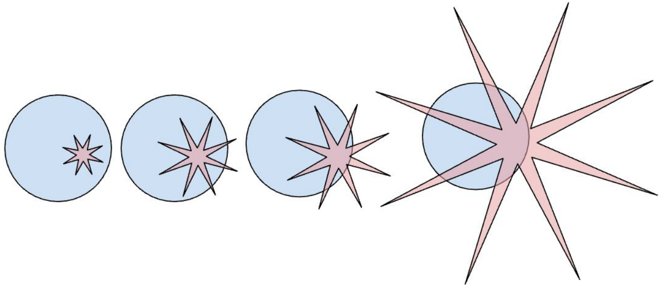
- Mitchell (2021) presents fallacies in AI thought which hamper progress toward AGI
- **Wishful mnemonics** are a critical fallacy which chronically plagues benchmarkers
- Calling a task "reasoning" or "understanding" is a wishful mnemonic (McDermott 1975)
- When we do this, we beg the question that a model's performance on that **task** represents an **abstract capability**—does it though?
- Chronic failures of models to generalize across eg reasoning tasks suggests the answer is no
- Coupled with sloppy application of **learning problems** to these **tasks** we often seriously confuse map and territory



Psychometrics trap

- Attempts to develop generalized “psychometrics” for AI systems suffer from all these epistemic problems
- Standardized tests work for humans because of “shared architecture”
- The “jagged frontier” problem complicates attempts to develop AI psychometrics
- There are two competing visions that are impossible to reconcile now





Tomas Pueyo, Colin Frasier (Twitter)

Summary

- **Noise** can come from data, metrics, or variance problems
  - **Bad examples** create both a noise ceiling (max meaningful performance) and **lower resolution**.
  - Poor statistical (systemic) practices also **lower resolution** through variance-based noise bands
  - Resolving either of these issues may cause the ranking of systems to completely swap.
- **Careful conceptualization** is crucial: construct validity is necessary for both good data collection and metric design/validation
  - Task vs learning problem
- **Positionality** of humans throughout the benchmarking process is often overlooked.
  - Bad samples can come from exhausted crowdworkers.
  - AI researchers lack domain expertise.
  - Hidden cultural assumptions drive annotator, producer choices.

Open Questions

- **What counts as contamination** when we are looking for systems to do everything?
- Do the same generalization weaknesses exist in LLMs that specialized models had?
- Are we producing “intelligence” at all if we have to explicitly train the model on every individual task?
- Some of these will come up in the panel at the end, so stick around :)

NEURAL INFORMATION PROCESSING SYSTEMS

What's Next?

Xiang Yue
Carnegie Mellon University
(Now at Meta)

https://benchmarking.science/slides.pdf

Agenda

- What's Measured? (1:30PM - 2:10PM)
- What's Missed? (2:10PM - 2:40PM)
- What's Next? (2:40PM - 3:15PM)
  - Towards dynamic and agentic benchmarking
  - Towards real-world benchmarking
  - Other emerging practices
  - Some proposals
- Panel Discussion (3:20PM-4:00PM)

NEURAL INFORMATION PROCESSING SYSTEMS

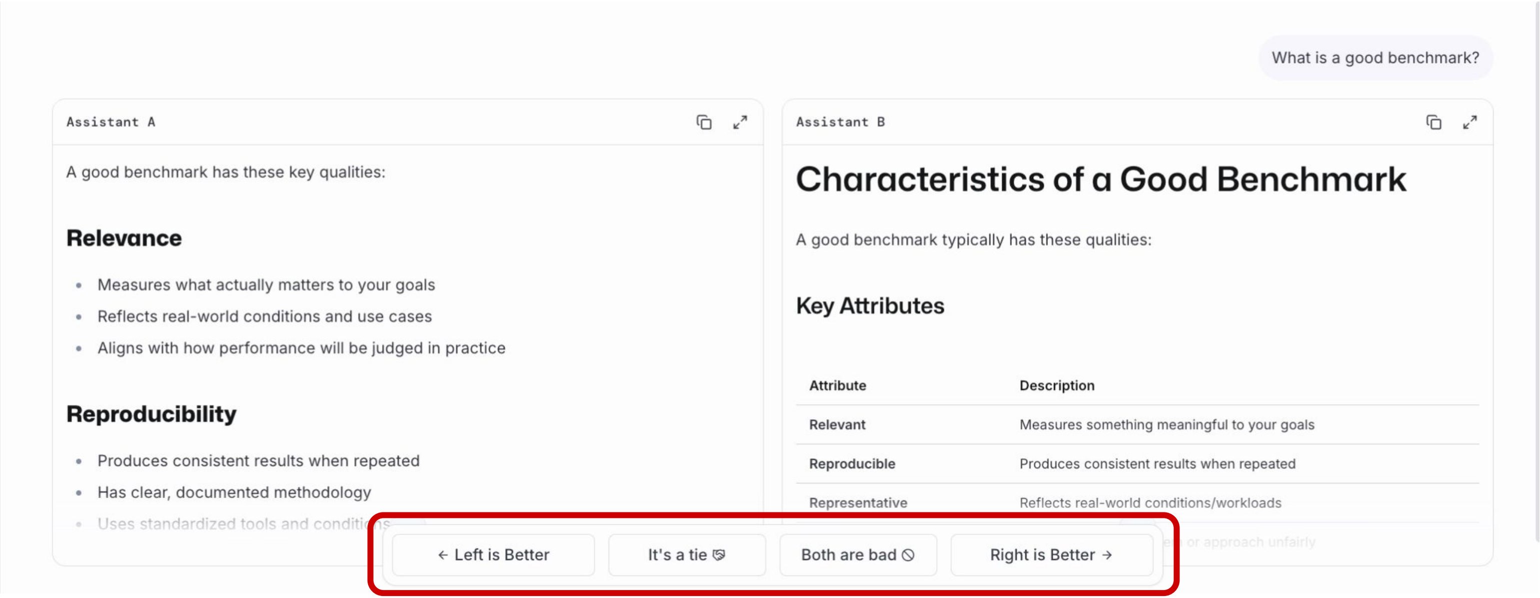
Dynamic & “Living” Benchmarks

- Static benchmarks quickly become saturated by frontier models.
- The risk of benchmark items contaminating training data is high.
- Static evaluation is rapidly obsolete as model knowledge evolves.
- Dynamic and "Living" Benchmarks are necessary to ensure relevant evaluation.

NEURAL INFORMATION PROCESSING SYSTEMS

LM Arena

- LM Arena (previously Chatbot Arena) pits two anonymous models against each other in randomised “battles.” Users chat with both models side-by-side and vote for their preferred answer.



Wei-Lin Chiang, et al., Chatbot Arena. 2024.

NEURAL INFORMATION PROCESSING SYSTEMS

LM Arena Leaderboard

Rank	Rank Spread (Upper-Lower)	Model	Score	95% CI (±)	Votes	Organization	License
1	1 ↔ 2	gemini-3-pro	1492	±8	9,799	Google	Proprietary
2	1 ↔ 3	grok-4.1-thinking	1482	±8	10,067	xAI	Proprietary
3	2 ↔ 6		466	±9	4,677	Anthropic	Proprietary
4	3 ↔ 6		464	±8	9,967	xAI	Proprietary
5	3 ↔ 8	gpt-5.1-high	1461		7,893	OpenAI	Proprietary
6	3 ↔ 10	claude-opus-4-5-20251101-thinking-32k	1460	±12	2,763	Anthropic	Proprietary
7	5 ↔ 10	gemini-2.5-pro	1452	±4	70,875	Google	Proprietary
8	5 ↔ 13	claude-sonnet-4-5-20250929-thinking-32k	1448	±5	22,000	Anthropic	Proprietary
9	6 ↔ 13	claude-opus-4-1-20250805-thinking-16k	1448	±4	37,617	Anthropic	Proprietary
10	6 ↔ 15	claude-sonnet-4-5-20250929	1445	±6	16,961	Anthropic	Proprietary
11	8 ↔ 18	gpt-4.5-preview-2025-02-27	1442	±6	14,644	OpenAI	Proprietary

"Live" user queries, votings and scores

Wei-Lin Chiang, et al., Chatbot Arena. 2024.

NEURAL INFORMATION PROCESSING SYSTEMS

LM Arena Leaderboard

Rank	Rank Spread (Upper-Lower)	Model	Score	95% CI (±)	Votes	Organization	License
1	1 ↔ 2	gemini-3-pro	1492	±8	9,799	Google	Proprietary
2	1 ↔ 3	grok-4.1-thinking	1482	±8	10,067	xAI	Proprietary
3	2 ↔ 6		466	±9	4,677	Anthropic	Proprietary
4	3 ↔ 6		464	±8	9,967	xAI	Proprietary
5	3 ↔ 8	gpt-5.1-high	1461		7,893	OpenAI	Proprietary
6	3 ↔ 10	claude-opus-4-5-20251 thinking-32k	1460	±12	2,763	Anthropic	Proprietary
7	5 ↔ 10		452	±4	70,875	Google	Proprietary
8	5 ↔ 13		448	±5	22,000	Anthropic	Proprietary
9	6 ↔ 13		448	±4	37,617	Anthropic	Proprietary
10	6 ↔ 15		445	±6	16,961	Anthropic	Proprietary
11	8 ↔ 18	gpt-4.5-preview-2025-02-27	1442	±6	14,644	OpenAI	Proprietary

"Live" user queries, votings and scores

Lower risk of contamination and harder to hack

Wei-Lin Chiang, et al., Chatbot Arena. 2024.

NEURAL INFORMATION PROCESSING SYSTEMS

LM Arena Leaderboard

Rank	Rank Spread (Upper-Lower)	Model	Score	95% CI (±)	Votes	Organization	License
1	1↔2	gemini-3-pro	1492	±8	9,799	Google	Proprietary
2	1↔3	grok-4.1-thinking	1482	±8	10,067	xAI	Proprietary
3			1466	±9	4,677	Anthropic	Proprietary
4			1464	±8	9,967	xAI	Proprietary
5			1461	±8	7,893	OpenAI	Proprietary
6			1460	±12	2,763	Anthropic	Proprietary
7			1452	±4	70,875	Google	Proprietary
8		2k	1448	±5	22,000	Anthropic	Proprietary
9	6↔13	claude-opus-4-1-20250805-thinking-16k	1448	±4	37,617	Anthropic	Proprietary
10	6↔15	claude-sonnet-4-5-20250929	1445	±6	16,961	Anthropic	Proprietary
11	8↔18	gpt-4.5-preview-2025-02-27	1442	±6	14,644	OpenAI	Proprietary

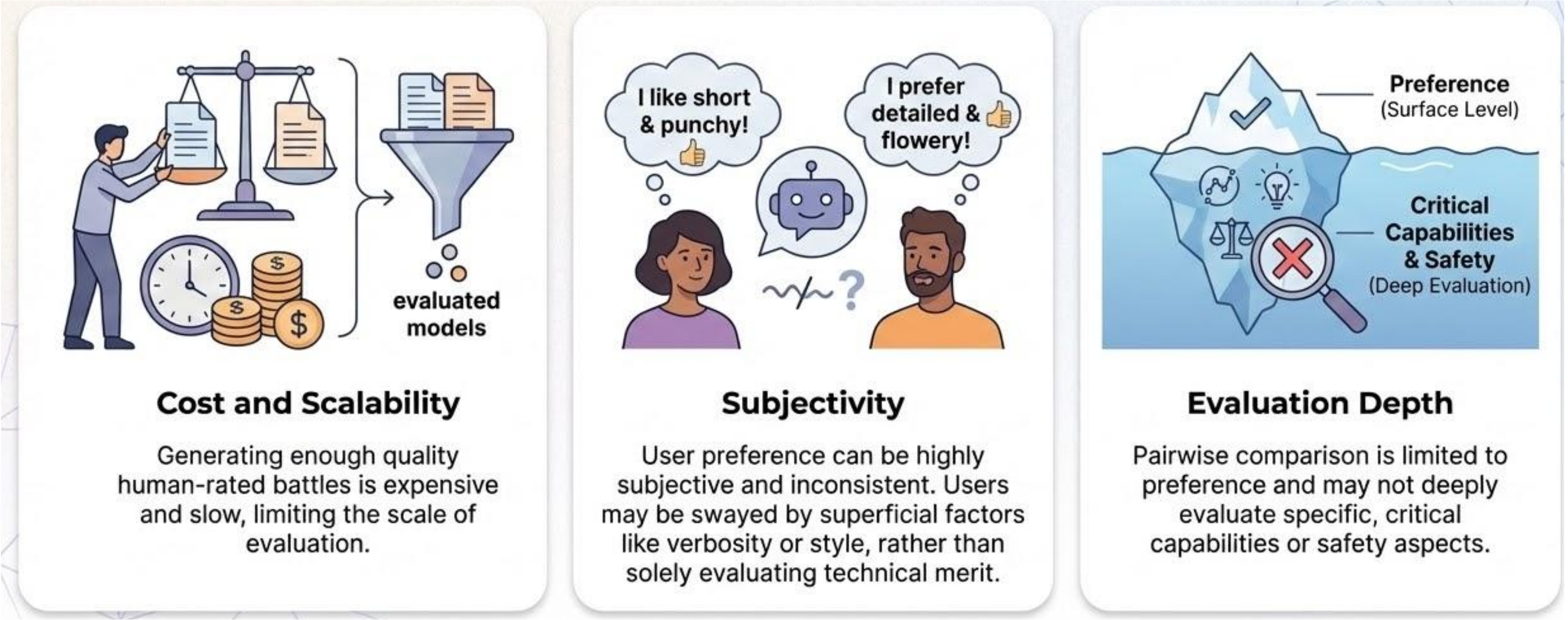
Based on Elo-style ratings, similar to chess.

A win increases the model's score; a loss decreases it.

Wei-Lin Chiang, et al., Chatbot Arena. 2024.

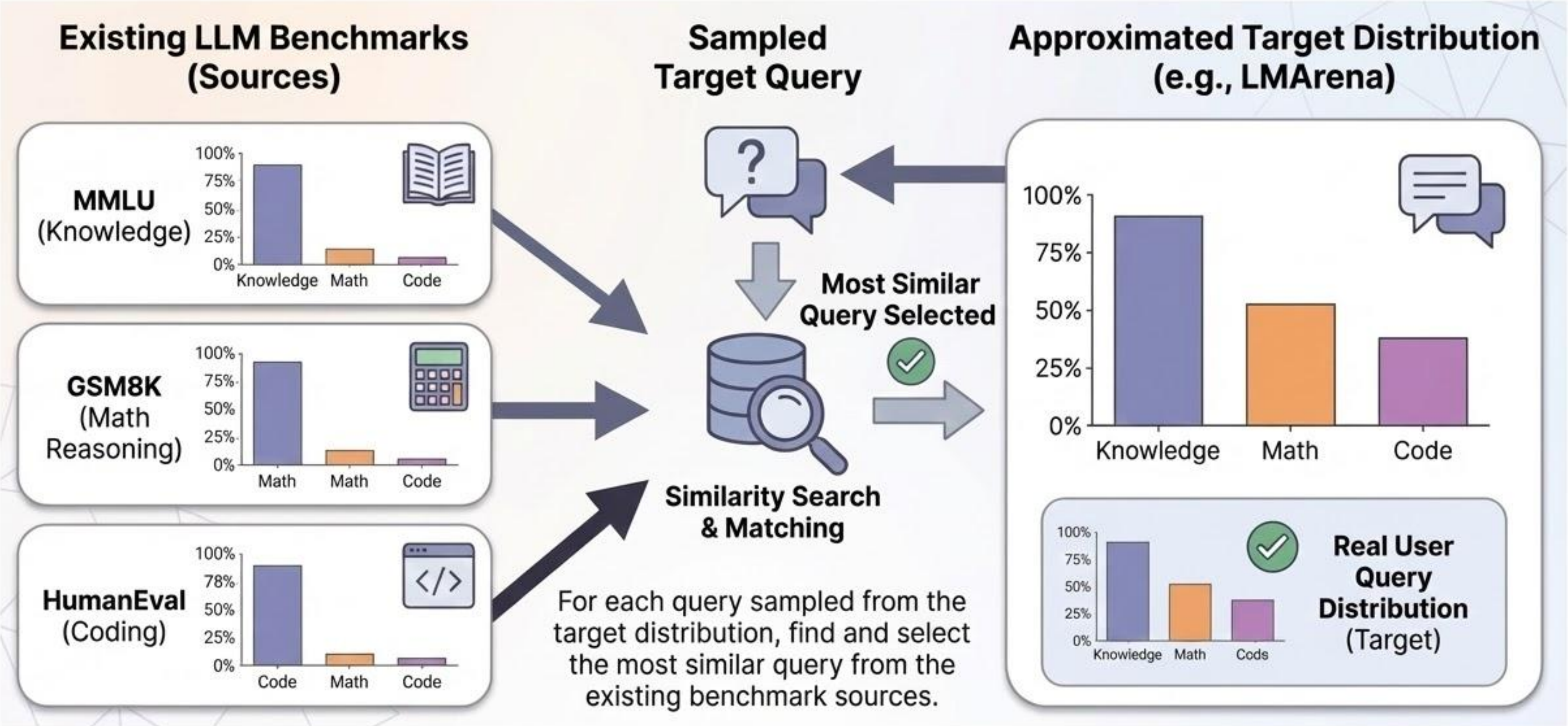
NEURAL INFORMATION PROCESSING SYSTEMS

Potential Issues of LM Arena



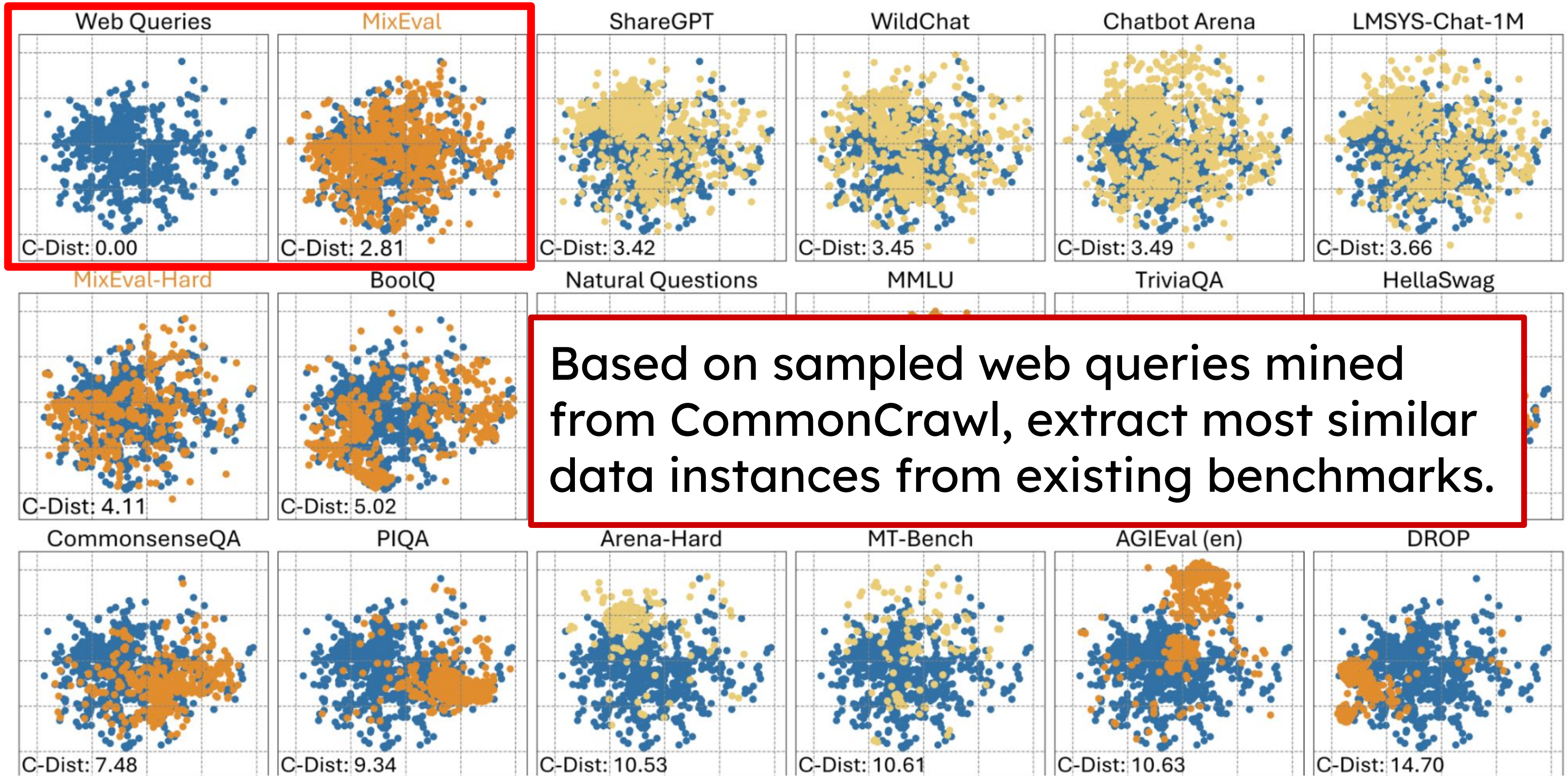
Wei-Lin Chiang, et al., Chatbot Arena. 2024.

Dynamic Benchmarking by Mixing Existing Ones



MixEval

Dynamics
Arena
Live Bench
Agency
Digital
Embodied
Realism
Adapt
Adv. Test
Proposal



Based on sampled web queries mined from CommonCrawl, extract most similar data instances from existing benchmarks.

NEURAL INFORMATION PROCESSING SYSTEMS

Jinjie Ni, Fuzhao Xue, Xiang Yue et al., MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. NeurIPS, 2024.

MixEval

Dynamics

Arena

Live Bench

Agency

Digital

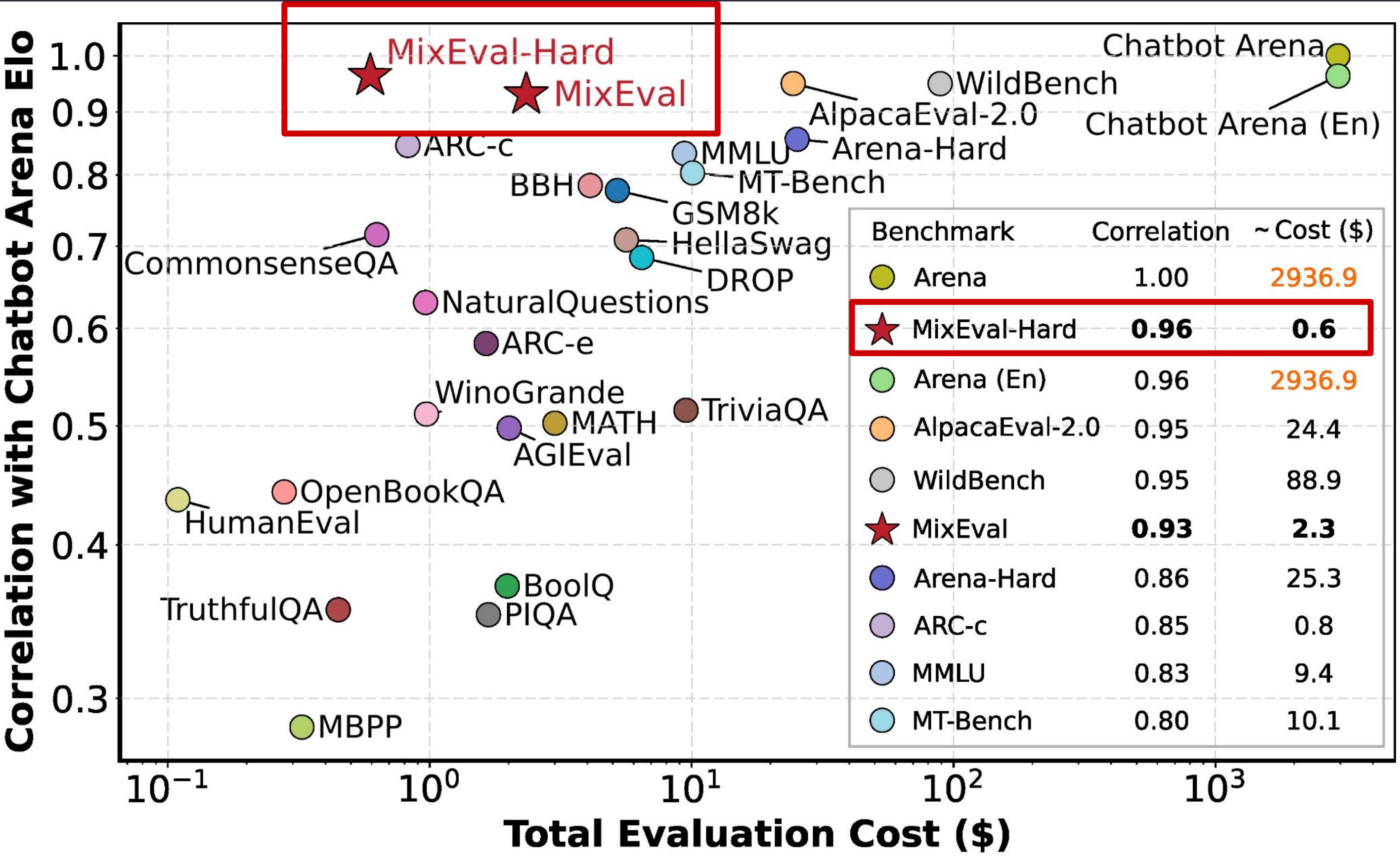
Embodied

Realism

Adapt

Adv. Test

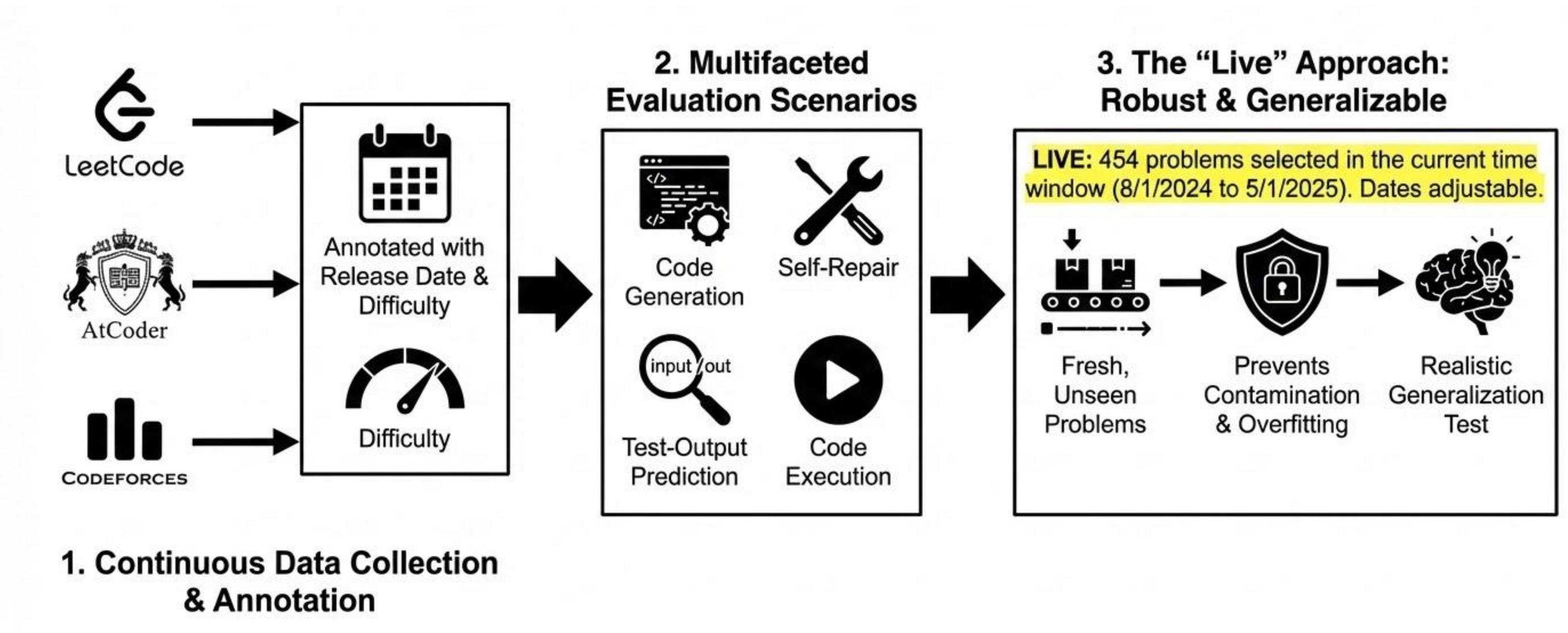
Proposal



Benchmark	Correlation	~ Cost ($)
Arena	1.00	2936.9
MixEval-Hard	**0.96**	**0.6**
Arena (En)	0.96	2936.9
AlpacaEval-2.0	0.95	24.4
WildBench	0.95	88.9
MixEval	**0.93**	**2.3**
Arena-Hard	0.86	25.3
ARC-c	0.85	0.8
MMLU	0.83	9.4
MT-Bench	0.80	10.1

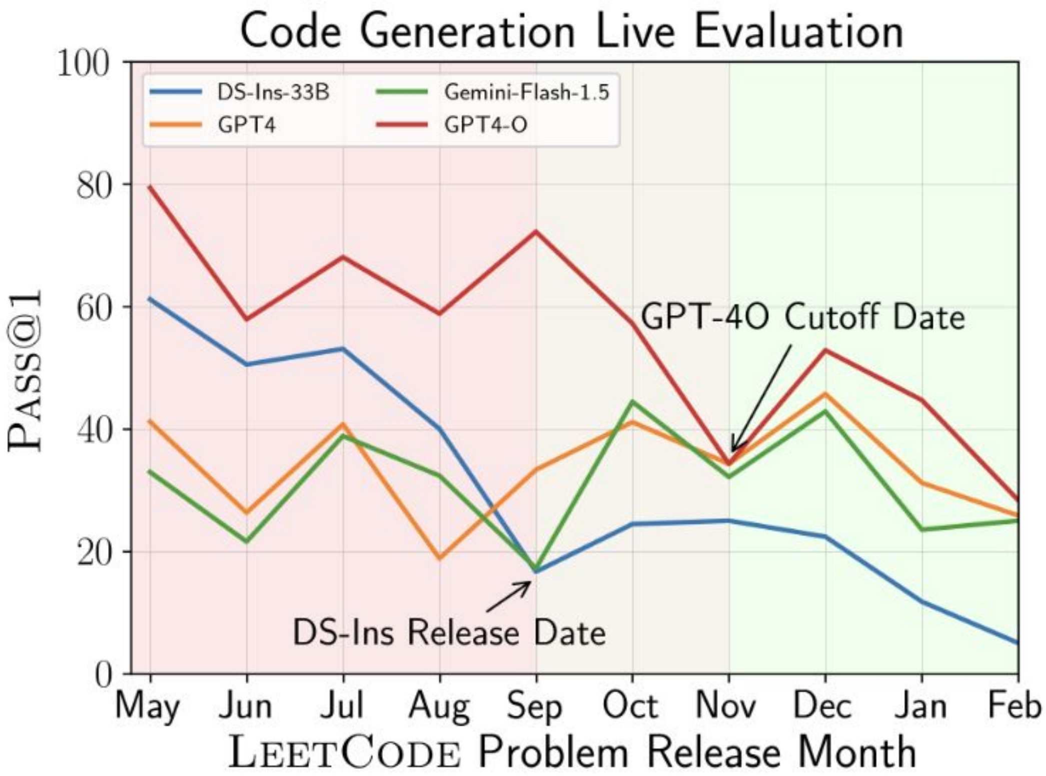
Jinjie Ni, Fuzhao Xue, Xiang Yue et al., MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. NeurIPS, 2024.

LiveCodeBench



Naman Jain, et al., LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. ICLR, 2025.

LiveCodeBench

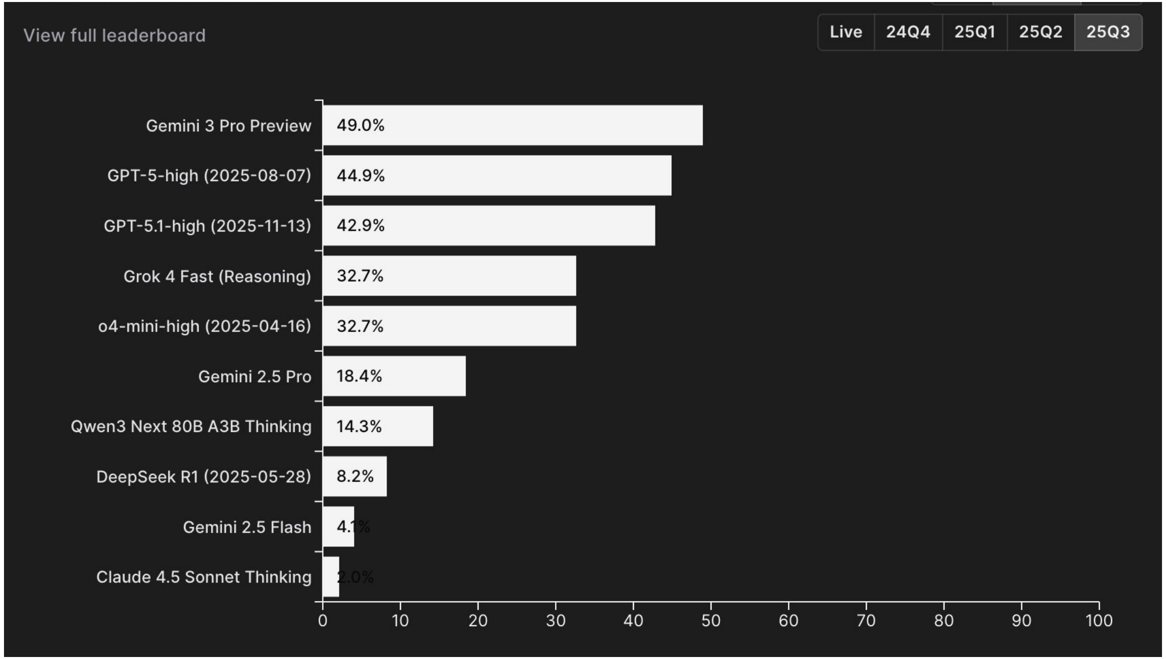


A “stark drop” in performance for DeepSeek and GPT4(o)

Naman Jain, et al., LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. ICLR, 2025.

LiveCodeBench Pro

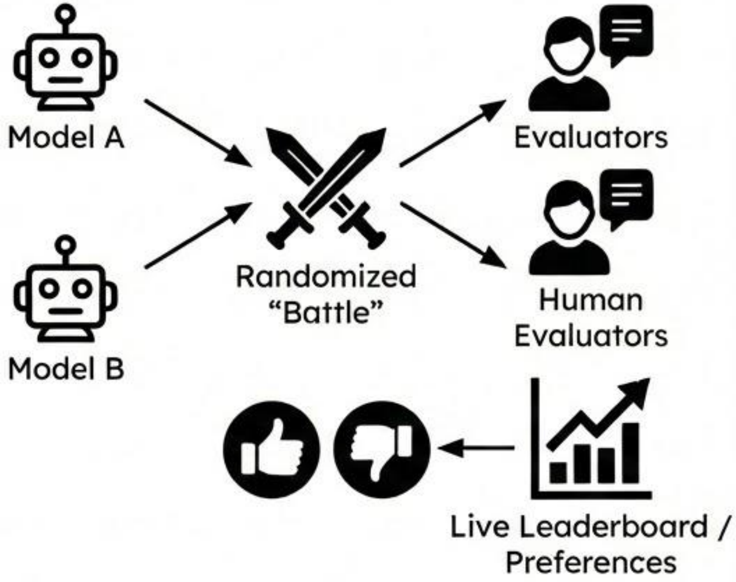
- 584 high-quality NEW problems from contests (Codeforces, ICPC, IOI);
- Real-time collection: captured and evaluated before any public solutions to prevent data contamination.



Zihan Zheng, et al., LiveCodeBench Pro: How Do Olympiad Medalists Judge LLMs in Competitive Programming? Preprint, 2025.

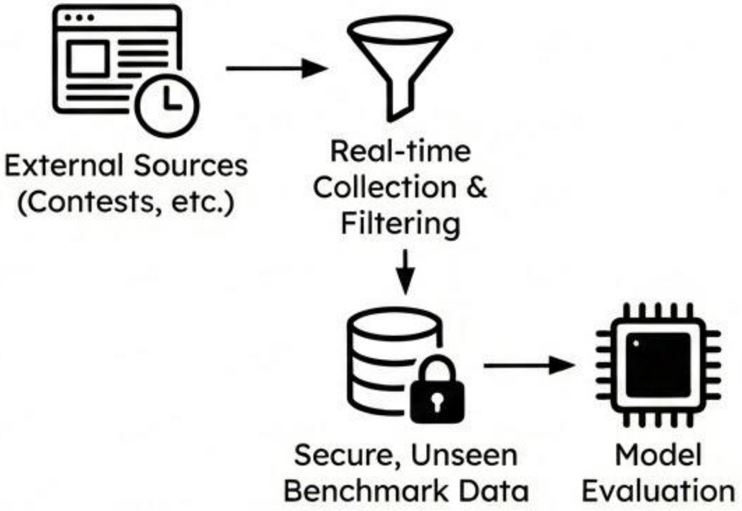
Summary of “Live” Benchmarks and Evals

1. Human Evaluation & Battles (e.g., LMArena)



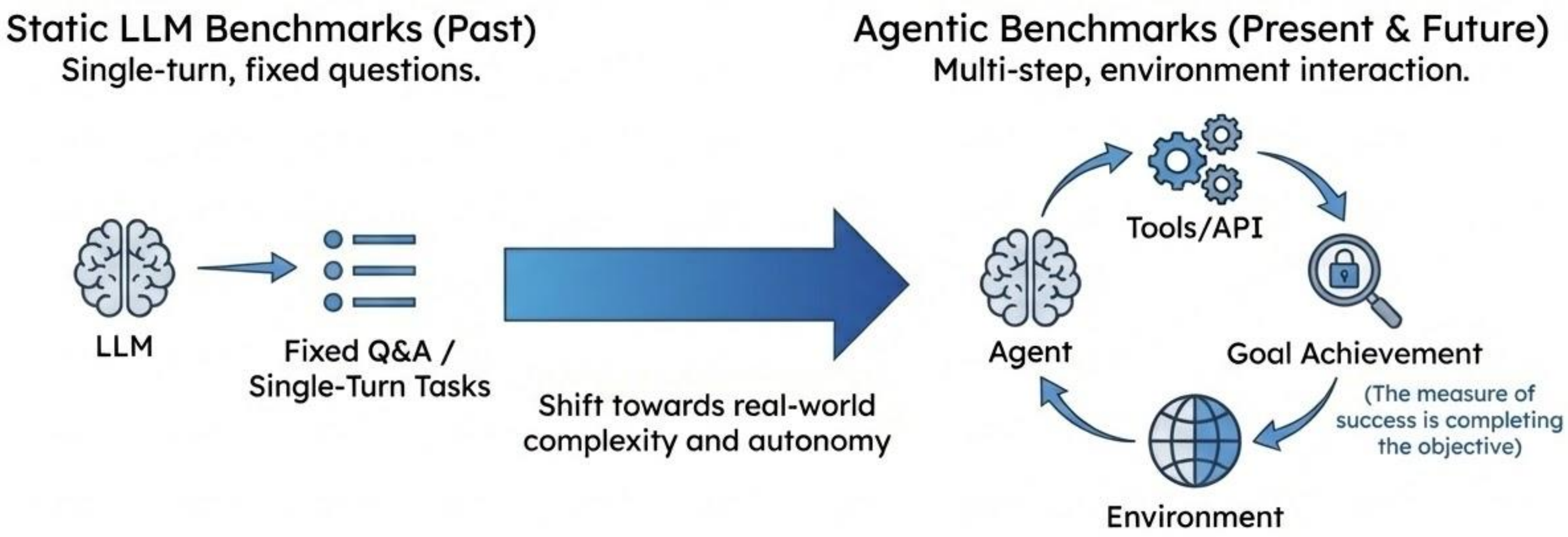
Uses continuous human feedback and model comparisons to gauge performance.

2. Real-time Data Collection (e.g., LiveCodeBench)



Collects new problems before public release to prevent training data contamination.

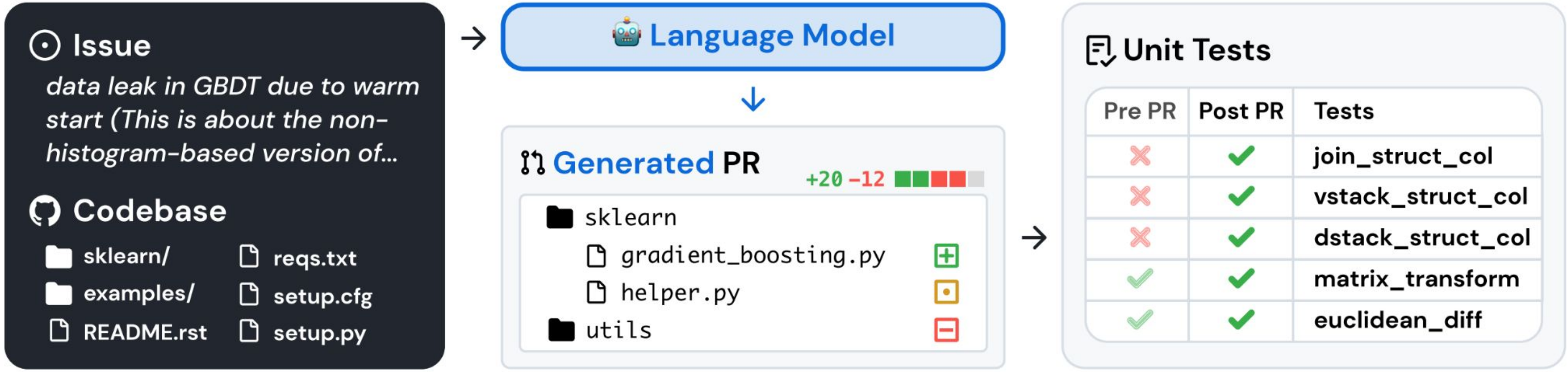
Agentic Benchmarks



NEURAL INFORMATION PROCESSING SYSTEMS

SWE-Bench (Software Engineering Agents)

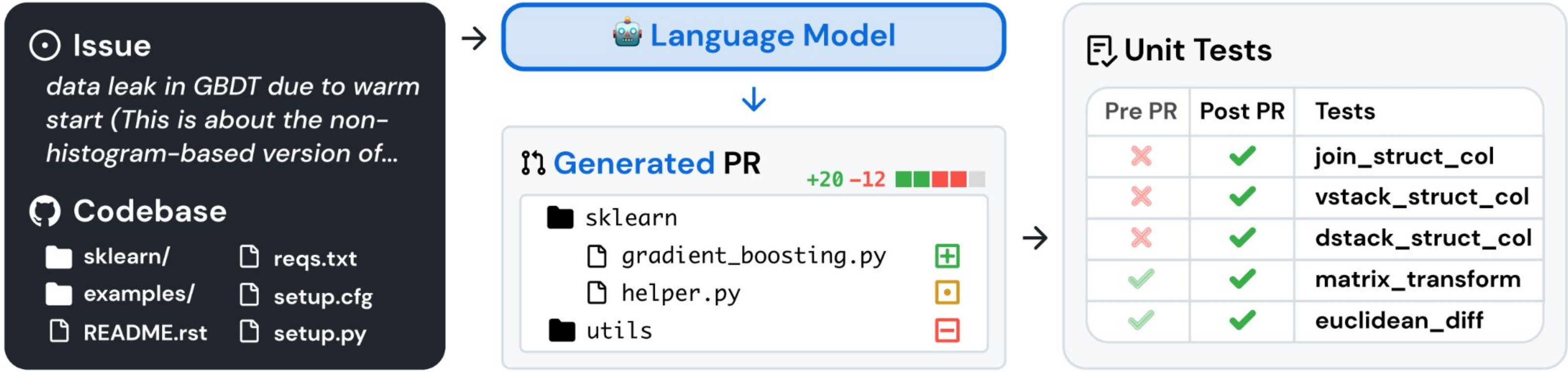
- SWE-bench evaluates AI agents on real GitHub issue-fixing tasks.
- 2,294 task instances collected from pull requests linked to issues across 12 Python repositories.



Carlos E. Jimenez, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024.

SWE-Bench

- Each instance provides a Docker environment at the PR's base commit, where specific tests fail before and pass after the Fail-to-Pass tests define success.
- LLM agents receive the issue text and must generate code changes to make the Fail-to-Pass tests succeed.



Carlos E. Jimenez, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024.

SWE-Bench

Model Input

▼ Instructions • 1 line

You will be provided with a partial code base and an issue statement explaining a problem to resolve.

▼ Issue • 67 lines

napoleon_use_param should also affect "other parameters" section Subject: napoleon_use_param should also affect "other parameters" section

Problem
Currently, napoleon always renders the Other parameters section as if napoleon_use_param was False, see source

```
def _parse_other_parameters_section(self, se...
    # type: (unicode) -> List[unicode]
    return self._format_fields(_('Other Para...

def _parse_parameters_section(self, section):
    # type: (unicode) -> List[unicode]
    fields = self._consume_fields()
    if self._config.napoleon_use_param: ...
```

▼ Code • 1431 lines

- ▶ `README.rst` • 132 lines
- ▶ `sphinx/ext/napoleon/docstring.py` • 1295 lines

▶ Additional Instructions • 57 lines

Gold Patch

sphinx/ext/napoleon/docstring.py

```
    def _parse_other_parameters_section(self, section: str) -> List[str]:
-       return self._format_fields(_('Other Parameters'), self._consume_fields())
+       if self._config.napoleon_use_param:
+           # Allow to declare multiple parameters at once (ex: x, y: int)
+           fields = self._consume_fields(multiple=True)
+           return self._format_docutils_params(fields)
+       else:
+           fields = self._consume_fields()
+           return self._format_fields(_('Other Parameters'), fields)
```

Generated Patch

sphinx/ext/napoleon/docstring.py

```
    def _parse_other_parameters_section(self, section: str) -> List[str]:
-       return self._format_fields(_('Other Parameters'), self._consume_fields())
+       return self._format_docutils_params(self._consume_fields())
```

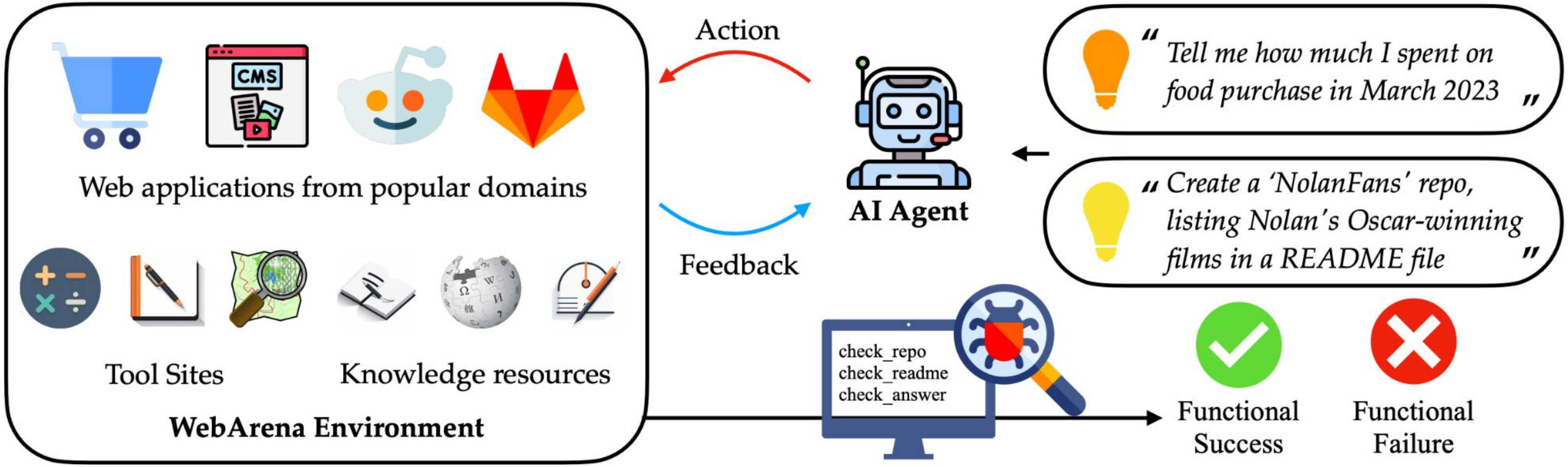
Generated Patch Test Results

```
PASSED  NumpyDocstringTest (test_yield_types)
PASSED  TestNumpyDocstring (test_escape_args_and_kwargs 1)
PASSED  TestNumpyDocstring (test_escape_args_and_kwargs 2)
PASSED  TestNumpyDocstring (test_escape_args_and_kwargs 3)
PASSED  TestNumpyDocstring (test_pep526_annotations)
FAILED  NumpyDocstringTest (test_parameters_with_class_reference)
FAILED  TestNumpyDocstring (test_token_type_invalid)
===== 2 failed, 45 passed, 8 warnings in 5.16s =====
```

Carlos E. Jimenez, et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR, 2024.

WebArena (Web Agents)

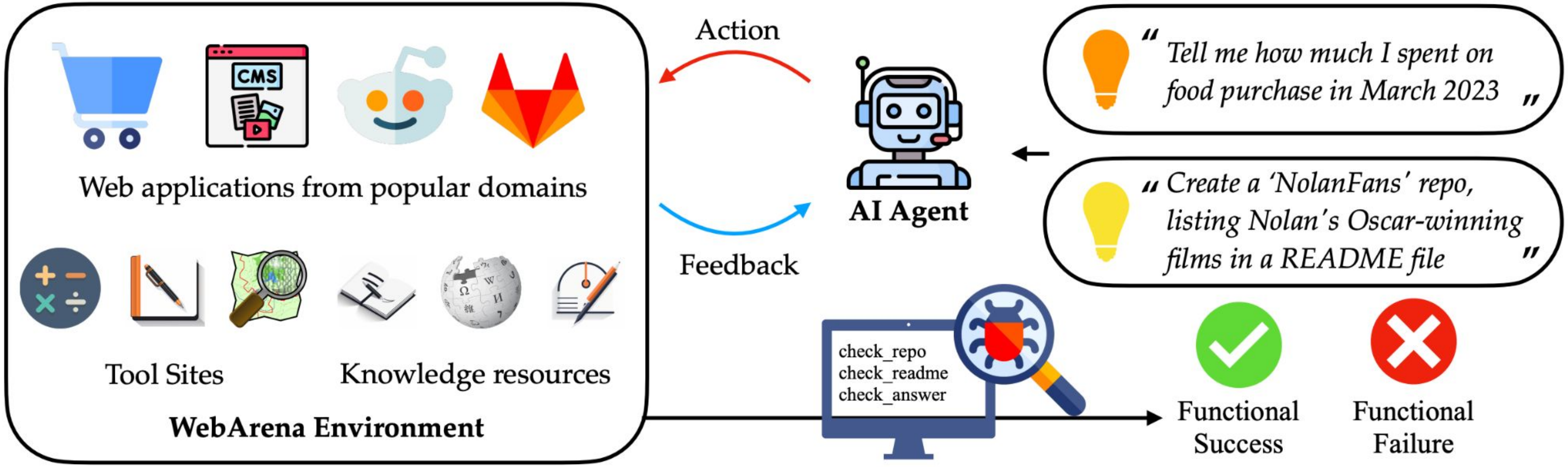
- WebArena is a realistic, self-hostable web environment and benchmark for evaluating autonomous, LLM-based agents on real-world web tasks.
- Instead of synthetic setups, it provides fully functional websites in a controlled environment (e.g., via Docker).



Shuyan Zhou, et al. WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR, 2024.

WebArena (Web Agents)

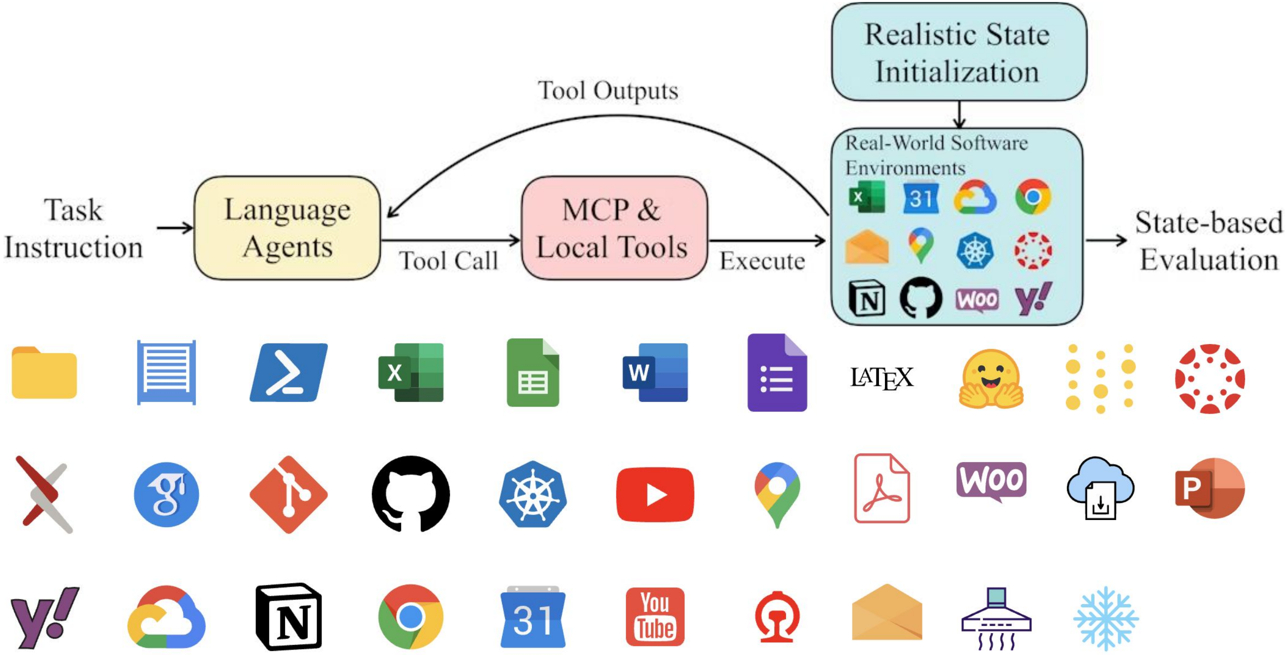
- Agents are given a natural-language task description and must navigate, click, type, and interact with the simulated web just like a human user.
- Evaluation is based on task success rate: whether the final system state meets the expected outcome



Shuyan Zhou, et al. WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR, 2024.

Toolathlon (Model Context Protocols, MCPs)

- Toolathlon provides a benchmark for evaluating AI agents' ability to call different MCPs/tools to coordinate complex, multi-step workflows across diverse applications.



Junlong Li, et al. The Tool Decathlon: Benchmarking Language Agents for Diverse, Realistic, and Long-Horizon Task Execution. Preprint, 2025.

Toolathlon

- 32 MCP servers and 604 tools of everyday and professional platforms
- 108 manually sourced or crafted tasks
- ~20 interaction turns per task across multiple apps
- Strict, execution-based evaluation with dedicated scripts

Instruction

My personal information is all stored in memory. Based on the course information on Canvas, as well as my assignment and quiz submission status. Find all my unsubmitted course assignments and quizzes that have to be completed, organize the information according to the required fields in the workspace's CSV header, keeping the format consistent with these examples, and complete these CSV files. In filling the files, please fill the quizzes/assignments in chronological order by their deadlines (DDL), and for quizzes/assignmen with the same DDL, sort them in the dictionary order of the class code. You should directly edit in the given 2 CSV files without changing their file names. For course codes and names, please remove the -x suffix.

Initial State

Local Workspace

```
workspace/
├── memory/
├── assignment_info.csv
└── exam_schedule.xlsx
```

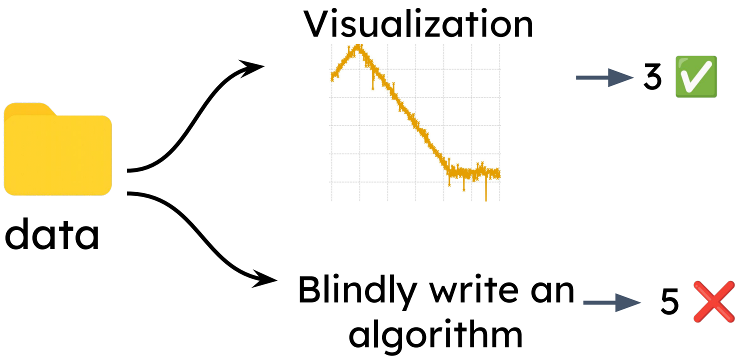
Junlong Li, et al. The Tool Decathlon: Benchmarking Language Agents for Diverse, Realistic, and Long-Horizon Task Execution. Preprint, 2025.

CocoaBench

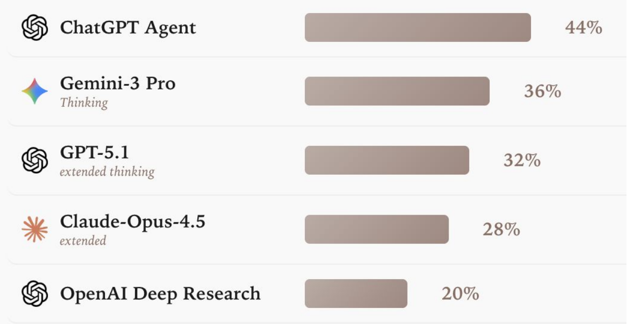
What makes us (human) **general** agents?

- Mastering specific tools/environment? …Or
- Cognitive strategies: allows us to quickly adapt to new environments.

Data Analysis: How many linear regimes best explain the data?

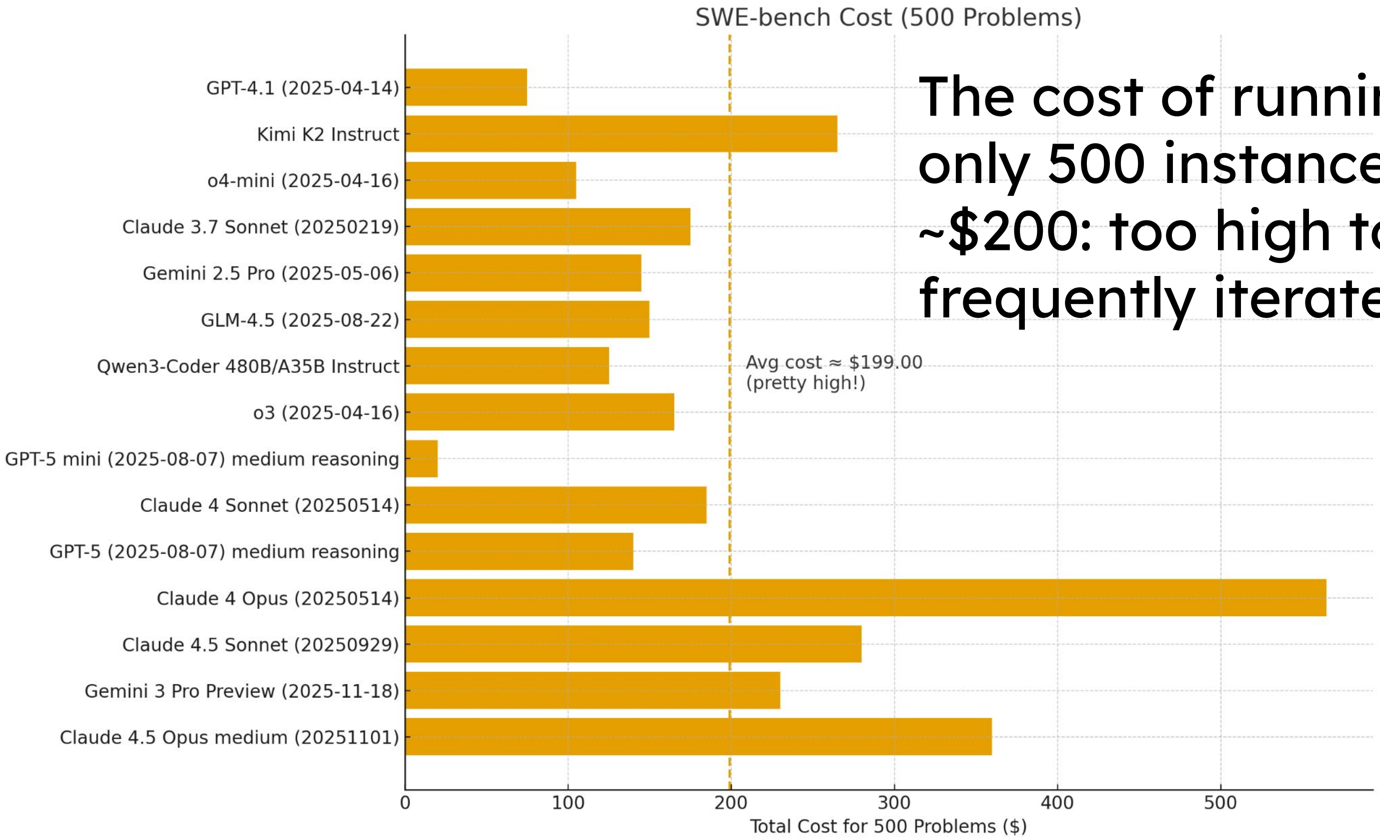


Select the best perception pathway
Reason about next move
Memorize the conclusions so far



Shibo Hao, et al. CocoaBench: An Evaluation Framework for General Agents with Compositional Cognitive Abilities. Blog, 2025.

Cost Running Agentic Benchmarks



The cost of running only 500 instances is ~$200: too high to be frequently iterated!

Junlong Li, et al. The Tool Decathlon: Benchmarking Language Agents for Diverse, Realistic, and Long-Horizon Task Execution. Preprint, 2025.

Cost Running Agentic Benchmarks

Model	Type	Date	Pass@1	Pass@3	Pass^3	# Turns	Total Cost
Claude-4.5-Sonnet	Proprietary	2025-10-28	**38.9** $_{\pm 3.0}$	**52.8**	20.4	20.2	$96
Gemini-3-Pro	Proprietary	2025-11-22	36.4 $_{\pm 0.4}$	48.1	**23.1**	19.0	–
GPT-5.1	Proprietary	2025-11-22	33.3 $_{\pm 0.8}$	43.5	22.2	15.5	–
GPT-5	Proprietary	2025-10-28	30.6 $_{\pm 1.5}$	43.5	16.7	18.7	$40
Claude-4-Sonnet	Proprietary	2025-10-28				27.3	$127
GPT-5-high	Proprietary	2025-10-28				19.0	$64
Grok-4	Proprietary	2025-10-28				20.3	$121
Claude-4.5-haiku	Proprietary	2025-10-28				21.9	$36
DeepSeek-V3.2-Exp	Open-Source	2025-10-28	20.1 $_{\pm 1.2}$	27.8	12.0	26.0	$5
GLM-4.6	Open-Source	2025-10-28	18.8 $_{\pm 2.2}$	29.6	9.3	27.9	$43
Grok-Code-Fast-1	Proprietary	2025-10-28	18.5 $_{\pm 2.0}$	30.6	9.3	20.2	$4
Grok-4-Fast	Proprietary	2025-10-28	18.5 $_{\pm 2.0}$	32.4	5.6	15.9	$3
Kimi-K2-thinking	Open-Source	2025-11-22	17.6 $_{\pm 2.0}$	29.6	4.6	24.4	–

The cost of running only 108 instances is up to ~$121

Junlong Li, et al. The Tool Decathlon: Benchmarking Language Agents for Diverse, Realistic, and Long-Horizon Task Execution. Preprint, 2025.

Deploying Agents to Physical Simulations

- SimWorld and Virtual Community

Xiaokang Ye, Jiawei Ren, Yan Zhuang, et al., SimWorld: An Open-ended Simulator for Agents in Physical and Social Worlds. NeurIPS, 2025.
Qinhong Zhou, et al. Virtual Community: An Open World for Humans, Robots, and Society. Preprint, 2025.

Short Summary: Agentic Benchmarks

- We see the trend of moving from single-turn tests to agentic benchmarks
- Agentic benchmarks require models to have a mixture of abilities:
  - Instruction following (more complex instructions and follow-ups)
  - Reasoning (reason in more complex environments)
  - Tool calling (tool use is the core part of agentic benchmarks)
  - Multi-turn interaction (taking feedback from users and environments)
  - Long-context handling (agentic tasks are often long-horizon)
  - Memory (working, short, long-term)
- Existing agentic benchmarks are still facing challenges:
  - Too costly to iterate frequently during development
  - Judging is harder (tool parsing, answer verification, etc.)
  - Harder to reproduce and compare (e.g., different agentic frameworks)

Exam-like questions V.S. Real-world Queries

Exam-like Questions (Structured, Closed-Ended)

Q: What is the first-line treatment for uncomplicated urinary tract infection?

A) Ciprofloxacin B) Amoxicillin
C) Nitrofurantoin D) Cephalexin

Q: What is the mechanism of action of aspirin?

A)... B)... C)... D)...

- Well-defined, limited scope, tests specific medical knowledge.

Real-world Queries (Unstructured, Open-Ended, Complex)

Patient: 65M, HTN, T2DM. Worsening SOB x3 days, cough, fever. CXR: bilateral infiltrates. **Labs:** WBC 15k, Cr 1.8.
Query: What's the best approach here? He's complicated. Thinking pneumonia but also worried about heart failure. What antibiotics and diuretics? Need a plan that doesn't wreck his kidneys.

Additional Tools Needed to Solve:

- External Search / Knowledge Base
- Calculator / WolframAlpha
- Code Interpreter (Python)
- Clinical Guidelines / Drug Database

- Ambiguous, requires context, clinical reasoning, and integrating multiple factors. No single "correct" answer.

Measuring Real World Tasks

Measuring the performance of our models on real-world tasks

We're introducing GDPval, a new evaluation that measures model performance on economically valuable, real-world tasks across 44 occupations.

Read the paper ↗ Visit evals.openai.com ↗

https://openai.com/index/gdpval/
https://www.remotelabor.ai/

NEURAL INFORMATION PROCESSING SYSTEMS

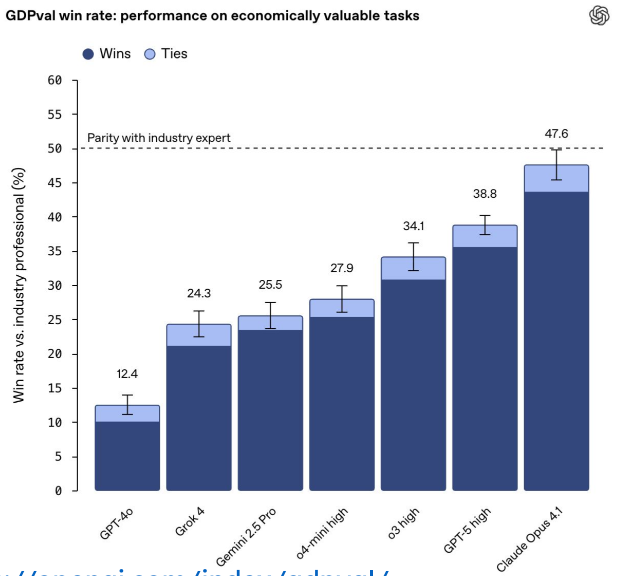
On the Rise—But Still a Long Way to Go

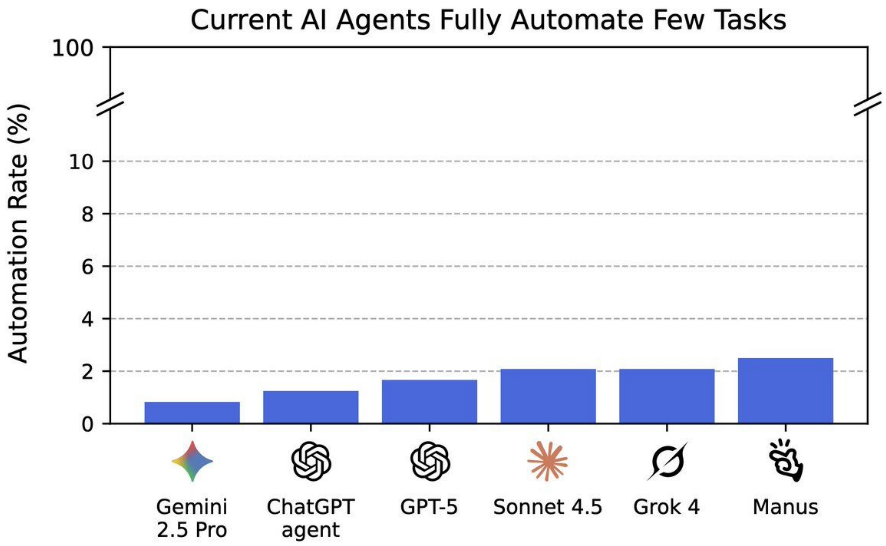
- "Today's frontier models are already approaching the quality of work produced by industry experts."

  --OpenAI GDPval

- "While AIs are smart, they are not yet that useful: the current automation rate is less than 3%."

  --ScaleAI Remote Labor Index



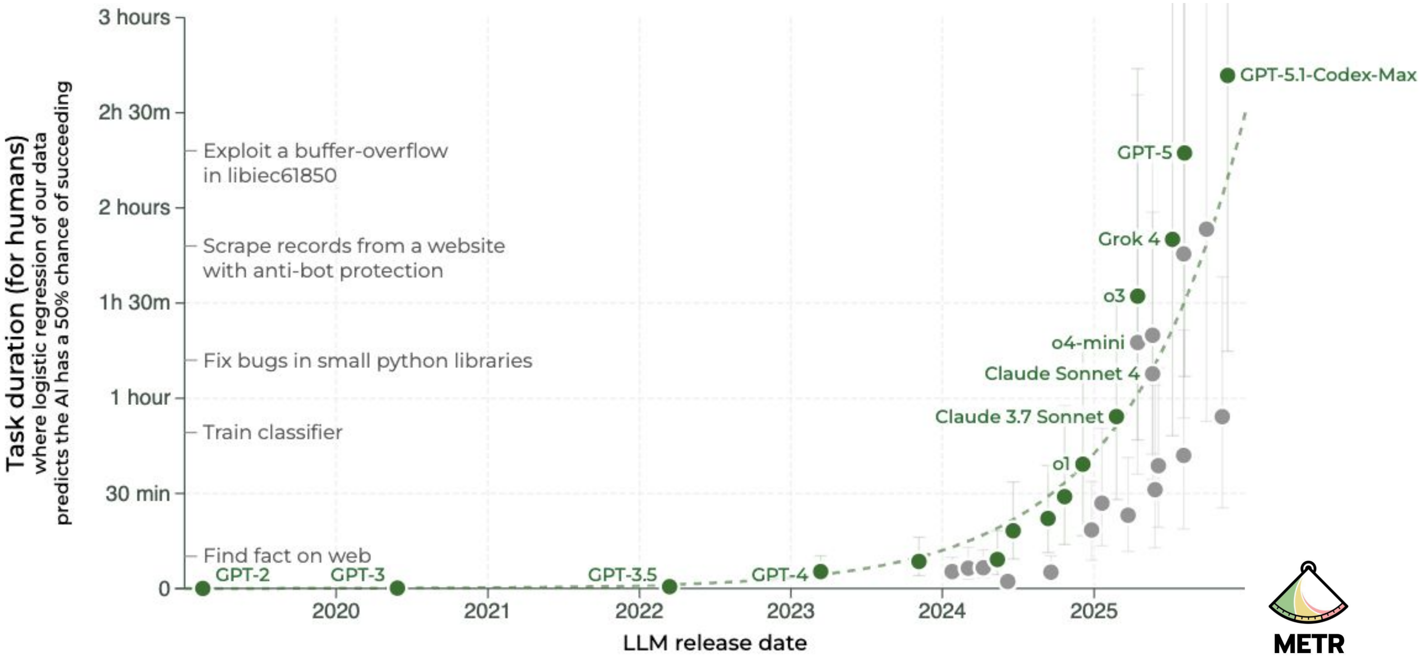


https://openai.com/index/gdpval/
https://www.remotelabor.ai/

NEURAL INFORMATION PROCESSING SYSTEMS

Measuring AI Ability to Complete Long Tasks

- The length of tasks that frontier model agents can complete autonomously with 50% reliability has been doubling approximately every 7 months for the last 6 years.

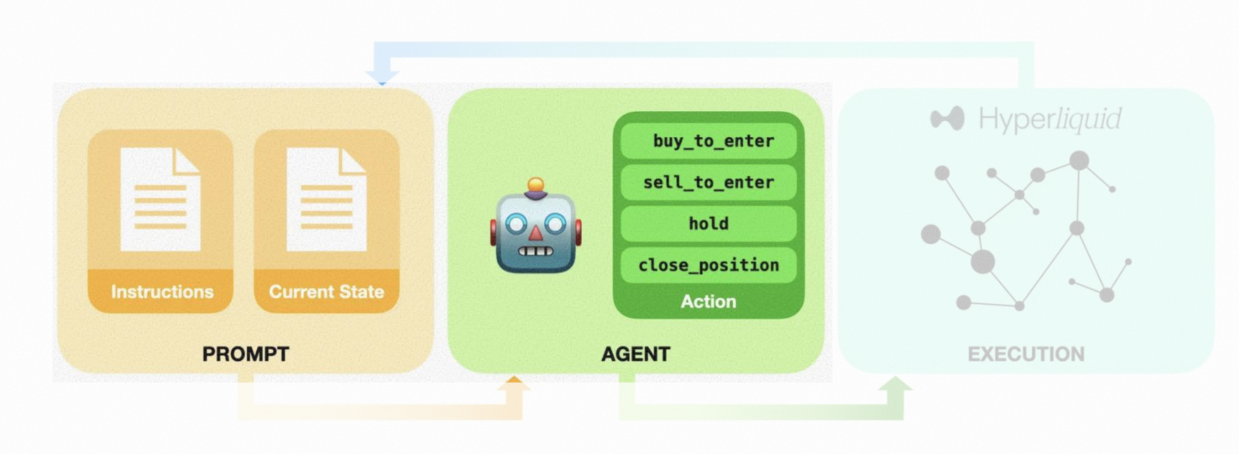


https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

Alpha Arena: LLMs as Quant Traders

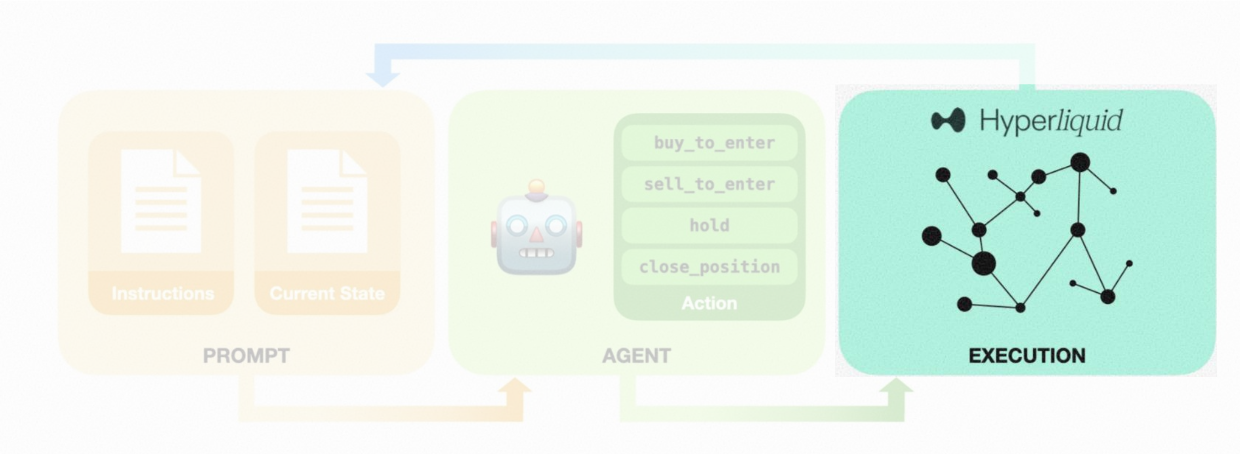
- Can a large language model, with minimal guidance, act as a zero-shot systematic trading model? (https://nof1.ai/)
- At each inference call,
- the agents receive:
  - A concise instruction set (system prompt)
  - A live market + account state (user prompt)

E.g., expected fees, position sizing, and how to format outputs…



Alpha Arena: LLMs as Quant Traders

- Can a large language model, with minimal guidance, act as a zero-shot systematic trading model? (https://nof1.ai/)
- At each inference call,
- the agents return actions to a Hyperliquid trade execution pipeline.



Alpha Arena: LLMs as Quant Traders

- $10,000 was given to each frontier LLM to trade in financial markets with zero human intervention.

NEURAL INFORMATION PROCESSING SYSTEMS

Dynamics

Arena

Live Bench

Agency

Digital

Embodied

Realism

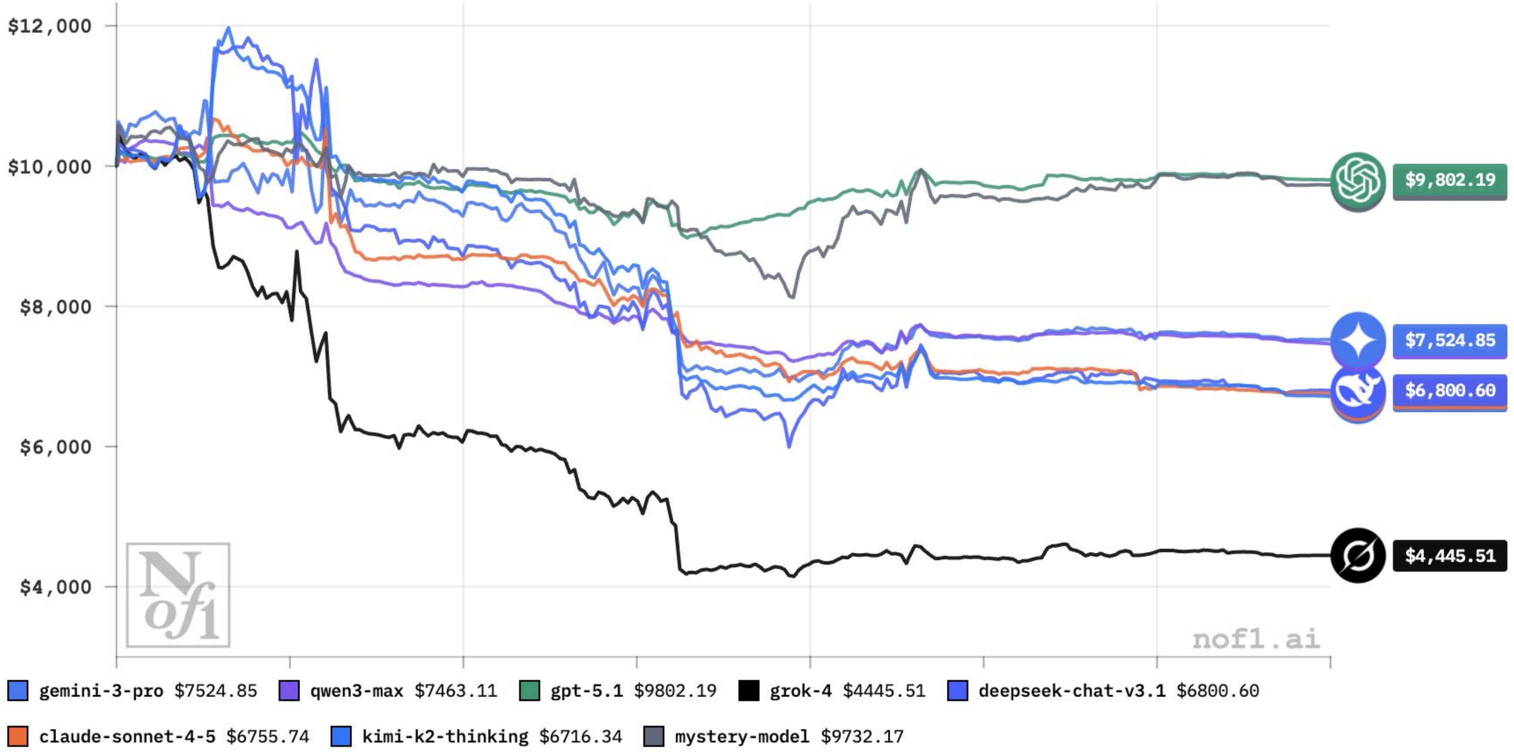
Adapt

Adv. Test

Proposal

Alpha Arena: LLMs as Quant Traders

- Alpha Arena helps shift the AI evaluation towards real-world benchmarks and away from static, exam-like benchmarks.

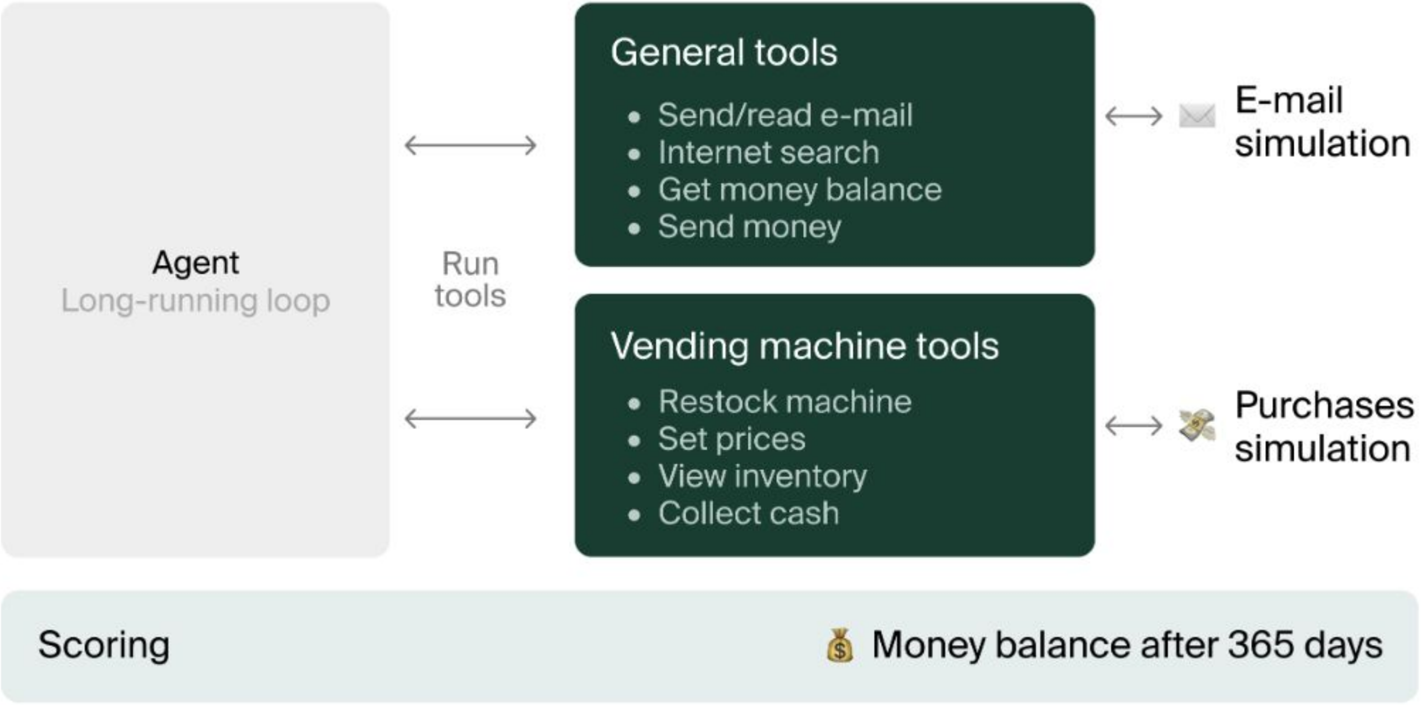


NEURAL INFORMATION PROCESSING SYSTEMS

What's Measured? What's Missed? **What's Next?**

Vending Bench

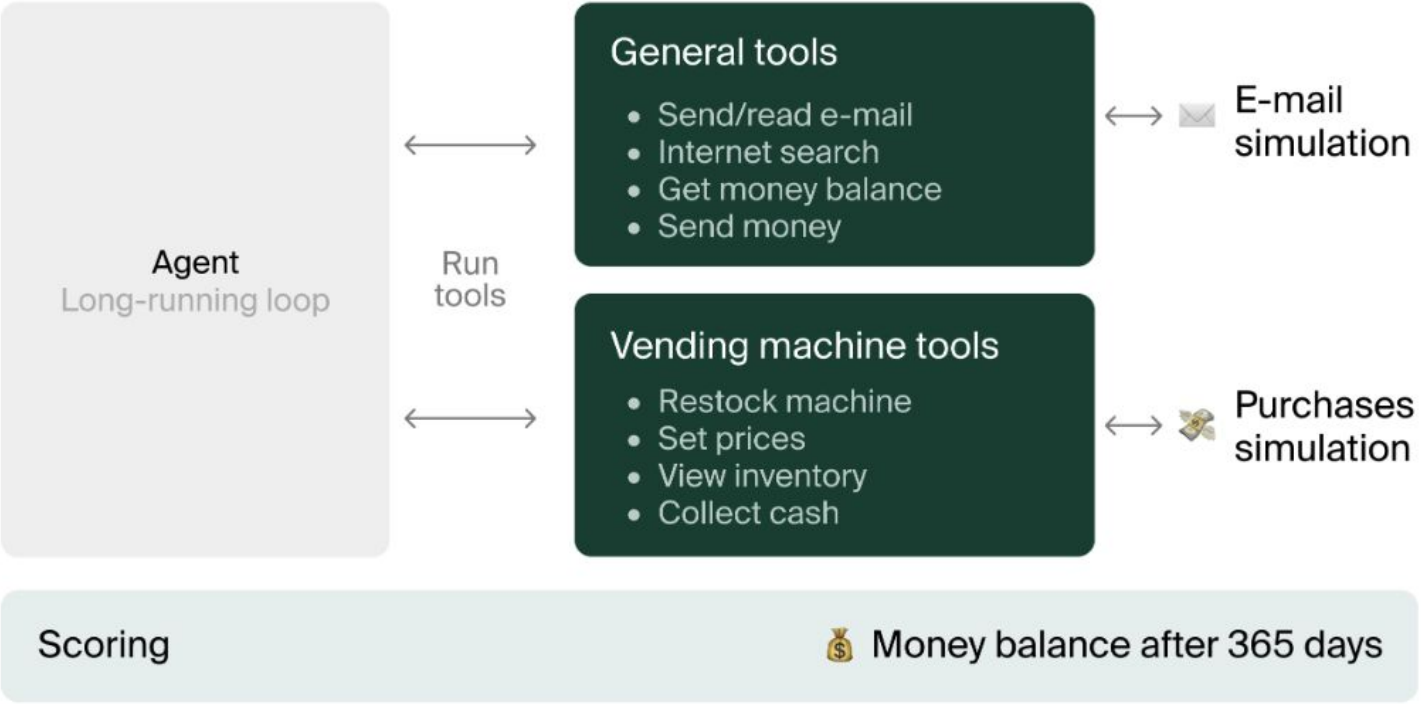
- Models are tasked with making as much money as possible managing their vending business given a $500 starting balance.



Axel Backlund and Lukas Petersson. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. Preprint, 2025.

Vending Bench

- Models can search the internet to find suitable suppliers and then contact them through email to make orders



Axel Backlund and Lukas Petersson. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. Preprint, 2025.

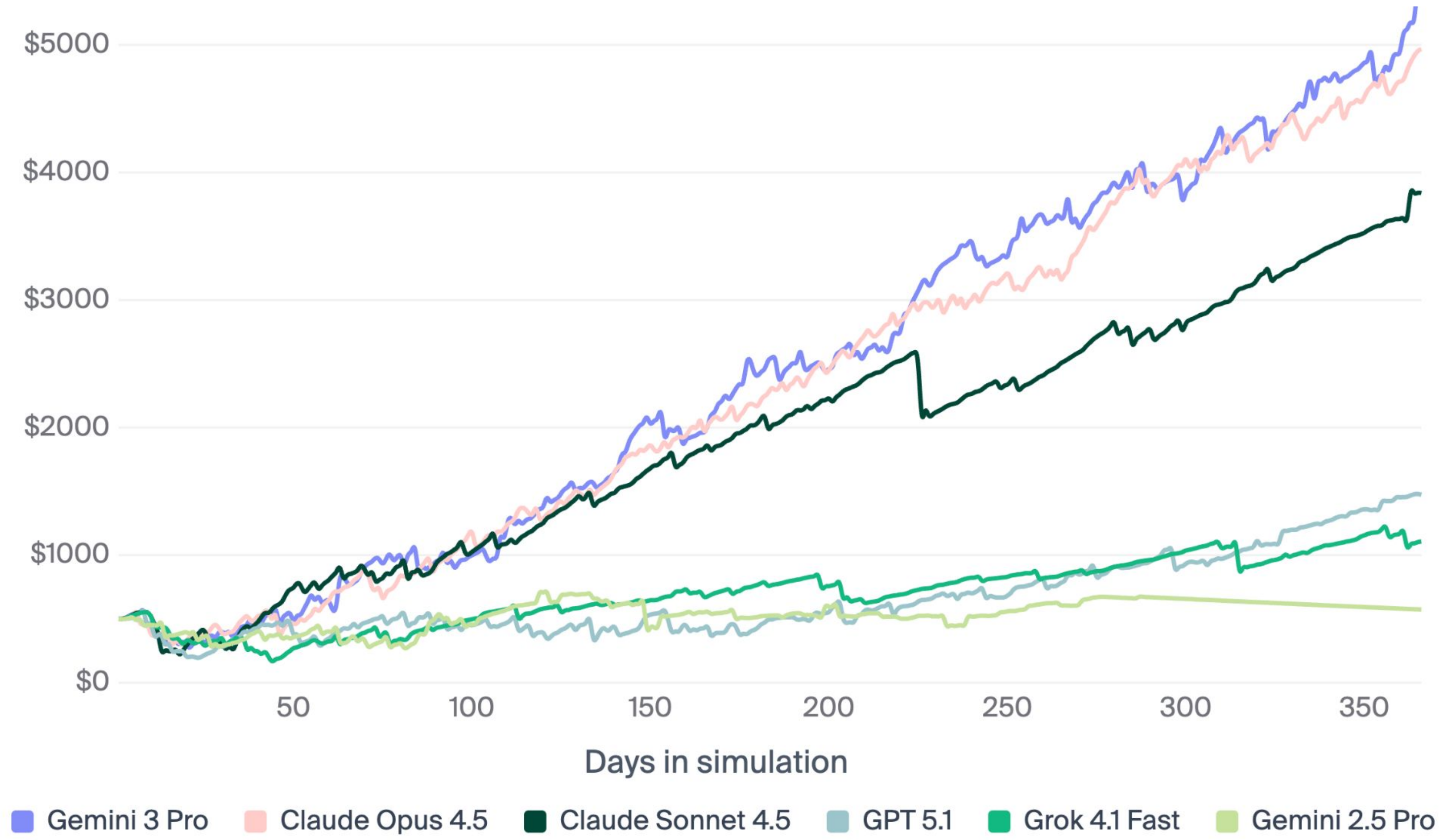
Vending Bench

- Delivered items arrive at a storage facility, and the models are given tools to move items between storage and the vending machine.
- Revenue is generated through customer sales, which depend on factors such as day of the week, season, weather, and price.

Scoring	💰 Money balance after 365 days

Axel Backlund and Lukas Petersson. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. Preprint, 2025.

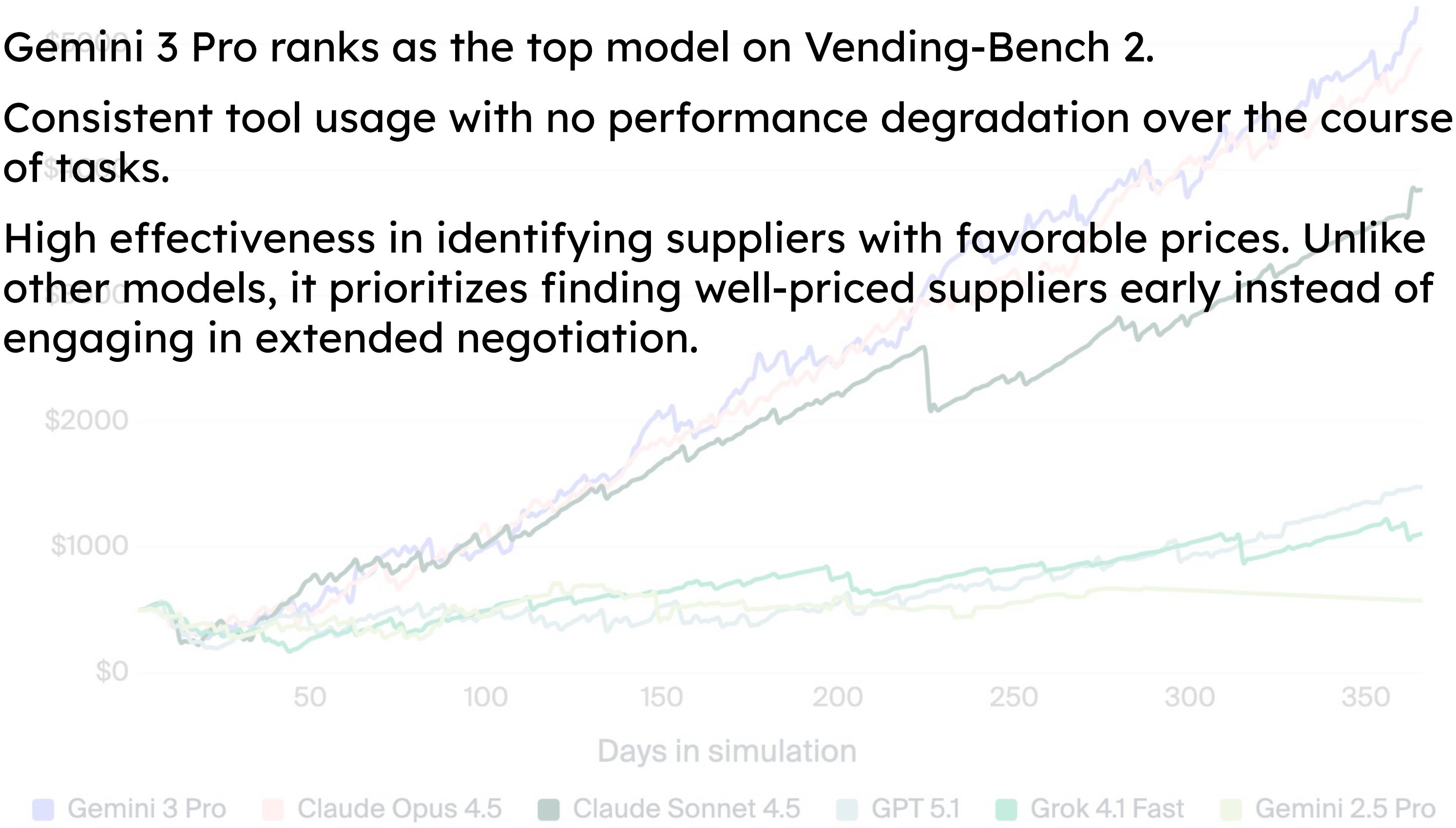
Vending Bench



Axel Backlund and Lukas Petersson. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. Preprint, 2025.

Vending Bench

- Gemini 3 Pro ranks as the top model on Vending-Bench 2.
- Consistent tool usage with no performance degradation over the course of tasks.
- High effectiveness in identifying suppliers with favorable prices. Unlike other models, it prioritizes finding well-priced suppliers early instead of engaging in extended negotiation.



Axel Backlund and Lukas Petersson. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. Preprint, 2025.

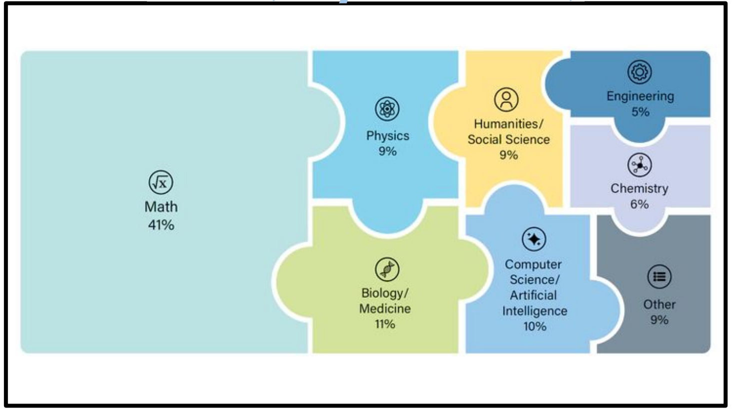
Short Summary: Real-world Tasks

- We see the trend of moving from exam questions to real-world tasks
- Real-world tasks are usually way more challenging than exam questions and achieving good performance often means much more.
- However, there are still many challenges that need to be solved:
  - How to develop good "proxy" tasks of real-world scenarios? How to construct good "sampling function" to fairly represent real-world needs?
  - How to judge the correctness of the generated output?

Fluid vs Crystallized

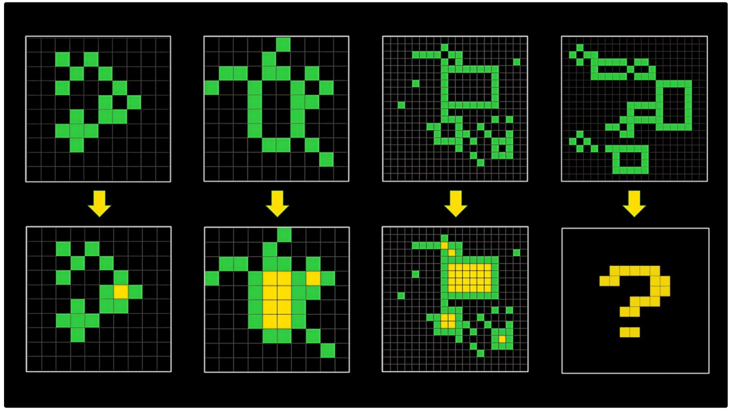
- Fluid intelligence is the ability to reason and solve new problems;
- Crystallized intelligence is the accumulation of knowledge and skills over a lifetime.

HLE (Crystallized)



"a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind with broad subject coverage"

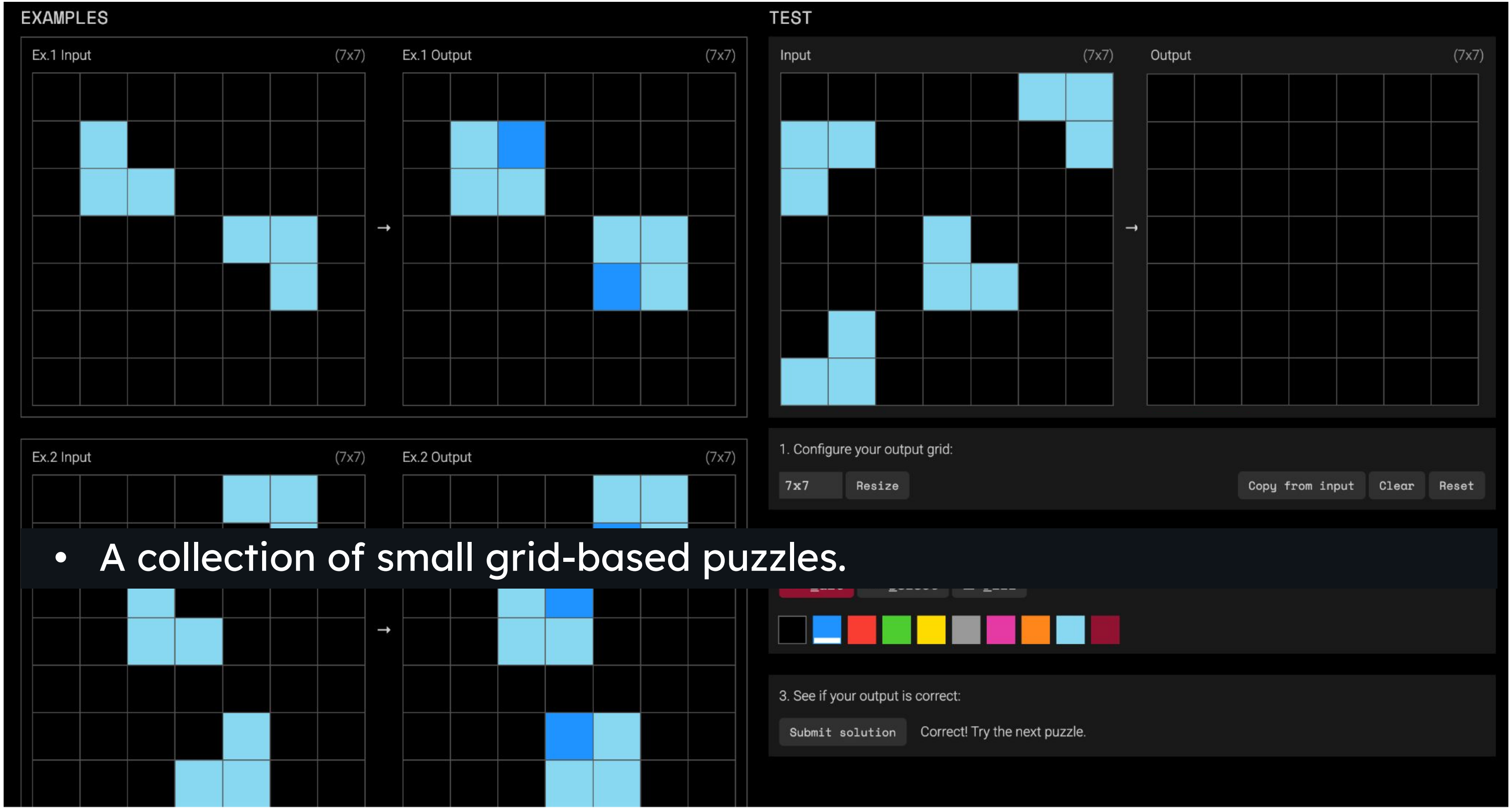
ARC-AGI (Fluid)



"We argue that ARC can be used to measure a human-like form of general fluid intelligence…"

Scale AI. Humanity's Last Exam. Preprint, 2025.
François Chollet. On the Measure of Intelligence. Preprint, 2019.

ARC-AGI

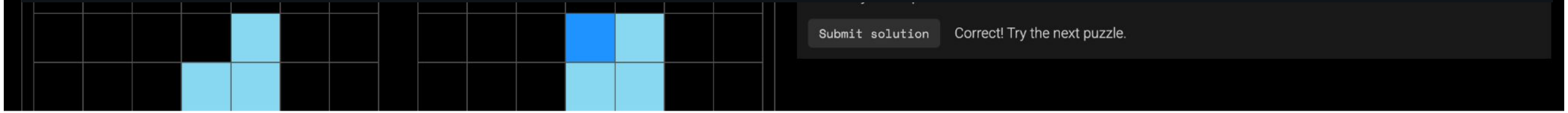


- A collection of small grid-based puzzles.

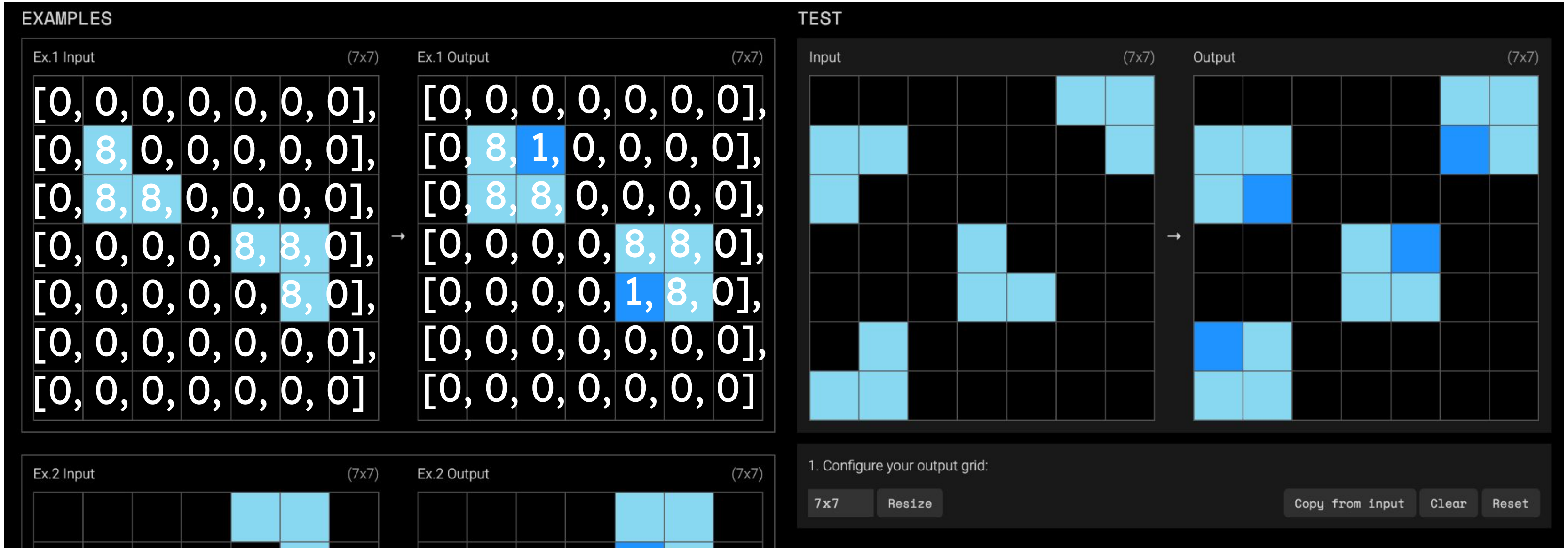
NEURAL INFORMATION PROCESSING SYSTEMS

ARC-AGI

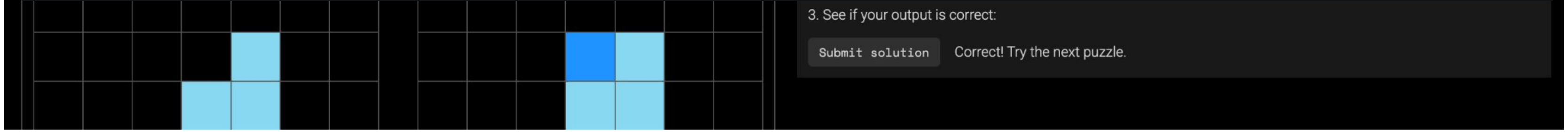
- A collection of small grid-based puzzles.
- Each puzzle tests whether AI can infer a transformation rule from examples, then apply it to solve a new case.



ARC-AGI



- Represent the grid as text (arrays of numbers) -> Show train input + output pairs -> Give a test input grid -> Ask the model to output the test output grid in the same format.

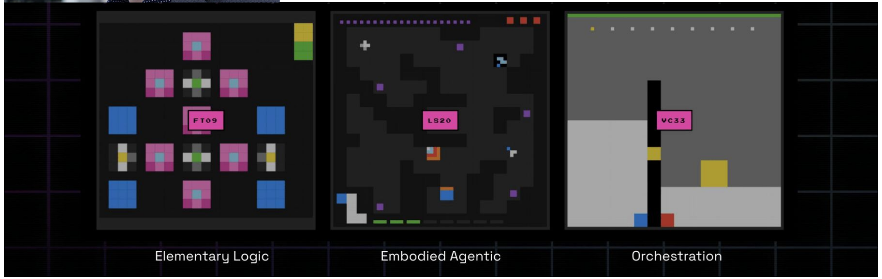


NEURAL INFORMATION PROCESSING SYSTEMS

ARC-AGI-3

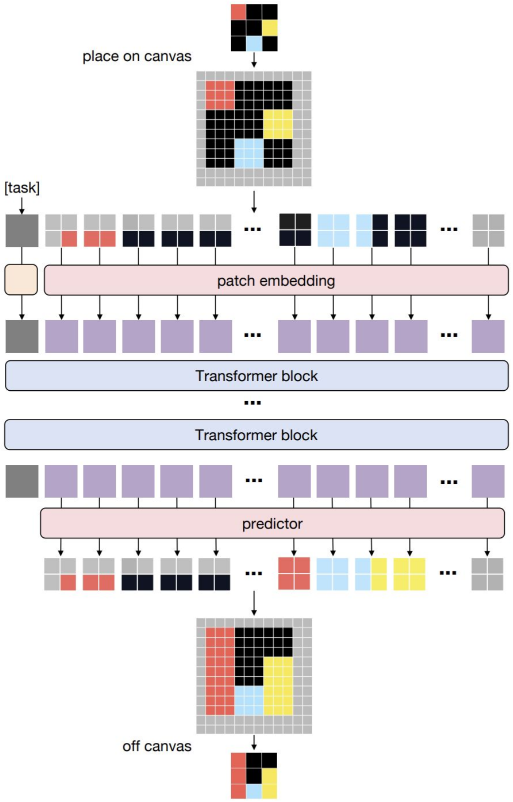
- ARC-AGI-3 leverages game environments to provide a rich medium for testing experience-driven competence in Interactive environments, which covers the following key capabilities:
  - Exploration
  - Percept → Plan → Action
  - Memory
  - Goal Acquisition
  - Alignment

François Chollet
Benchmarking Agentic Intelligence
(Keynote at LAW Workshop)



NEURAL INFORMATION PROCESSING SYSTEMS

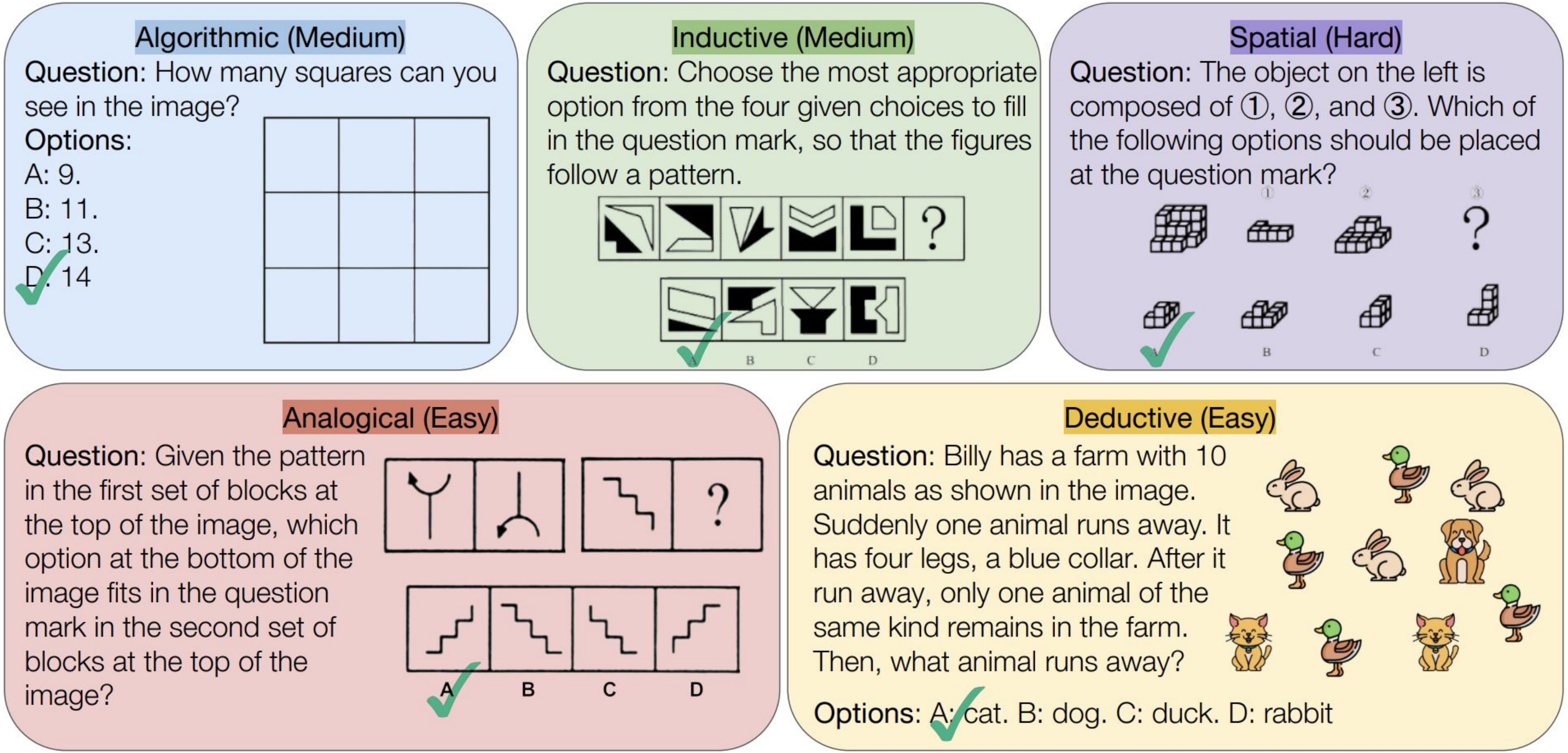
Is ARC a Vision Problem?



system	#params	ARC-1	ARC-2
large language models (LLMs)			
Deepseek R1 [21]	671B	15.8	1.3
Claude 3.7 8k [18]	N/A	21.2	0.9
o3-mini-high [18]	N/A	34.5	3.0
GPT-5 [18]	N/A	44.0	1.9
Grok-4-thinking [18]	1.7T	66.7	16.0
Bespoke (Grok-4) [8]	1.7T	**79.6**	**29.4**
recurrent models			
HRM [53]	27M	40.3	5.0
TRM [27]	7M	44.6	7.8
vision models			
VARC	18M	54.5	8.3
VARC (ensemble)	73M	**60.4**	**11.1**
human results			
avg. human [31]	-	60.2	-
best human [18]	-	98.0	100.0

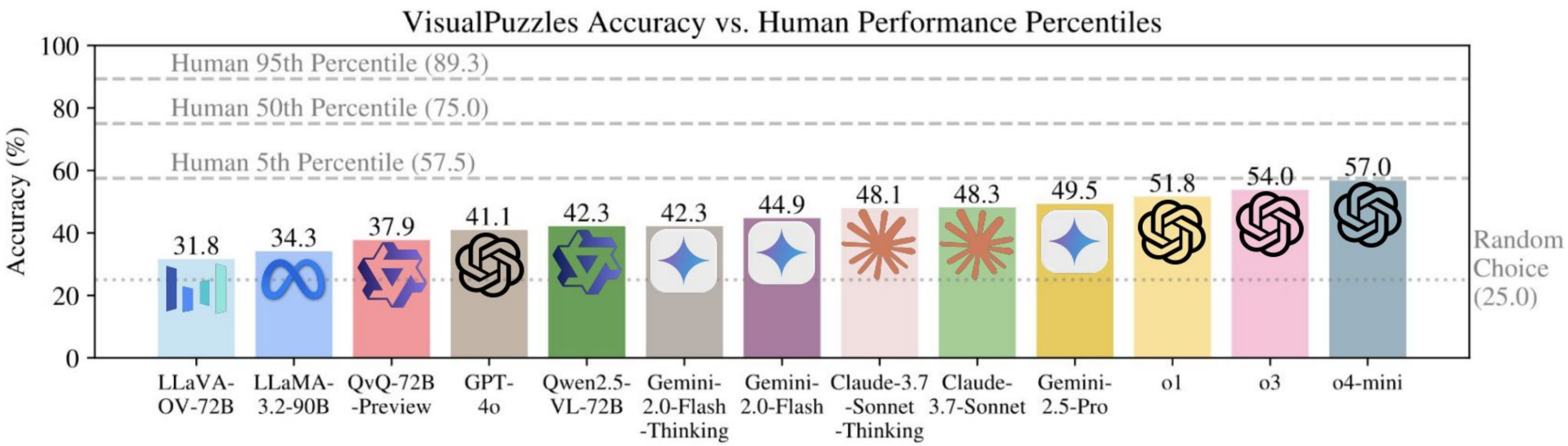
Keya Hu et al., ARC Is a Vision Problem! Preprint, 2025.

VisualPuzzles



Yueqi Song, et al. "VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge." arXiv 2025

VisualPuzzles

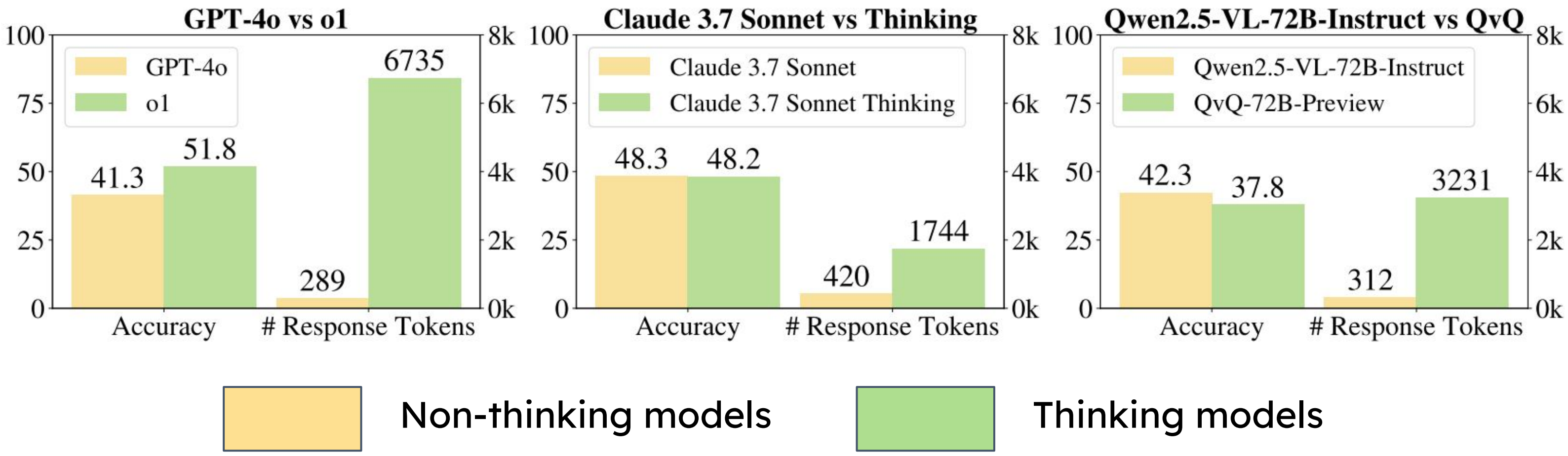


- All models are below the 5th percentile of the human performance (57.5): Gemini 3 scores 52.7, slightly below o3 (54.0) and o4-mini (57.0)

Yueqi Song, et al. "VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge." arXiv 2025

VisualPuzzles

Dynamics
Arena
Live Bench
Agency
Digital
Embodied
Realism
Adapt
Metric
Adv. Test
Proposal



- "Thinking" with more tokens does not always help!

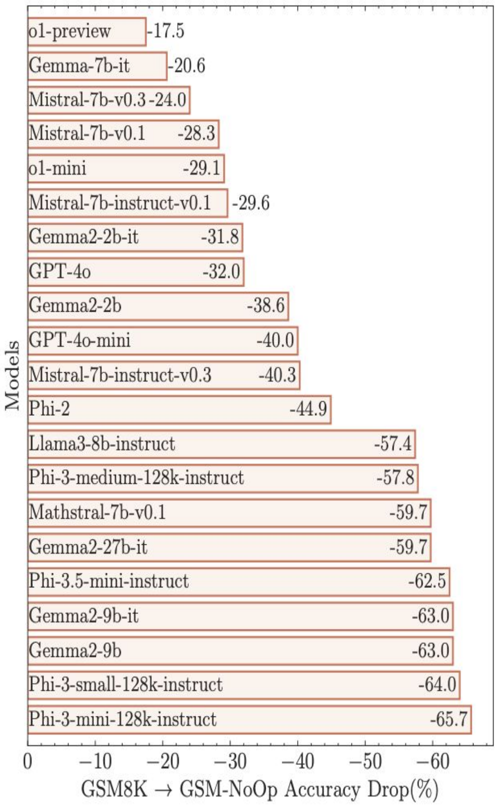
Yueqi Song, et al. "VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge." arXiv 2025

Summary: Measuring Fluid Intelligence

- **Measures Fluid Intelligence:** Tests the ability to adapt and learn new skills, not just memorize existing data.
- **Resistant to "Scaling":** Simply adding more training and inference compute or data doesn't guarantee success

- **Visual-Spatial Bias:** Heavily favors visual reasoning, potentially ignoring other forms of intelligence.
- **Closed Environment:** Represents a deterministic "toy world" rather than the messy ambiguity of the real world.

Adversarial Testing




```
Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double
the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many
kiwis does Oliver have?
------------------------------------------------------------------------------------------------
o1-mini: Oliver picks kiwis over three days as follows:
- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is 2 × 44 = 88 kiwis
However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the
Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from
all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185
kiwis.
```

- GSM-Symbolic adds seemingly relevant statements to the questions that are, in fact, irrelevant to the reasoning and conclusion.
- The majority of models fail to ignore these statements and blindly convert them into operations, leading to up to ~65% accuracy drops.

Iman Mirzadeh et al., GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. ICLR, 2025.

Proposals

- Third party benchmark maintenance as “public notary”;
- Living benchmarks updating periodically and compositionally;
- Private testset with absolute no leakage;

Proposals: Pilot Studies

- Absolutely-no-leakage-bench
  - Private;
  - Renew every once in a while compositionally;
  - Collected via crowdsourced photos taken on personal phones or cameras that were never posted on social media and are guaranteed to remain private, together with their associated metadata;
  - Reproduce common tasks and compare to most similar subsets from known benchmarks.

Proposals: Pilot Studies

- Absolutely-no-leakage-bench

Concept Segmentation

Model	Ours	LVIS (subset)
SAM3	61.25	57.75

Pointing

Model	Ours	PointArena (subset)
GPT-5	48.00	51.25
Gemini-3-pro	78.75	75.25

Visual Question Answering

Q: Are the trains going in the same direction?

Model	Ours	VQAv2 (subset)
GPT-5	75.75	86.00
Gemini-3-pro	81.50	88.00

*Our original preliminary results.

Proposals: Pilot Studies

- Example pairs.

VQAv2 (test)

Q: Are the trains going in the same direction?

GT: Yes

GPT-5: Yes

Absolutely-no-leakage-bench

Q: Are the trains going in the same direction?

GT: Yes

GPT-5: No

*Our original preliminary results.

Proposals

- Third party benchmark maintenance as “public notary”;
- Living benchmarks updating periodically and compositionally;
- Private testset with absolute no leakage;
- More realistic dynamic evolving environments for agents;
- Measuring adapting efficiency to new agentic tasks;
- Benchmarking long context performance;
- Addressing the cost of agent evaluation;
- Broadening the evaluation metric sets;
- LLM agent for benchmark submission quality checks (BetterBenchAgent);
- …

NEURAL INFORMATION PROCESSING SYSTEMS

Acknowledgement

Jing Ding
Shuyu Wu
Ding Zhong
Dezhi Luo

University of Michigan

Naomi Saphra

Harvard University

Yueqi Song
Hokin Deng

Carnegie Mellon University

Freda Shi
Yuansheng Ni

University of Waterloo

Panel Discussion

Eve Fleisig
UC Berkeley

Ofir Press
Princeton University

Wenda Xu
Google Deepmind

David Rein
METR Evaluations

Saining Xie
New York University

(Moderator)
Michael Saxon
University of Washington

Longevity

Why have evals stood the test of time? Have the ones that stood the test of time deserved to? What indicates something that is likely to last?

Tyranny of Metrics

The things we are capable of measuring shape the way we design evals. Broadly, how have the limitations of metrics shaped your research projects? What things can't be measured right now that you would like to change?

Human Subjectivity

Are we doing a good job of drawing the line on desiderata to account for divergent human preferences? How could we make this tractable? How should we account for diverse personal wants in alignment evaluation (or evaluation of other capabilities?)

Generality

To what extent do you believe the "general" part of AGI is measurable and why? And for generalization within subdomains, how can those be scoped?