

The Leaderboard as Galton’s Fair

Michael P. Laing

`mllaing@post.harvard.edu`

ORCID 0009-0004-5593-5759

Written with Anthropic Claude Fable 5.1

August 2026; revised September 28, 2026

Companion essay to “What a Deployment Correlates” (September 2026) and to “Measure-Relational Algebra: The Core” (September 2026; doi:10.5281/zenodo.22999911). Sibling of “The Committee That Never Disagrees” (realql.org), which tells the committee half of this story. Canonical reference: realql.org.

The purchase

An airliner carries two engines not because one is insufficient but because two can fail *separately*. If each engine quits on one flight in fifty thousand, and the failures have nothing to do with each other, both quit together on one flight in two and a half billion. That leap — from one-in-fifty-thousand to one-in-billions — is not bought with metal. It is bought with *independence*. The second engine is only worth its weight if whatever kills the first one leaves it alone.

Engineers know this so deeply it is almost reflex, which is why the exceptions are famous. Feed both engines from a tank of contaminated fuel and the arithmetic collapses: the failures were never independent, the multiplication was never licensed, and the second engine was, on the day it mattered, decoration. The whole discipline of redundancy is a discipline of hunting shared causes — common fuel, common wiring, common software — because a shared cause silently refunds the purchase and keeps the receipt.

Hold that thought, because our era is conducting a very large experiment in shared causes, and most of the parties to it believe they are flying on separate engines.

The fair

In 1907 Francis Galton reported that nearly eight hundred fairgoers, guessing the weight of an ox, had missed individually by wide margins — while the *average* of their guesses landed within a percent of the truth. The wisdom of crowds is real, and it is the same purchase as the second engine: it works because the guesses were wrong *separately*. Each fairgoer erred out of private ignorance, and eight hundred private ignorances, pointing every which way, cancel.

A sibling essay on this site follows what happens when the crowd’s members stop being separate — when a committee’s members all consult the same acquaintance. This one follows

the purchase into two places where modern machine practice spends it every day: getting a second opinion, and keeping score.

The second opinion

Ask one model a question, get an answer, and prudence suggests the obvious: ask another. A different product, a different company, a different color scheme. If both say the same thing, that is two witnesses, and confidence rightly rises. This is Galton's move — and it is licensed by Galton's condition. Two witnesses are two witnesses only if they err separately.

Do they? In 2025 a group at Cornell measured it, across more than three hundred and fifty language models, large and small, from many companies. The question was simple: when two models are both wrong on the same test item, how often are they wrong *in the same way* — the same wrong answer, not merely some wrong answer? On one widely used evaluation suite: about sixty percent of the time. Different engines; same fuel. The models are trained on records that overlap enormously — much of the written internet, the same reference works, the same benchmark-adjacent corpora, and increasingly each other's output — and models built by rivals inherit the same tendencies from the same shared record. Asking a second model is very often not calling a second witness. It is asking the first witness's archive a second time, through a different reader.

The study's sharpest finding is the one that inverts intuition: the *larger and more accurate* the models, the more their errors agreed — across different companies, different architectures. If errors were private noise, better machines should err more idiosyncratically. They err more identically. The technical companions to this essay argue that is exactly what faithful learning from a shared record must produce: the best students in the class miss the same questions the same way, because the questions they miss are the ones the syllabus itself gets wrong.

The scoreboard

Now to the fair's public face: the leaderboard. Dozens of models, one test, a ranked column of scores — the ox-guessing contest of the machine age, updated hourly. Two things about it deserve more suspicion than they get.

The first is what the scoreboard *measures*. It grades each answer right or wrong and tallies. Researchers at OpenAI and Georgia Tech observed in 2025 what this scoring does: a blank scores zero, a lucky guess scores one, so the test rewards the confident guesser over the honest doubter — and since models are tuned to climb these boards, the industry has been breeding for confident guessing. The fair, in other words, does not merely rank the contestants; it trains them.

The second is what the scoreboard *cannot see*. Its column of scores presents the contestants as so many independent fairgoers, and consumers of the board read agreement among high scorers as convergence on truth. But sixty percent identical errors means the contestants are not eight hundred fairgoers; they are closer to a handful of fairgoers wearing hundreds of coats. The board has no column for that. Nowhere in its record does it store the one number that governs how much a consensus of its contestants is worth — how *coupled* they are, and through

what. The fair posts every guess and forgets that the guessers all read the same almanac on the way in.

The refund, collected

Return to the two engines, because the stakes are not academic. Redundancy reasoning — multiply the failure probabilities — is built into how institutions consume machine judgment. A hospital system cross-checks one model’s read with another’s. A firm ensembles several models and takes the vote. A defense program puts the same perception stack on many platforms and counts on numbers. A regulator sees a market of many vendors and infers resilience. In every case the arithmetic is Galton’s, and the license is independence.

A shared model — or shared training data, shared components, shared upstream sources — is contaminated fuel. It does not make any single judgment worse. It makes the *combination* worth less than the arithmetic claims, and it does so invisibly, because the coupling lives between the records, where no institution’s own books can show it. Two hospitals that have never exchanged a byte can be wrong together, about the same patients, for the same upstream reason, while each one’s audit comes back clean. That is the technical companion’s central point, and its title is its thesis: what a deployment correlates. One instrument, written into everyone’s records, couples everyone’s errors — and today’s records have no place to write that fact down.

The question to ask

None of this says the crowd is finished, or the second opinion worthless, or the leaderboard a fraud. Galton’s arithmetic is not wrong; it is *conditional*, and the condition has quietly stopped holding in more places than we are accustomed to checking. The repair, as everywhere in this project, is not to trust less but to record more: agreement is evidence exactly in proportion to the independence behind it, and the independence behind it is a fact — causal, estimable, storable — that our records currently discard at the moment of writing.

So the question this essay would leave in the reader’s pocket, for the next time several confident machines agree: not “how many said so?” but “*how many engines am I actually flying on?*” The fair only works when you count the fairgoers right.

Sources and precision for the claims sketched here — the Cornell measurement of correlated errors (ICML 2025), the OpenAI / Georgia Tech analysis of grading incentives (2025), the aggregation-floor and floor-propagation propositions, and the century of prior mathematics on combining dependent judgments — are given in the technical papers at realql.org.