AI Model Leaderboards: Myth vs Reality, 2 Votes Flipped #1 and 78% of Codeforces Was in Training

“Dropping just a handful of preferences can change top large language model rankings.” That is the title of an ICLR 2026 poster from MIT, and the handful turns out to be two votes out of 57,477.

Two preferences. 0.00348% of the data. Remove them and Chatbot Arena’s top model changes from GPT-4-0125-preview to GPT-4-1106-preview. The MIT team (Jenny Y. Huang, Yunyi Shen, Dennis Wei, Tamara Broderick) ran the same procedure across the other arenas and got numbers of the same order: 9 of 49,938 on the LLM-judge Arena, 18 of 10,501 on Webdev, 28 of 29,845 on Vision, 61 of 24,469 on Search. None of these leaderboards needs a scandal to produce a different winner. A rounding error is enough.

That is a targeted deletion, which is the worst case. The random case is not much better. Drop 1% of Chatbot Arena’s preferences at random and the top-1 model survives in 77% of

trials. Roughly one in four alternate universes has a different champion at the top of the page your procurement team screenshotted.

Two votes can move the number your architecture decision is anchored to

I have watched more than one architecture decision get anchored to a leaderboard position that was inside the noise floor. The cost is rarely the model choice itself (the frontier models are close enough that most workloads barely notice), it is the reasoning chain that follows. You pick model A because it is #1, you build the harness around its quirks, you negotiate a committed-spend contract, and eight months later you have a migration cost that exists entirely because of a number that two votes could have moved.

The second cost is worse and lands in August. EU AI Act GPAI obligations have applied since 2 August 2025, and Commission enforcement powers, including fines, apply from 2 August 2026. The General-Purpose AI Code of Practice requires model evaluations carried out with high scientific and technical rigour, specifically naming internal validity, external validity, and reproducibility. The 2026 literature I am about to walk through shows public leaderboards failing all three. If your compliance evidence for a systemic-risk model consists of citing a public Elo score, you are citing something a peer-reviewed paper has shown to be unstable under a 0.003% perturbation.

Four leaderboard claims that the 2026 papers take apart

The first claim: a leaderboard built on tens of thousands of human preferences is statistically settled, and small changes cannot move the top. The MIT result says otherwise, and the mechanism is not mysterious. Elo-style rankings compress a huge pairwise comparison graph into a single scalar, and when two models are close, the ordering hangs on a small number of decisive comparisons. The paper does not claim the arenas are wrong about which models are good. It claims the ordinal top of the list, which is the only part anyone reads, is unstable.

The second claim is about contamination, and this is where I changed my own mind this year. I used to treat contamination as a binary: either the exact test string is in the training corpus or it is not, and dedup catches it. Spiesberger et al. embedded the full Olmo3 training corpus and looked for near-matches rather than exact ones. They found semantic duplicates for 78% of Codeforces benchmark problems and exact duplicates for 50% of

ZebraLogic problems.

The interesting part is what happened when they finetuned on those duplicates. Scores rose on the trained items, which is expected. Scores also rose on held-out items from the same benchmark, which is the part that should worry you. And the gains "rarely transferred to related benchmarks." So the benchmark number went up without the underlying capability going up. It measured how well the model had absorbed the shape of that particular test. A GEM 2026 systematic review puts contamination-driven inflation at roughly 6% to 40% depending on benchmark and auditing assumptions. That range is wide, and I want to be honest that the width is the finding. Different auditing assumptions give different answers, which means nobody currently knows the true inflation on a given benchmark. That uncertainty is itself disqualifying for a compliance artifact.

Detection is also shakier than the tooling suggests. Test of Time (ACL 2026, with Schölkopf, Sachan and Jin among the authors) tested the popular post-cutoff-decay detector, the one that infers contamination from a model performing worse on questions about events after its training cutoff. Questions built from the same source documents produced markedly different temporal signals depending on whether they were LLM-transformed or cloze-style. The detector is measuring question construction as much as contamination.

The third claim is that a high benchmark score transfers to the capability the benchmark names. The ARC living survey is the cleanest counterexample I have seen. One frontier model scores 93.0% on ARC-AGI-1, 68.8% on ARC-AGI-2, and 13% on ARC-AGI-3, while humans stay near-perfect on all three versions. Same task family, same abstract reasoning label, three different worlds. Across 82 approaches the survey documents a consistent 2 to 3x drop from ARC-AGI-1 to ARC-AGI-2 in program synthesis, neuro-symbolic and neural paradigms alike. That detail matters more than the headline number: **the degradation is architectural, not model-specific.** It shows up regardless of paradigm, which rules out "this particular lab overfit" and points at the benchmark version boundary itself. ARC Prize 2025's top Kaggle score reached 24% on ARC-AGI-2, and the winners needed hundreds of thousands of synthetic examples to get there.

The fourth claim is that a single global score answers your question. It does not, because your question is narrower than the average. "Who Defines Best?" (FAccT 2026) analysed the LMArena dataset and found it heavily skewed toward certain topics, with model

rankings varying substantially across prompt slices. One Elo number averages over disagreements that would reverse the ordering if you sliced by the workload you actually run.

Leaderboards are a coarse filter, and that is all they are

Public leaderboards are useful as a coarse filter and useless as a decision. They will tell you which five or six models are in the frontier tier. They will not tell you which of those five is best for your traffic, because the ordering within the tier is unstable to a rounding error (MIT), partly manufactured by training-data overlap (Olmo3, GEM), unlikely to transfer across benchmark versions of the same task (ARC), and averaged over a prompt distribution that is not yours (FAccT).

There is a fifth failure mode that these papers do not cover and I have written about separately: the scaffold. The same model on the same benchmark can score 65% or 74% on SWE-bench depending on the harness around it. Stack that variance on top of a 6% to 40% contamination band and a top-1 position that two votes can move, and no footnote closes the gap between a leaderboard number and a production expectation.

> ✓ **What to build instead**
> Take 200 to 500 real requests from your own logs, redact them, and score candidate models on those with a rubric your domain experts wrote. Freeze the set, version it, and re-run it on every model upgrade. It costs a few engineer-days, it is private (so it cannot leak into anyone's training corpus), and unlike an Elo score it is reproducible in the sense the EU Code of Practice means. Report the confidence interval, not just the mean.

If you cannot build that, at minimum stop treating rank order as information. Treat the frontier tier as a set and choose on latency, price, data residency, and how the vendor behaves when you file a support ticket. Those are measurable and they do not move when someone drops two votes.

Why the myths hold, and why August is what breaks them

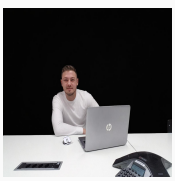
My take: leaderboards persist because they are the only artifact that is free, public, and comparable across labs, and because every party has a reason to keep them. Labs need a headline. Buyers need a defensible line in a procurement memo. Journalists need an

ordering. The MIT result is not new mathematics. It is a stability analysis nobody had bothered to run on a number that had already been cited in thousands of decisions. I expect that pattern to repeat: the technical rebuttals to leaderboard culture will keep being cheap experiments that were simply never done, because nobody was paid to do them.

What changes it is the 2 August 2026 enforcement date rather than the papers. Once a systemic-risk provider has to submit a Model Report under the Code of Practice and stand behind evaluations claiming internal validity, external validity and reproducibility, "we scored well on Arena" stops being an answer. The likely outcome is a split, where labs publish rigorous internal evaluations for regulators and keep publishing leaderboard positions for marketing, and the two numbers drift apart. Some labs are already skipping the marketing half: Alibaba shipped a 2.4-trillion-parameter model with no official benchmarks at all, which I read as a signal about how much weight the numbers still carry internally.

Unconfirmed, and I want to flag it as my speculation rather than a finding: I suspect the semantic-duplicate rate that Spiesberger et al. found in Olmo3 is on the low end for the industry, because Olmo3 is an openly documented corpus with a published decontamination process. Closed corpora have no such constraint and no external auditor. **If 78% of Codeforces has semantic duplicates in the corpus that tried hardest to be clean, I would not assume better elsewhere.**

The practical version of all this is short. Stop asking which model is best. Start asking which model is best on your 300 prompts, with your harness, at your latency budget, and write down the confidence interval so that next quarter you can tell whether the upgrade actually helped. If you want to talk through what that evaluation set looks like for your workload, or what your Model Report needs to survive August, I am happy to have that conversation. Bring your logs, not your leaderboard screenshots.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →