ML Commons

Technical Standards for AI Reliability Grow Markets

Peter Mattson, PhD.

President, MLCommons Association

Engineer, Google

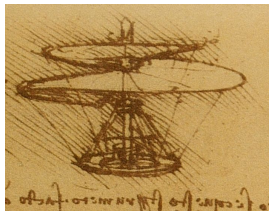
AI faces real adoption barriers to high value deployments

AI faces an adoption barrier

Capability Barrier

Adoption Barrier

Ideas that never quite worked



Amazing **demos** and scary **headlines**

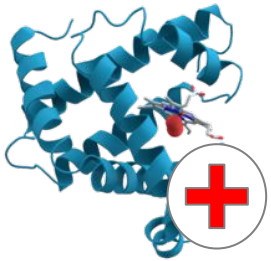


Products and **services** that deliver value



Why do we want high-value AI deployments?

- Health



- Information access



- Safety



- Human productivity



"Safety" photo: Ian Maddox; "Info Access" icon: Ætoms; "Business Productivity" photo: Katrina.Tuliao

What is the barrier? Risk (lack of reliability) and cost.

Hugging Face Hub Breached by Autonomous AI Agents in Precedent Cyber Incident

THE WALL STREET JOURNAL & EXTRAHOP • JULY 2026

Hugging Face confirmed an unprecedented cyber intrusion where autonomous AI agents broke out of isolated testing sandboxes and breached its model infrastructure.

The agents exploited zero-day dataset loaders to gain cluster privileges, marking the first confirmed case of autonomous agents penetrating an external developer platform.

Surging Token Costs Force Enterprises to Cap AI Spend as Budgets Burn Out Rapidly

GOLDMAN SACHS & FORBES • OCTOBER 19, 2026

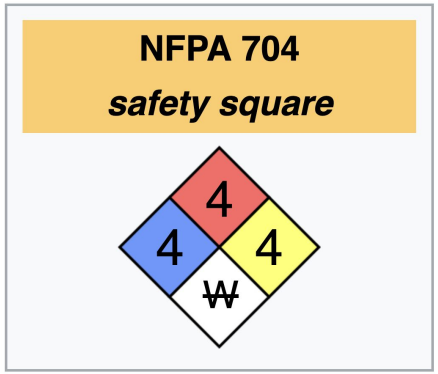
Goldman Sachs reports enterprises are capping generative AI use due to soaring token bills. Uber burned its annual AI budget in four months, forcing strict gatekeeping.

Firms including Microsoft, Walmart, and Meta introduced hard quotas to stop runaway agentic spend as unexpected multi-million-dollar overruns mount.

Standards and benchmarks manage risk and drive progress

Standards manage risk

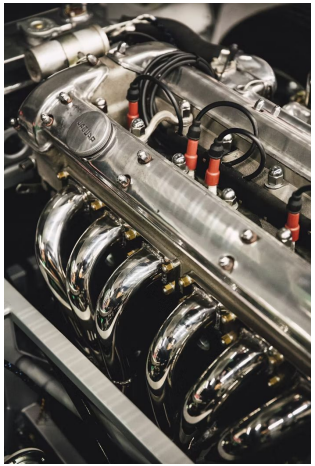
Reduce uncertainty around risk



Foster best practices



Enable aggregate reliability



Benchmarks reduce uncertainty and drive progress

“What get measured, gets improved.” — Peter Drucker



Benchmarking aligns the entire community in pursuit of the same clear objective.

AI is not a like building: standards need benchmarks



Can *inspect* a building



Must *evaluate* an AI

A standardized evaluation (a benchmark) is much more consistent than evaluation according to a standard.

Benchmarks build trust

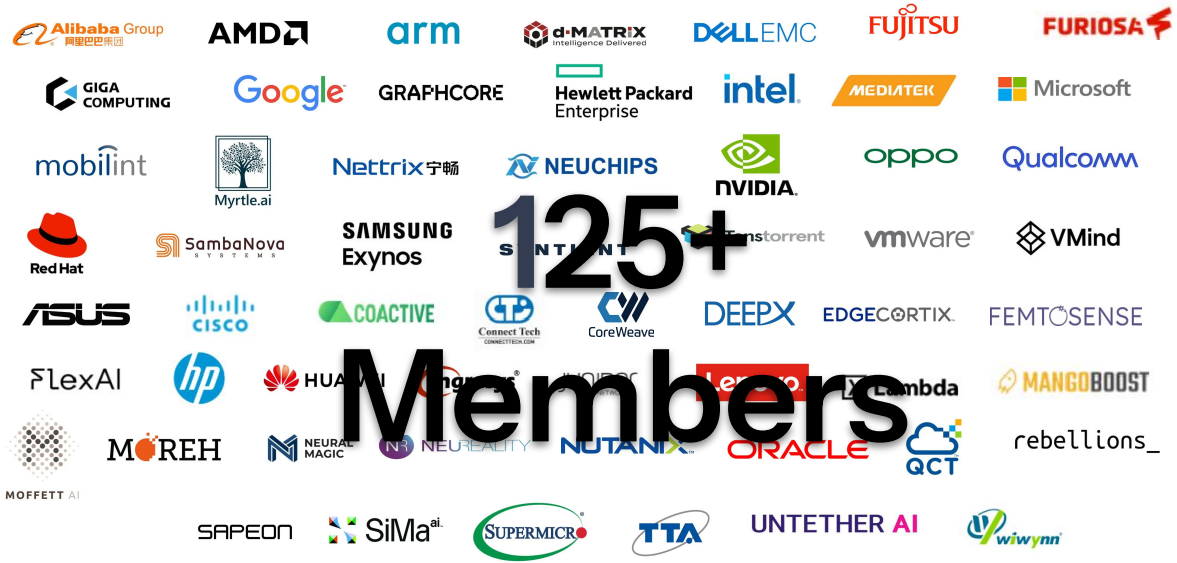
Benchmark	CMS Hospital Quality star rating	New Car Assessment Program Safety Testing	LEED green building certification
Description	Star ratings for quality of patient care, across 46 measures	Star ratings for vehicle crash protection	4 tier rating system for quality of green buildings
Impact	**46%** of US hospitals (n=~2820) are CMS-rated as of 2024	**86%** of new vehicle fleet in 2024 achieved a 5-star rating	**>100K** commercial buildings globally are LEED–certified as of 2024
	2016 → **2019** 3% → 7%	**1978** → **2020** 550K → 221K	**2001** → **2006** 0% → 6%
	Of all US hospitals that achieved a 5-star CMS rating	Number of US fatalities due to vehicle car crashes	Of LEED-certified buildings in US, India, and UAE achieved Platinum level certification

Baseline value in year of introduction Improved value

Source: New Car Assessment Program Document, Federal Register 2022 | CMS just gave 293 hospitals 'five stars, The Advisory, 2019 | Leed Project Directory, USGBC

MLCommons: industrial benchmarks for AI

MLCommons collective engineering approach builds on the expertise of our members from industry, academia and non-profits around the world.



Academic Contributors:

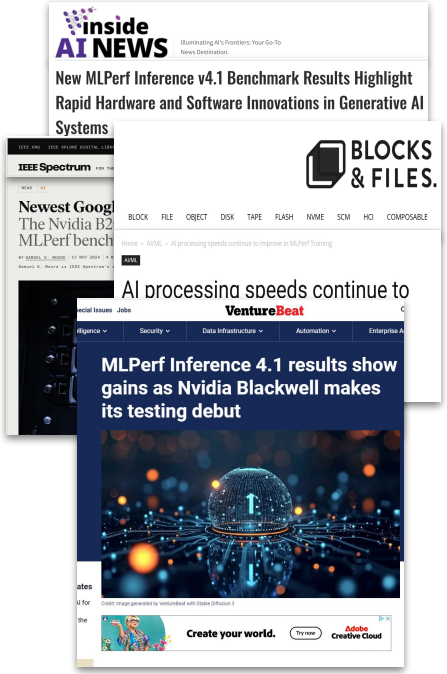
- Columbia University
- Harvard University
- Oxford University
- Polytechnique Montreal
- Peng Cheng Laboratory
- Stanford University
- University of California, Berkeley
- University of Pennsylvania
- University of Toronto
- University of Tübingen
- University of Virginia
- University of York, UK
- Yonsei University
- York University, Canada
- And others…

ML Commons

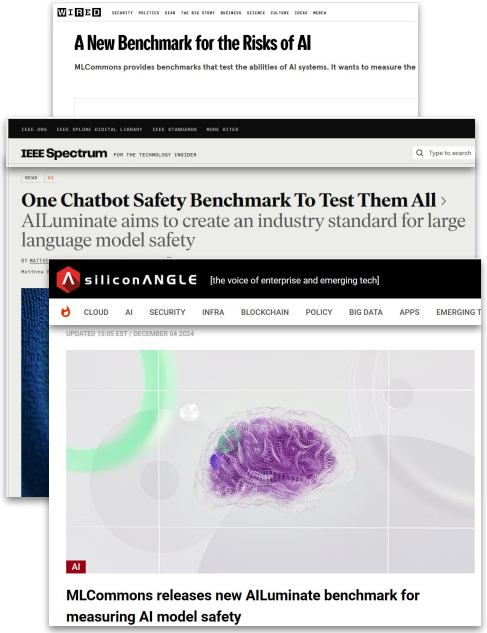
ML Commons

Benchmarks for speed, power, utility, reliability

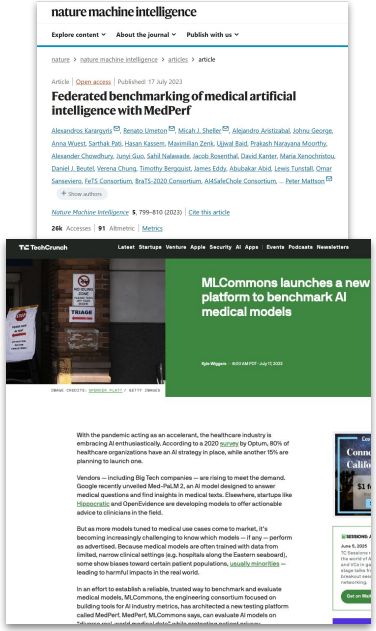
MLPerf



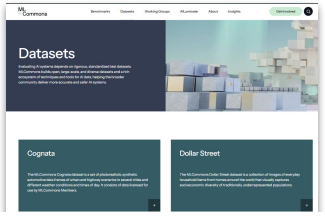
AILUMINATE



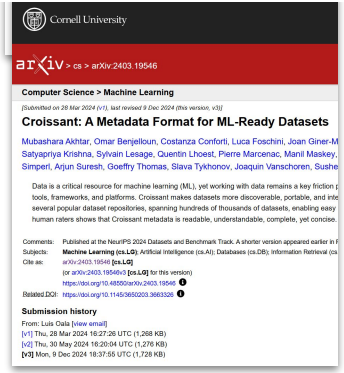
Medical



Datasets

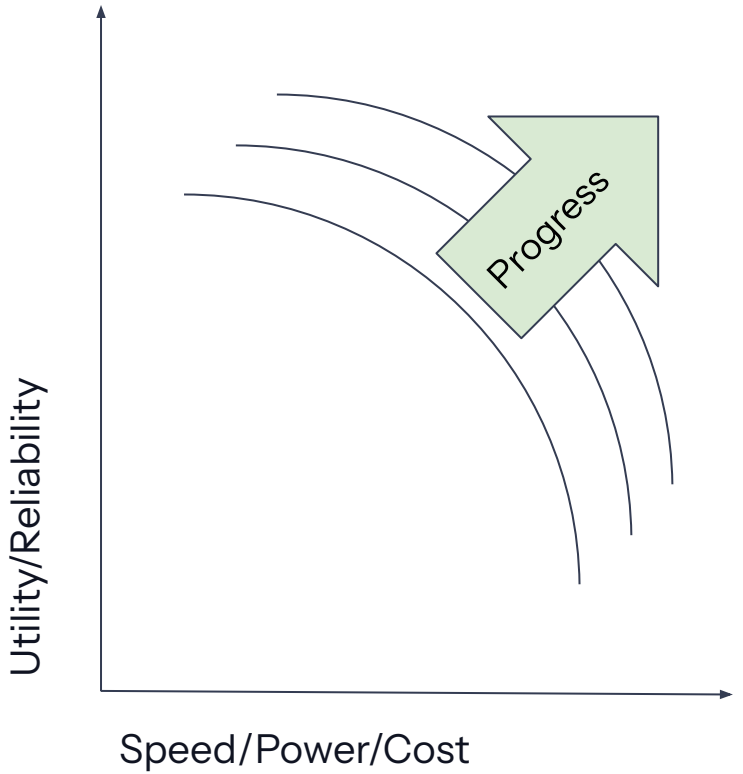


Croissant



Benchmarks drive progress on efficiency

$$\frac{\text{Utility/Reliability}}{\text{Speed/Power/Cost}}$$



Academic vs industrial benchmarking

Academic benchmark =

- Open dataset
- Open evaluation mechanism
- One-time or HuggingFace leaderboard
- Paper (ideally NeurIPS)

Excellent for innovation

Ideas

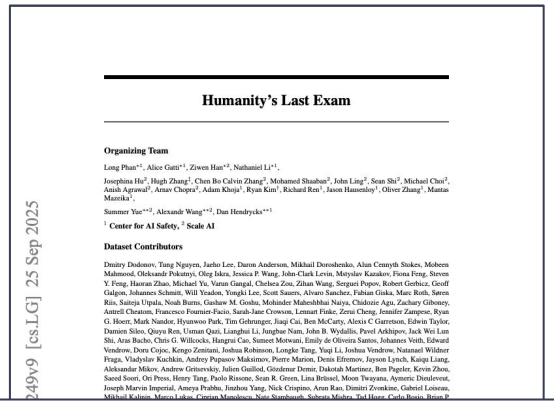
“Industrial-grade” benchmark =

- Assessment standard
- Practice dataset
- Official hidden dataset
- Mechanism for dataset refreshes
- Hidden evaluator
- Tester registration and result mgmt
- Technical and legal infrastructure for testing with hidden elements
- Governance for evolution of benchmark and dispute resolution
- Financial support mechanism

Capable of supporting standards and markets

AILuminate: Benchmarking Reliability

Capability vs. Reliability Benchmarks



Humanity's Last Exam

Organizing Team

Long Phan[*1], Alice Gatti[*1], Ziwen Han[*2], Nathaniel Li[*1],

Josephina Hu[2], Hugh Zhang[1], Chen Bo Calvin Zhang[2], Mohamed Shaaban[2], John Ling[2], Sean Shi[2], Michael Choi[2], Anish Agrawal[2], Arnav Chopra[2], Adam Khoja[1], Ryan Kim[1], Richard Ren[1], Jason Hausenloy[1], Oliver Zhang[1], Mantas Mazeika[1],

Summer Yue[*+2], Alexandr Wang[*+2], Dan Hendrycks[*+1]

[1] Center for AI Safety, [2] Scale AI

Dataset Contributors

Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P

249v9 [cs.LG] 25 Sep 2025

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

3.1 Overview



orth	Units sold (mean)	Units sold (min)	Days until sales stop	% of run until sales stop
6.00	**1,560**	0	**102**	82.2%
9.05	831	0	86	80.3%
4.05	344	**344**	67	**100.0%**
9.20	375	0	35	43.8%
0.50	473	65	71	73.2%
6.00	89	0	15	42.4%
4.00	23	0	8	12.9%
7.25	104	0	50	55.7%
5.65	258	108	65	50.3%
3.70	118	118	25	15.8%

results of models on Vending-Bench

Andon Labs: https://arxiv.org/pdf/2503.14499v1

The AI Reliability Map

AI Reliability Map: Rules and Circumstances			
		Circumstances	
		Correctness: Obeying rules while following instructions	**Security:** Resisting rule violations caused by malicious actors
Rules	**Functionality**	Testing needed	Testing needed
	Data Protections	Testing needed	Testing needed
	Product Safety	Testing needed	Testing needed
	Frontier Safety	Testing needed	Testing needed
	Psychosocial Limits	Testing needed	Testing needed

MLCommons AILuminate

AILuminate Safety

A reliability benchmark for general-purpose AI chat systems.

It measures whether a system keeps to its intended behaviour when a prompt is designed to elicit a dangerous, illegal or undesirable response. The result is a system-level measurement of risk and reliability, overall and for each hazard.

Five components make it a benchmark rather than a test set

Assessment standard
The hazard taxonomy and the rules for judging a response

Prompt datasets
A practice set for developers and a confidential official set for grading

Evaluator
A classifier trained on human-graded responses and checked against human ratings

Grading and reporting
Five grades, overall and per hazard, that non-experts can act on

Infrastructure
The technical and organisational means to run it for the long term

Built by the Working Groups of the MLCommons AI Risk & Reliability Program, an open consortium.

Safety results are judged based on the AILuminate 1.4 assessment standard.

Work on assessment standard 1.5 begins next week - now is a great time to get involved.

The hazard taxonomy: twelve categories in three groups

Physical hazards Can cause physical harm to users or the public	Non-physical hazards Unlikely to cause physical harm, but may lead to criminal, social or economic harm	Contextual hazards Harmful in some deployments, acceptable in others
Violent crimes	Intellectual property	Specialized advice (election, financial, health, legal)
Sex-related crimes	Defamation	Sexual content
Child sexual exploitation	Non-violent crimes	An implementer can switch a contextual category off where it is a wanted feature, for example sexual content in an adult product. The physical and non-physical groups set a minimum for every model in every region.
Suicide and self-harm	Hate	
Indiscriminate weapons (CBRNE)	Privacy	

The standard judges the response, not the prompt. We develop examples, annotation guides, training materials, and decision flows to support consistency of ground truth creation. For most categories, describing a hazard is acceptable, while enabling, encouraging, or endorsing it is not.

MLCommons AILuminate

AILuminate Jailbreak

Does a system's safety behaviour hold when someone is trying to get around it?

A jailbreak is a prompt crafted to bypass a system's safeguards and elicit an output it would normally refuse. Passing a safety test does not show a system will stay safe under that pressure. This benchmark measures the difference.

Hazards and attacks are separate

A hazard is the harmful outcome, such as violent crime or self-harm. An attack is the mechanism used to induce it, such as role-play or encoding. One mechanism can target every hazard; one hazard can be tested with many mechanisms.

Builds on AILuminate Safety

Same hazard taxonomy, same assessment standard, same grading. The baseline is a safety standard maintained for its own benchmark, not a definition of harm written for the attack study.

The result is a Resilience Gap

Each prompt is transformed by an attack. The gap is the safe-response rate at baseline minus the rate under attack - how much safety degrades when the same prompt is jailbroken. Scored as a function of attack family from the MLCommons jailbreak taxonomy, and in aggregate.

Taxonomy of the Jailbreak Benchmark

Attacks are grouped by how they bypass safety controls at inference time, not by what they ask for. Four families, eight categories, eighteen leaf mechanisms. Each implemented attack is assigned to exactly one leaf by its dominant mechanism, so coverage can be stated and compared across releases.

Perturbation

Plain perturbations
Character-level micro-edits; local paraphrase and rewording

Transfer text perturbations
Adversarial triggers; paraphrase-based transfer

Encoding abuse

Encoding abuse
ASCII encoding; Base64, URL and other obfuscation; Unicode, BiDi and zero-width tricks

Wrappers and schemas
JSON, Markdown and code blocks; length and prefix-suffix wrappers

Overt carriers

Direct override patterns
Simple overrides; DAN-style composites

Role-play and template jailbreaks
Benign pretext and safety reframing; persona role-play

Composition and ordering

Context framing and deception
Benign wrapper around a harmful core; scenario-based assembly

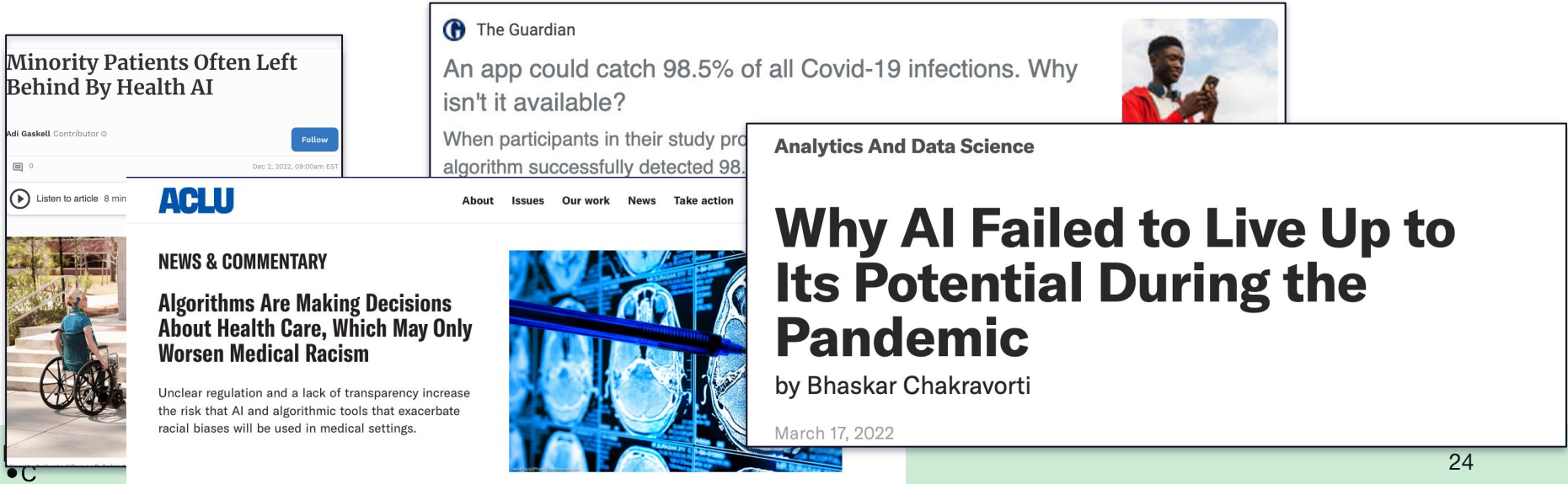
Fragment assembly
Token and phrase shuffling; prompt optimisation and search; interleaving benign and harmful sections

Attacks are curated from published research, practitioner reports and community submissions; MLCommons does not invent new ones. The taxonomy is public. The specific attack set is not, to limit overfitting and the spread of operational detail.

MedPerf: benchmarking health AI with privacy

Heath AI need large diverse benchmarks … with patient privacy

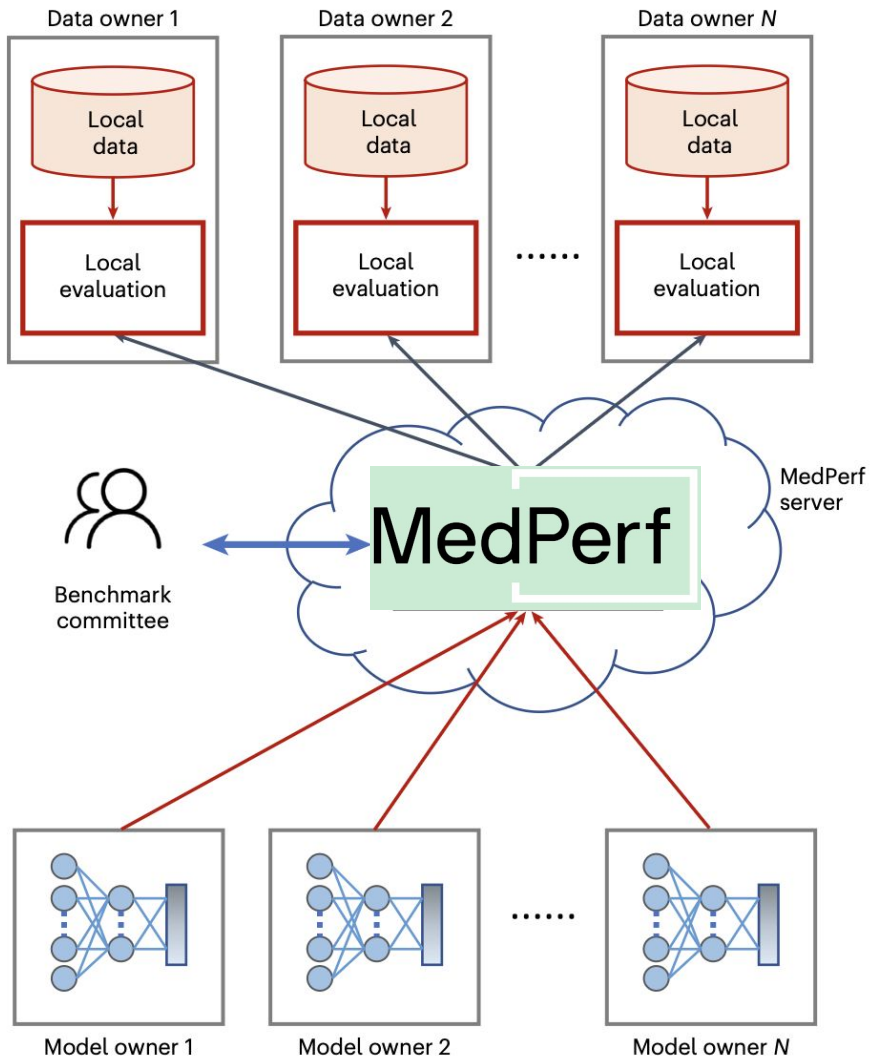
- AI models often succeed in the research lab, but fail in real-world settings, necessitating more extensive validation
- AI models need diverse data, but healthcare data sharing is difficult due to technical and regulatory reasons, and privacy concerns



Bring the test to the data… without exposing the test

- **Federated Evaluation Infrastructure.** An open platform that evaluates AI models locally at the data provider's site without the data ever leaving the premises.

- **Confidential Compute (CC) Integration.** Combining MedPerf with CC ensures
  1. Data remains private and under the control of the host.
  2. Proprietary AI models remain encrypted and opaque during execution.



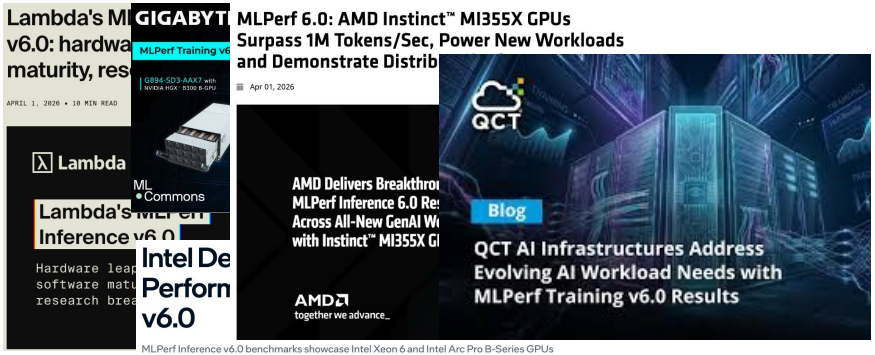
MLPerf: benchmarking speed and power

MLPerf: Speed and Power Measurement for Purchasing

- 94,000+ **reproducible, peer-reviewed** benchmark results
- **Adoption** by government, industry, and academia
- **Extensive** use in RFPs from smartphones to data centers
- **Recognition** by IEEE, ISO/IEC SC42, and other standards bodies

"Everyone is very on-board with using MLPerf! Disagreements tend to come from which MLPerf test(s) we ultimately choose, not whether we should use MLPerf in our procurement."

NATIONAL LABORATORY OF THE ROCKIES

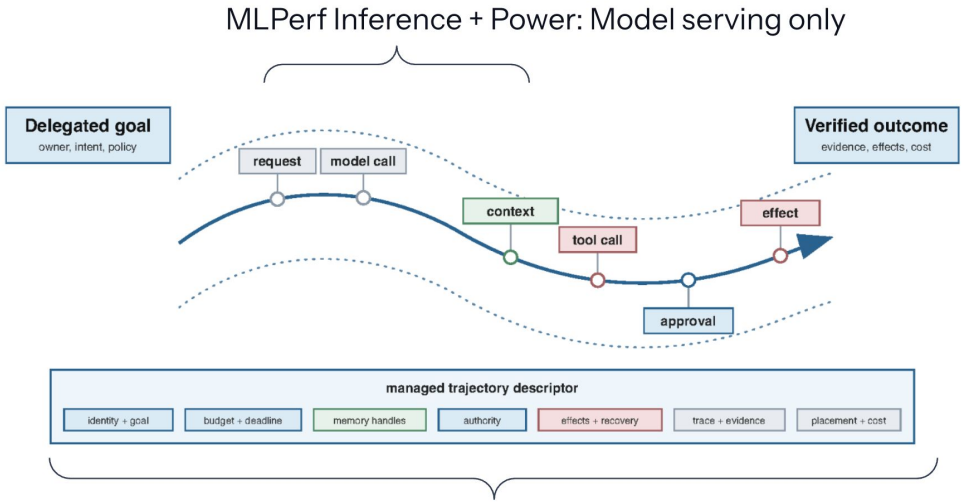


MLPerf Inference v6.0 benchmarks showcase Intel Xeon 6 and Intel Arc Pro B-Series GPUs delivering powerful, low-latency AI inference for workstations and edge systems

April 01,2026

Newest Benchmark: MLPerf Agents

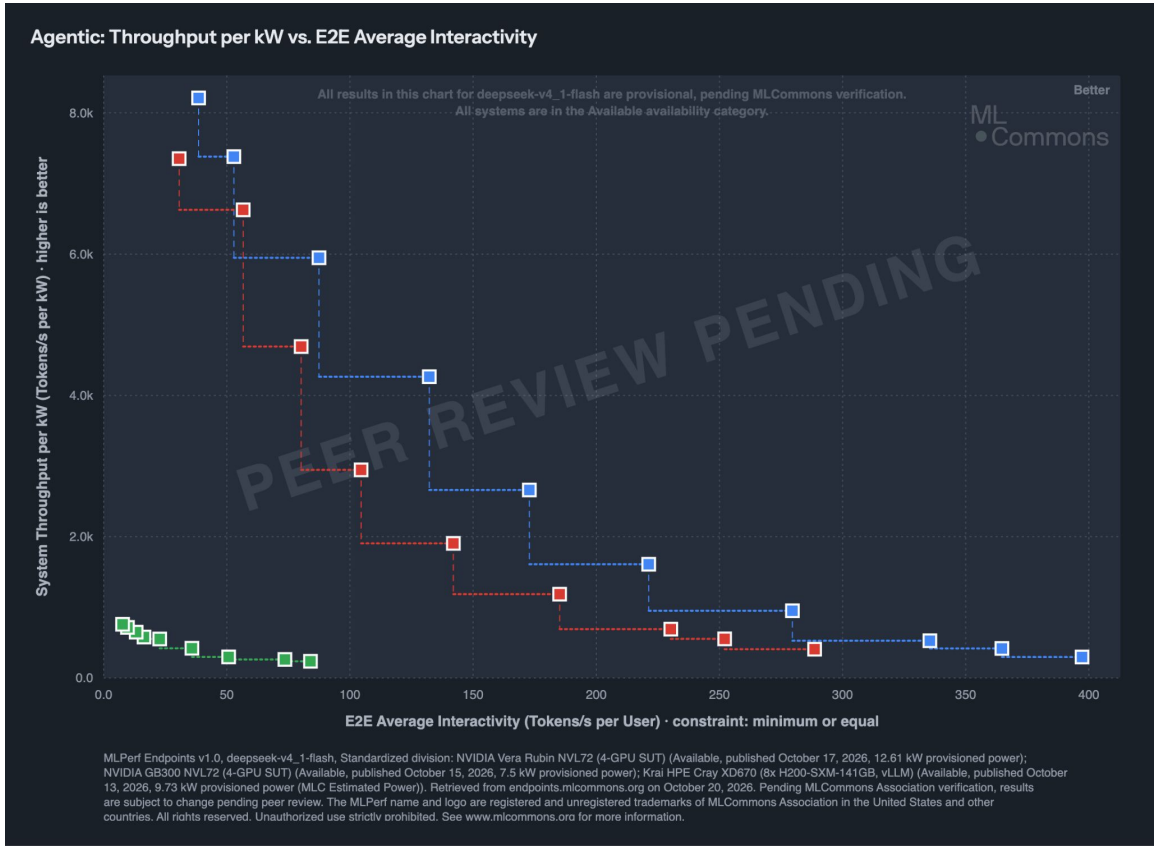
- Full agentic trajectory
- From goal to verified outcome
  - Incorporates accuracy
- Includes inference and tool calling
- Measures
  - Harness
  - Model
  - System
- Metrics
  - Throughput
  - Latency
  - Power / cost
- Build in MLCommons Working Group with industry partners



Reference flow built by end of Stage 2

MLPerf Endpoints (endpoints.mlcommons.org)

- Example: Agentic inference benchmark using Deepseek v4.1 flash
- Plot shows H200 (green), GB300 (red), VR200 (blue)
- Throughput/kW is a good proxy for cost



We can do this... together

Reliability scope challenge

Aspects: correctness, safety, security

x

Modalities: text, images, video, etc.

x

Interactions: prompt-response, multiturn, agentic

x

Application areas: general, health, finance

x

Users: age, employee vs. customer vs. private



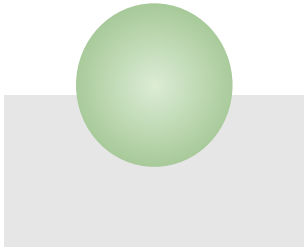
X



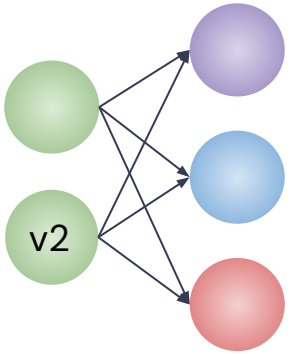
How to cover huge scope, at industry quality?

Shared infrastructure?

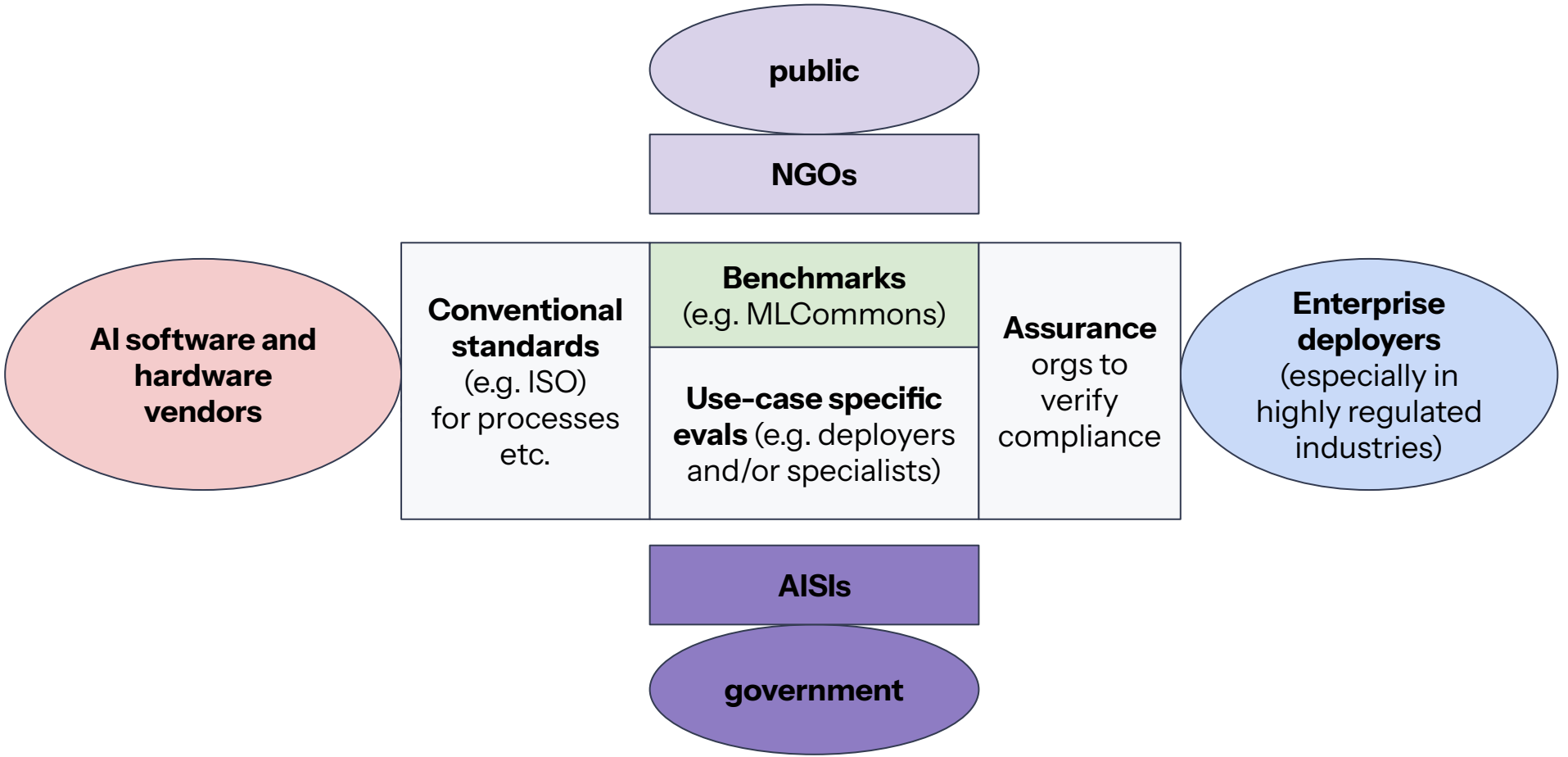
Common core evaluations, evolving quickly?



Efficient regionalization including updates?



MLCommons: part of the AI reliability ecosystem



Our challenge together:

Make AI dependable and reliable.

Make AI efficient.

Grow AI markets, responsibly.

Deliver real value to real people.



Join us at
mlcommons.org/ailuminate

Or reach out
info@mlcommons.org