— 15 —
Epilogue

Source: The Emerging Science of Machine Learning Benchmarks. M. Hardt, 2025. URL: `https://mlbenchmarks.org`. Compiled on 2025-11-07.

In 1916, the mechanical engineers Durand and Lesley set out to find a better aircraft propeller. The old *blade element theory* for predicting the performance of propellers only went so far; it couldn't provide actionable guidance on propeller design in practice. Operating outside the confines of established physical theory, Durand and Lesley instead subjected dozens of propeller models to a sequence of quantitative tests. After pinning down a metric for *propeller performance*, the engineers systematically varied model parameters to evaluate performance in a wind tunnel.

In *What Engineers Know and How They Know It*, Vincenti describes the process as a form of *experimental parameter variation*, repeatedly testing the performance of a device while systematically varying its parameters. Already deployed by catapult designers in Ancient Greece, parameter variation has long played a central role in the creation of engineering knowledge. But the crucial point of experimental parameter variation, according to Vincenti, is to "circumvent science" altogether:

> Indeed, the strength of experimental parameter variation is precisely its ability to provide solid results where no useful quantitative theory exists. [...] the *function* of experimental parameter variation may be to free engineering from limitations of science.[1]

Paradoxically, the knowledge empirical testing creates ultimately becomes science. In fact, Strevens takes things a step further in *The Knowledge Machine: How an Unreasonable Idea Created Modern Science*. Competitive empirical testing *is* the iron rule of modern science. Scientists are free to do whatever they want—*anything goes*—but empirical testing must resolve all disagreements. The anything goes creates variation, the iron rule enforces consistent selection. It defines the rules of the game that scientists stick to as they climb the leaderboard of knowledge.

In committing to benchmarking, machine learning undoubtedly borrows from its engineering roots. More than that, benchmarking is how machine learning research evidently implements the iron rule. These two facts paint a deceptively simple picture of machine learning as a modern empirical science rooted in engineering principles. But there's a wrinkle to the clean story of machine learning as a modern engineering science.

Even though machine learning benchmarking seems to stand on a reliable engineering foundation, it differs in one crucial respect. Where engineering ultimately meets its guardrails in the laws of physics, machine learning benchmarks only have the laws of statistics to lean on. And those aren't as strong.

Benchmarking is practically statistical testing. A vast scientific crisis engulfed statistical testing, taking with it volumes of irreproducible results from numerous fields. Although severe, the crisis was hardly a surprise. Goodhart's law predicts that statistical measurement breaks down under competitive pressure. Statistical testing is no match for the relentless competition that modern science incentivizes. The same competitive pressure should have corrupted machine learning.

And yet, machine learning got away with it. In fact, it thrived.

The ImageNet era was a triumph for benchmarking. Aggressively competing over dog breed classifiers somehow promoted real progress in computer vision. It's an almost absurd outcome that we've all just gotten used to. Likewise, bringing the perplexity down on Penn Treebank somehow created the fertile ground from which Transformers emerged, kicking off the AI revolution we're witnessing. Measured by its own success, it seems machine learning escaped all crisis.

The truth is that it has... and it hasn't.

In some meaningful ways, benchmarks don't actually work—at least not reliably. The absolute numbers on a leaderboard typically mean very little. Benchmark numbers change erratically from one dataset to another, even between datasets that are seemingly similar. Dataset biases and artifacts further distort the picture. As a result, it's hard to assign any deeper meaning to benchmark numbers. In particular, benchmarks have a poor track record in measuring latent constructs, such as skill, ability, or inclination. The many sound critiques of benchmarking often boil down to a failure of valid measurement one way or another.

What rescued machine learning research is its essential dependence not on absolute numbers, but on rankings. Reliable rankings are relatively easy to come by when all models train on the same data. In this case, good model rankings don't even require realistic datasets. Rankings on toy data often match those on other datasets. What makes rankings special is that they have resilience to competitive pressure. Even if all benchmark scores are strongly inflated due to competition, the resulting ranking can still be valid.

Valid rankings are enough to implement the basic mechanic of the iron rule—*beating the previous best*. It's what's needed for continual improvement. To the extent that machine learning thwarted a scientific crisis, it is because benchmarking runs on rankings and not on precise quantitative measurement. This is also a salient difference between machine learning and the physical engineering disciplines. Where engineering has the luxury of precise quantitative measurement, machine learning runs on a weaker currency: comparisons and rankings.

The real threat to ranking isn't competition—it's *inequality*.

Any good competition must be clear on what it equalizes. Runners must *both* start on the same line and finish on the same line. If you fix the finish line but let runners have varying starting positions, you get meaningless results. If half the students in your class take expensive tutoring lessons and the other half doesn't, your exam will tell you more about family wealth than about ability.

Current LLM evaluation is a bit like a running competition with no clear start or finish line. It's like an exam that only some prepared for; others didn't even know about it—and we can't tell who did and who didn't. Querying proprietary black boxes through API calls often resembles a game of Twenty Questions more than a scientific competition. What's missing in those cases is a clear understanding of what factors should be equalized and how. When we take care to sort these out, however, valid rankings do emerge.

Rankings aren't invincible; the LLM era has shown us *how*. The chaos that is LLM evaluation has jolted confidence in the benchmarking enterprise. Some say it's about time we moved on from benchmarks. Others double down and vow to build them bigger, better, and harder. Benchmarking is increasingly a lucrative business model with startups raising venture capital for model evaluations that promise to fix the mess.

Still others hope for a future where the real world might solve the benchmarking crisis. Perhaps benchmarking will no longer be necessary for AI agents so advanced that they can carry out real-world tasks. Just put a coding agent on a gig labor platform and count how many dollars it brings home. Who could argue with real monetary objectives as a means of comparing AI systems? Perhaps the ultimate benchmark is how many jobs AI has taken.

Business applications, however, don't seem to make benchmarking obsolete. On the contrary, benchmarking has lived a parallel life in the business world of operations management and quality control. After scrutinizing forty-nine definitions of benchmarking in the management literature, benchmarking proponent Spendolini came up with his own. Published in 1992, *The Benchmarking Book* defines *benchmarking* as "a continuous, systematic process for evaluating the products, services, and work processes of organizations that are recognized as representing best practices for the purpose of organizational improvement."[2]

In the management view, there is no such thing as a single benchmark. There's only *benchmarking*. Benchmarking encompasses the collaborative practices of individuals and organizations aiming to improve their products. Benchmarking is never static—it's a process. Thinking of it as a process shifts the focus from any single test to setting up a benchmarking ecosystem that promotes progress. Putting agents into real-world settings won't make benchmarking obsolete—rather the opposite—it may reunite the practice with its management relatives.

There's another good reason to expect a new wave of benchmarking. Smitten by Nobel Prize-winning AI models, scientists are rushing to apply AI to whatever scientific challenge seems suitable. After benchmarking powered breakthroughs in protein structure prediction, scientists hope to achieve similar triumphs elsewhere. Benchmarking has already transformed weather prediction from mostly physics-based models to physics-informed AI models. In chemistry, the Open Catalyst project, built on terabytes of data, maintains a leaderboard for predicting how molecules stick to metal surfaces.

AI is the Trojan horse that will carry benchmarking behind the walls of the higher sciences. Rather than making AI more scientific, the sciences will adopt more of the engineers' science-circumventing behavior.

For the foreseeable future, competitive empirical testing will be the social institution in charge of adjudicating scientific and industrial progress. This mandate reflects Vincenti's analysis of engineering history as much as Spendolini's management wisdom and Strevens' philosophy of science. And if engineers, managers, and philosophers can agree on a way of doing things, there must be something to it.

As Strevens argues, researchers understandably grow resentful of the chokehold the iron rule has on knowledge production. They'll try to change it. Perhaps exceptions should be granted. Perhaps it should be abandoned altogether. Resist the urge, he pleads:

> Do not, then, meddle with the iron rule. Do not tamper with the workings of the knowledge machine. Set its agenda, and then step back; let it run its course.[3]

If we're stuck with this machine—handed down to us without a blueprint or a manual—it will take some care to understand how it works and how to maintain it. I hope this text will help the future *engineers of science* keep it running.

Bibliography

1. Vincenti, W. G. *What engineers know and how they know it* (Johns Hopkins University Press, 1990) (↑ 2).
2. Spendolini, M. J. *The benchmarking book* (American Management Association, 1992) (↑ 5).
3. Strevens, M. *The knowledge machine: How an unreasonable idea created modern science* (Penguin UK, 2020) (↑ 5).