How to Benchmark AI Without Getting Lost in the Buzz

Wed, 02/11/2026 - 09:00
|
Blaine Fisher, Ph.D.

Have you ever woken up and realized you do not remember a time before the smartphone? You remember what it was like to go to a library and look something up, not just open Google. You remember dial-up tones, Ask Jeeves, Hotbot, Yahoo, AOL. You remember unfolding a paper Rand McNally map instead of opening a Magellan GPS or printing directions from MapQuest.

A moment came when all that just was not a thing anymore. One day you went to sleep and the next morning the world had changed and you do not remember exactly when it happened. You just know it has always been that way now. That shift is what is happening with AI. It is already here. It is already woven into our lives.

But I think we lack awareness of the change. Imagine a simple example. You open an AI tool and ask it to help plan a family vacation. Instead of accepting the first answer, you start asking follow-up questions. You wonder why it suggested one hotel over another, or why the travel times look different from what you expected. You experiment with different prompts and compare results. That small act of poking, testing, and asking why is curiosity in action. It is the habit that keeps you engaged instead of drifting along while technology changes around you. That lack of awareness comes from a lack of curiosity and interest. Curiosity is a human instinct we need to preserve in a world where AI is becoming more common. Using AI well is not just about getting AI to do things for you. It is about using AI to learn, to grow, to understand more about the world and about yourself. Access to knowledge is one thing. Actually using that access to think is another.

Now let us talk about something practical. How do you keep up with AI? How do you know if a new model is worth paying attention to? How do you know if a model is good or if it is hype? The answer is benchmarking.

When we talk about benchmarks in AI, we mean tests and leaderboards that measure how well models perform on specific tasks. Companies use these benchmarks when they want to show that their model is better, faster, smarter, or more capable than the rest.

Here are some of the key benchmarks you should know about. These benchmarks were chosen because they represent different dimensions of AI performance, including reasoning, knowledge, practical usefulness, and human preference. I included direct links with the URLs so you can click through and explore for yourself.

ARC Prize: This is a benchmark and competition focused on general intelligence. The people behind it crowdsource prize money and invite teams to build models capable of solving spatial and visual reasoning tasks that humans find simple but AI struggles with. These tasks measure the kind of common sense understanding that even advanced models historically had trouble with. This is where you see if a model can reason with simple abstract scenes the way humans do, not just repeat information.

Website: *ARC AGI Prize (**https://arcprize.org**)*

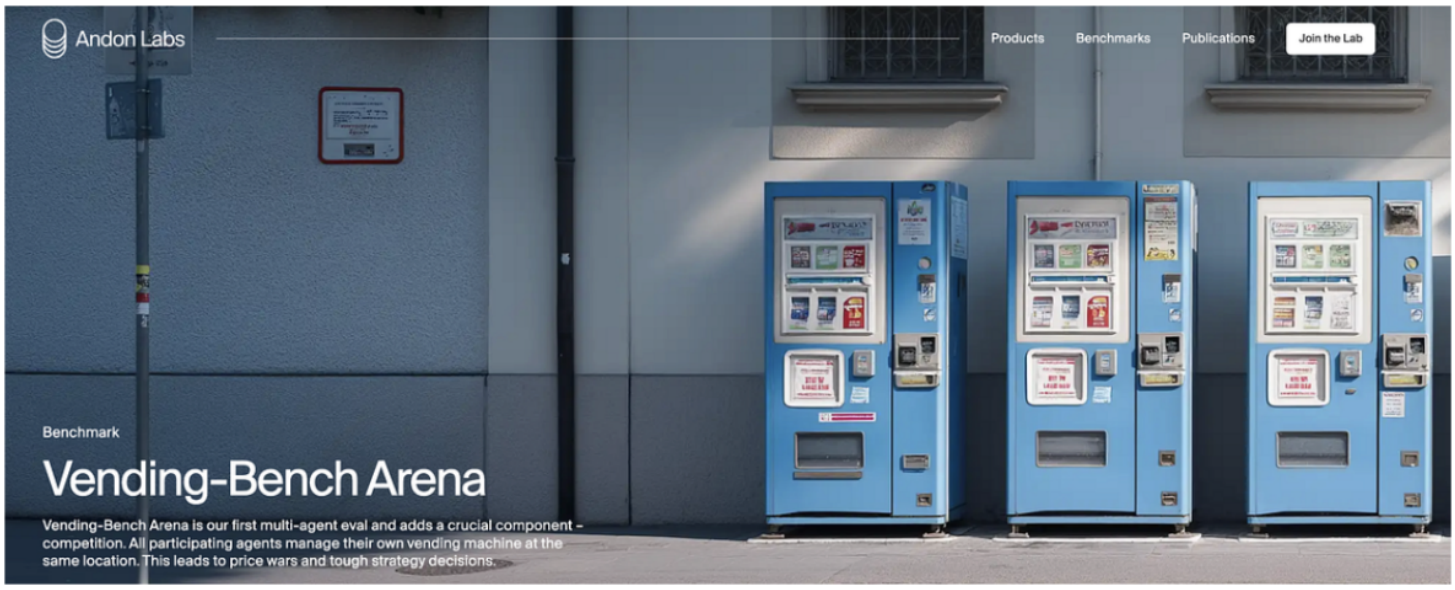
Humanity's Last Exam

Humanity's Last Exam: This one sounds dramatic but it is real and worth watching. Subject matter experts, researchers, academics, and leaders in science, technology, engineering, math, and other fields submit hard questions that really

test depth of understanding and problem solving. You can even register and submit your own questions if you think you have what it takes.

Website: *Humanity's Last Exam (**https://agi.safe.ai**)*

Think of this as the SAT or MCAT of AI. If ARC is about everyday reasoning and common sense, this is about intellectual depth and problem solving.



Vending Bench: Sometimes you do not care how smart an AI is in the abstract. You care how useful it is in real world tasks. Benchmarking Vending Bench is a creative example of that. Think about everyday workplace problems like planning a budget, scheduling deliveries, deciding how much product to order, or negotiating a better deal with a supplier. Those are the kinds of decisions this benchmark tries to

simulate. Models are put in charge of running a fictional vending machine company. Its tasks include managing inventory, negotiating contracts, and making business decisions. Instead of grading right or wrong answers, the leaderboard measures how much money the AI makes. That is a practical, concrete way to judge whether an AI would actually help a normal business operate more effectively.

Website: *Benchmarking Vending Bench (* ***https://andonlabs.com/evals/vending-bench-2****)*

LM Arena: This benchmark is different from all the rest because it tests preference and response quality. It works a bit like what you might think of as a Rotten Tomatoes audience score. You enter a prompt, and LLM Arena shows you two responses from different models. You choose which one you like better. After you choose, it tells you which model generated the response. Over time, this creates a leaderboard based on what people prefer rather than what a test says is correct. It is a crowd-sourced measure of how people like the results, not just how accurate or smart they are.

Website: *LLM Arena (****https://arena.ai****)*

Coding and Math Competitions There are benchmarks that measure how well AIs perform on human coding competitions, math olympiads, and professional exams. These include competitive programming challenges and tests where humans traditionally strive to excel. Watching how models perform here highlights areas where AI is beginning to surpass human capability and where it still struggles.

Examples you can explore:

- **Codeforces** for coding challenges (**https://codeforces.com**)
- **Project Euler** for math and logic problems (**https://projecteuler.net**)

_____

Many of these platforms have leaderboards showing AI entries and how they stack up against humans.

Benchmarking is how you see where AI innovation is right now. It covers common sense reasoning, deep academic problem solving, real world usefulness, user preference, creativity, coding ability, and more.

The next time you hear that a new model has launched and a company claims it is the best, do not take it at face value. Look at the benchmarks that matter to you. Are you building tools for business? Are you curious about pure reasoning? Are you more interested in how humans feel about the responses? Check the corresponding leaderboards, explore the links, and form your own opinion.

Then send me a message and tell me what you found. Curiosity and critical thinking are still the best tools for making sense of the AI world. Tell me which models impressed you. Tell me which ones you want to use in your work. Let us stay curious together.