Episode Book

EPISODE #6189

What "Real" Improvement Actually Means in AI Evals

PUBLISHED	21 October 2026 at 17:23 UTC
DURATION	30:16
LISTEN	myweirdprompts.com/episode/chance-statistical-significance-ai-evals-28610/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.12
SOURCES	10 cited by the research (listed at the end)
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

A leaderboard shows a bolded number: one model beat another. But is that a real improvement or ordinary statistical wobble? This episode digs into what "chance" actually means in AI evaluations — the hypothetical worlds behind a null hypothesis, why repeating the same questions with fresh answers is a different experiment than fresh questions, and why the American Statistical Association had to publicly warn that a p-value does not measure the probability that a hypothesis is true. Along the way: random-guessing baselines, Cohen's kappa, noise in human judgment, and the quiet damage done by selecting the best of a hundred prompts. If you've ever bolded a decimal point and hoped, this one's for you.

TRANSCRIPT



Corn

...the bold type is doing rather a lot of work, Herman. Daniel wants to get underneath that moment when somebody calls an improvement real rather than ordinary statistical wobble. What does chance actually mean, especially in AI evaluations, and how do you work out its size? He wonders whether standard deviation gives you the answer, or whether there are general rules.



Herman

The first thing I would ask the person bolding the number is what they imagine repeating. The same questions with fresh answers? Fresh questions from the same subject? A new training run? Those produce different kinds of wobble, even when the score printed in the table is identical.



Corn

Your first response to a leaderboard is a cross-examination.



Herman

You were going to object to the font.



Corn

The font has overstated its evidence.



Herman

Chance becomes concrete when we specify a hypothetical world and a procedure. For a comparison, that world might contain two equally capable models. We then ask how much their measured scores could differ anyway. The answer concerns a statistic under that setup, not a mysterious quantity hiding somewhere inside the dataset.



Corn

That already changes how I hear “within normal variability.” Normal for which experiment? If I keep the questions nailed down, I have quietly removed one source of uncertainty.



Herman

You've removed it from the repetition, yes. Whether you should remove it depends on the claim. “This model answered these questions correctly” needs little inference. “This model is better at the subject” reaches beyond those questions. Now their selection matters.



Corn

I can establish that you beat me at the crossword without establishing intellectual supremacy.



Herman

Especially if you selected the crossword.



Corn

An entirely legitimate editorial responsibility.



Herman

Suppose our statistic is the difference in average accuracy. Under a null hypothesis of no performance difference, repeated evaluations still produce positive and negative gaps. Their distribution tells us what ordinary variation looks like under that hypothesis. An observed gap far out in its tail puts pressure on the null and the assumptions accompanying it.



Corn

Not every surprising result represents an improvement, though. A broken answer key could be surprising. A benchmark full of familiar questions could make a model look remarkably capable.



Herman

Statistical uncertainty doesn't absorb all possible mistakes. Bias, contamination, poor task coverage, and scoring errors can persist across every repetition. You could measure the wrong thing with extraordinary precision.



Corn

A very steady compass pointing at the fridge.



Herman

The practical distinction is between uncertainty we model through repetition and errors that require changing the design. Calling everything noise hides that distinction. Even the repetition needs care. If questions arrive in families, sampling individual questions as though they were unrelated invents more independent evidence than we possess.



Corn

Before you reach for a formula, I predict you're going to ask how many independent observations there really are.



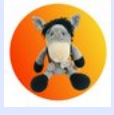
Herman

I was going to ask what the observations represent.



Corn

Worse. You've gone upstream.



Herman

You know my habits, Corn. But that's productive upstream travel. We need the target, the null, and the repetition before standard deviation can tell us anything useful.



Corn

Then “chance” is partly a modelling decision. Not arbitrary, but chosen in relation to the claim. I can't look at a column of scores and expect it to announce which hypothetical world I meant.



Herman

Take the conventional definition of statistical significance. A result is significant when a result at least as extreme would be sufficiently infrequent if the null hypothesis were true. “At least as extreme” matters. We count outcomes as incompatible with the null as the one observed, not merely the probability of reproducing our exact data.



Corn

Otherwise almost any detailed dataset would qualify as miraculous. The exact order of my mistakes is very distinctive.



Herman

A p-value summarizes that tail probability, conditional on the null and the statistical assumptions. If it is small, our observation would be unusual under that package. It does not tell us the probability that the package is true.



Corn

That's where the everyday reading goes wrong. People hear “five percent” and mentally attach it to “this was just luck.”



Herman

The American Statistical Association addressed that explicitly in its two thousand sixteen statement. “A p-value does not measure the probability that the studied hypothesis is true or the probability that the data were produced by random chance alone.” That warning is doing real work. The calculation assumes a hypothesis to assess possible data. People reverse it into a probability of the hypothesis after seeing data.



Corn

The calculation has answered its question. We've stolen it for another.



Herman

Consider a screening alarm. Knowing how often an alarm activates when nothing is wrong doesn't by itself tell you how often something is wrong when the alarm activates. You also need information about how common the underlying condition is and how the alarm behaves when it is present.



Corn

You spent years explaining that to worried parents. Did saying “conditional probability” ever improve the room?



Herman

Never as much as I hoped. Saying what the number actually counted helped. The same applies here. Instead of “a two percent chance this improvement is luck,” say, “under equal performance and these assumptions, gaps this extreme occur about two percent of the time.”



Corn

Longer, but considerably less fictional.



Herman

It also leaves room for alternatives. A small p-value could reflect a genuine effect, a wrong uncertainty model, dependence you ignored, selective reporting, or some combination. It doesn't identify the cause for you.



Corn

Nor does a large one establish equality. If I test two models on a handful of questions, my uncertainty could swallow a useful improvement.



Herman

What would you write beside that result, Corn?



Corn

“Inconclusive at this resolution.” I'd show the estimated gap and an interval. If the interval includes both a worthwhile gain and a worthwhile loss, I don't get to call the models equivalent merely because the test declined to settle it.



Herman

That's the power issue. A test with little ability to detect the effect you care about can return an unremarkable result even when that effect exists. Equivalence needs its own criterion, including a margin small enough to count as practically negligible.



Corn

“Not proven different” is a very cheap tie.



Herman

The usual five percent cutoff is a convention, historically associated with Fisher, not a natural boundary. An observation just either side of it hasn't crossed a physical phase transition. Yet institutions built publication and purchasing decisions around that crossing.



Corn

One decimal place gets tenure. The next goes home.



Herman

In two thousand nineteen, an editorial introducing a special issue of the association's journal urged moving beyond declarations of statistical significance and abandoning that threshold as a decision rule. I should narrow that... it was an influential recommendation, not every statistician voting to abolish hypothesis testing.



Corn

What replaces the stamp matters more than removing it. I still have to buy a model, ship a feature, or decide whether an experiment deserves another month.



Herman

Estimated effects, uncertainty intervals, practical costs, design quality, prior knowledge, and sometimes explicit decision models. Bayesian analysis can express uncertainty about hypotheses or parameters, but brings assumptions of its own. None of these removes judgment. They make its ingredients more visible.



Corn

I prefer an argument I can inspect to a threshold I can worship.



Herman

There's another meaning of chance that we should separate before it causes trouble. Four-option multiple choice has a random-guessing baseline of twenty-five percent, assuming one correct choice and uniform guessing. That's a level of expected performance.



Corn

Whereas the uncertainty is how tightly the observed score sits around that level. A random guesser can be reliably mediocre.



Herman

With enough independent questions, very reliably. Accounts of training evaluations include models remaining around that MMLU baseline even after roughly two hundred ten billion training tokens. The training quantity doesn't turn the baseline into a measure of variability. A score near chance and a noisy score are different descriptions.



Corn

An enormous education in guessing.



Herman

A model can also exceed a naive guessing baseline without demonstrating the skill we intended. Answer-position preferences or dataset regularities may help. The baseline must match the mechanism we're ruling out.



Corn

That comes up with judges, too. If two AI graders agree frequently, some agreement could follow from both mostly saying “acceptable.” Their raw agreement alone might flatter them.



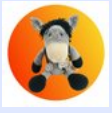
Herman

Cohen's kappa adjusts agreement against a baseline derived from the graders' marginal label frequencies. But that baseline has assumptions, and kappa can behave awkwardly when one label dominates. “Corrected for chance” should make us ask which chance model was used, not relax immediately.



Corn

Even the correction needs a label on the tin.



Herman

There's a broader use of noise in human judgment. Kahneman, Sibony, and Sunstein describe it as undesirable variability in judgments of the same problem. Different judges, or the same judge at different times, may give different answers where consistency would be valuable.



Corn

That helps me distinguish variation we want to study from variation we want to suppress. If graders disagree because the question is ambiguous, averaging their scores may conceal a defect in the question.



Herman

Or reveal a legitimate range of interpretations. The point is to examine it rather than automatically demand one number. In AI evaluation, a supposedly objective score can contain randomness in question selection, model generation, tool behavior, and the grader's own generation.



Corn

Four opportunities to put a decimal on uncertainty.



Herman

Repeatedly asking the grader about one fixed answer estimates one component. It doesn't tell you how representative your questions are. Repeating the entire pipeline estimates more components, but still only under the environment and sampling scheme you actually repeated.



Corn

I think that's why Daniel's standard-deviation instinct is useful but incomplete. Spread matters. Yet first we have to say the spread of what. Individual answers, averages, differences between averages, or a decision made after searching a hundred variants.



Herman

That last one changes the null distribution substantially. If you try many prompts and report the best, the winner has benefited from selection even when none is superior. Fresh evaluation often brings its score back toward the pack.



Corn

Regression to the mean. The victory lap has a return leg.



Herman

The search itself belongs in the analysis. Preselect the comparison, adjust for multiplicity where appropriate, or reserve untouched evaluation data. Repeatedly checking until something crosses your preferred threshold also changes how often you get a false alarm.



Corn

A p-value doesn't remember all the unsuccessful runs you omitted.



Herman

Now put that into a model leaderboard. One model beats another on the printed questions. Are we seeing a capability difference, or good fortune in the questions selected? Anthropic frames its statistical evaluation guidance around exactly that distinction between observed averages and performance across the intended population of possible questions.



Corn

The bold number answers “who won this sheet.” Buyers hear “who will win my workload.”



Herman

Traditional question-answer evaluations often operate with a highest-number-wins mentality, while uncertainty estimates go unreported. A slightly higher score may deserve investigation without deserving a crown.



Corn

Bold is cheaper than another evaluation run.



Herman

To compute uncertainty, separate question variability from answer variability. Data noise comes from which questions enter the sample and how difficulty varies across them. Prediction noise comes from a model producing different answers to the same question.



Corn

Repeating an answer can reduce the second. It cannot introduce the kind of question I forgot to include.



Herman

Data variance is sometimes described as irreducible in that decomposition. More precisely, repeating predictions on fixed questions cannot reduce it. Sampling more independent questions can reduce uncertainty about the population mean.



Corn

Good. “Irreducible” was about to acquire a mystical reputation.



Herman

For independent, similarly sampled scores with finite variance, the standard error of their mean is the standard deviation divided by the square root of the number of scores. That's the direct answer to Daniel's standard-deviation question. The spread of individual observations enters, but sample size changes the uncertainty of the average.



Corn

Four times as many independent questions halves that standard error. It doesn't halve the diversity of individual questions. We haven't made the world tidier, only estimated its average more precisely.



Herman

For correct-or-incorrect scoring, use the square root of accuracy times one minus accuracy, divided by sample size. Accuracy is expressed as a fraction there. At fifty percent accuracy and a hundred independent questions, the standard error is about five percentage points.



Corn

A one-point lead has arrived wearing very large shoes.



Herman

Under a reasonable normal approximation, a ninety-five percent confidence interval extends roughly one point nine six standard errors either side of the estimate. Here that means nearly ten percentage points each way. That is an interval for one accuracy estimate, not automatically the interval for the difference between two models.



Corn

Herman, before you decorate the whole table with intervals, paired observations should help. The models face the same questions. If both fail on the awful ones, some difficulty variation cancels when we subtract their scores question by question.



Herman

The comparison should exploit that. Frontier models' question scores have been reported with correlations from about zero point three to zero point seven. Paired differences can be much less variable than treating the two averages as unrelated.



Corn

For once, sharing the exam helps everybody.



Herman

At the same sample size, accuracy near ninety-five percent has lower Bernoulli variance than accuracy near fifty percent. But… fewer errors can also make a normal approximation less trustworthy. Lower variance doesn't grant immunity from small-sample problems.



Corn

Nor does it make the benchmark useful if every model passes. I'd rather measure the errors that matter than minimize an error bar by asking whether water is wet.



Herman

Hold that thought. And now, as promised, we have Dorothy on the line for her regular segment, Dorothy's Kitchen. Hello Dorothy.

D

Dorothy

Hello dear, it's your mother. I'm here for Dorothy's Kitchen.



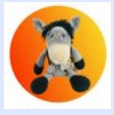
Corn

Hello, Mum.

D

Dorothy

Today I want to share my recipe for a glass of tap water.



Herman

I'm ready.

D

Dorothy

You will need one clean glass, and water from the tap. That is all. It sounds simple, but it is not.

D

Dorothy

First, hold the glass up to the light. If it is cloudy, put it back. I won't say why. Then fill it slowly, watching the water. If you rush, you will spill it, and things have gone wrong before.

D

Dorothy

Let it sit for a moment. The water must settle. Then drink it. Do not sip. Drink.

D

Dorothy

There. That is a glass of water. Now eat something, you look thin. I have to go, my program is on. Goodbye, dear.



Herman

Benchmark usefulness and precision can pull in different directions. How would you spend a limited evaluation budget, Corn? More questions, or more answers per question?



Corn

I'd repeat a small representative subset first to find out how unstable the answers are. If generations disagree constantly, repetitions buy information. If answers hardly change, I'd spend on broader coverage. Your instinct will be to build a perfectly aligned apparatus before admitting the target is too narrow.



Herman

Archery has rewarded that instinct.



Corn

The target generally stays in the same place.



Herman

Fair criticism. Though even broad coverage needs honest dependence accounting. Reading-comprehension benchmarks can ask several questions about one passage. Those questions share vocabulary, facts, and difficulty. Counting each as independent can make uncertainty estimates much too small.



Corn

Ten questions about the same misunderstood paragraph may represent one failure with ten invoices.



Herman

Clustered standard errors on popular evaluations can exceed naive ones by more than a factor of three. Preserve passage groups in the calculation or resampling. The effective amount of information depends on the dependence, not merely the row count.



Corn

This is where a shuffle test appeals to me. Keep the real question-level structure, then generate a world where neither model has a systematic advantage. Compare our observed gap with gaps from those rearrangements.



Herman

A permutation or randomization test can do that under an appropriate exchangeability assumption. For paired scores, one construction swaps model labels within pairs under a suitable null. But don't swap individual questions independently when the randomization unit should be a passage or another cluster.



Corn

Assumption-light still has assumptions.



Herman

Yes. Simulation is a way to make the hypothetical world explicit, not a way to avoid specifying one. Bootstrap resampling is another tool, usually aimed at estimating sampling uncertainty rather than directly imposing a null. Again, resample the independent units.



Corn

I like being able to inspect the procedure. If the simulated world permits combinations my actual evaluation could never produce, the resulting chance band answers the wrong question.



Herman

Tiny benchmarks add another problem. An AIME exam has fifteen problems. In the cited small-benchmark discussion, FrontierMath had around three hundred, with some categories containing fewer than three examples. One additional solved problem can have considerable leverage.



Corn

Fifteen questions. Every mistake gets a speaking role.



Herman

The central limit theorem doesn't promise that any average you calculate already has a well-approximated normal distribution. Sparse successes, unequal weights, dependence, or heavy-tailed scores can all make approximation difficult.



Corn

Yet there is a genuine methodological disagreement here, rather than a universal instruction to abandon standard errors.



Herman

One camp recommends normal-approximation standard errors as a practical baseline, with clustering and pairing done properly. Another argues those intervals often understate uncertainty on small specialized evaluations and favors alternatives, including Bayesian models. Exact methods can help in suitable discrete settings. Method choice depends on the statistic and design.



Corn

I'd distrust anybody who claims Bayesian on the label automatically fixes fifteen unrepresentative questions.



Herman

Priors and likelihoods need scrutiny too. No interval can conjure unobserved task families. A precise estimate of performance on competition mathematics doesn't establish performance across every kind of reasoning.



Corn

The scope of the claim needs an error bar of its own.



Herman

We can still give Daniel useful guidelines. Define the target and null. Estimate the sampling distribution analytically when the assumptions fit, or simulate a justified null process. Use paired comparisons, preserve clusters, and check whether the evaluation can detect a practically important gap.



Corn

That's power analysis before the expensive run. Choose the improvement worth noticing first, rather than declaring whatever improvement appeared to be important afterward.



Herman

Reduce within-question randomness where it helps. Repeated generations can estimate expected success under a specified sampling policy. For some multiple-choice setups, answer probabilities provide smoother measurements than a single hard choice.



Corn

Though improving the probability assigned to the right answer needn't improve the fraction of tasks actually completed. A smoother metric can quietly change what we're rewarding.



Herman

It must remain relevant to the application. Continuous measures have shown substantially higher signal-to-noise ratios in some benchmark comparisons, but lower measurement noise alone doesn't establish validity.



Corn

What about the proposal to characterize noise across many model pairs? That sounds more economical than treating every leaderboard row as a fresh statistical universe.



Herman

All-pairs analysis has found characteristic, fairly predictable noise levels within many evaluations. That can support reusable estimates instead of bespoke testing for every comparison. It still depends on the models, evaluation protocol, and assumptions staying sufficiently similar.



Corn

A useful calibration, not a perpetual warranty.



Herman

Improvements combining pairing and reduced prediction noise have cut noise severalfold in published examples. The gain is real, but the factors aren't universal constants. Would a purchasing team actually use a calibrated noise band, Corn, or keep sorting by the highest score?



Hilbert

...provided the band is printed beside the price, yes. My neighbour Bernard wants one for the laundromat accounts. He has already ordered a brass pointer. That is the consequence of explaining uncertainty to a man who makes ceremonial weather instruments. He now expects every uncertain quantity to have a needle. We are fitting his first indoor takings barometer on Friday. It will occupy the change machine alcove. Customers will have to receive coins through the mouth of a painted heron.



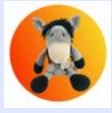
Corn

The heron should have a queue limit.



Hilbert

Bernard has allowed for six customers and one inspector. The inspector is a plywood outline. It reserves the space. He dislikes being surprised by official shoulders. The accounts come from a notebook I inherited from my uncle, who ran the laundromats and changed detergent whenever weekly takings fell. He called every dip significant. Forty years of weekly figures. No record of the weather. Excellent records of which spoon he used.



Herman

Detergent changes triggered by low takings would make the comparison awkward before you even decoded it.



Hilbert

I have decoded fourteen notebooks' worth. Three years on paper. The original uses his own code. A triangle means pounds unless it touches a swan, in which case it means the boiler was sulking. Bernard has commissioned a swan stamp for Friday. I told him we are measuring money, not extending the alphabet. He has ordered it anyway. It cost eighteen pounds, including a little velvet nest.



Corn

The accounts now have habitat requirements.



Hilbert

The weekly takings have a standard deviation of forty-seven pounds and thirty pence. Exactly, in my calculation. Bernard wants that engraved across the dial. I have given him the figure twice. Forty-seven pounds and thirty pence. He says it sounds like a deposit. He has put down that amount on a second heron, which is not part of my plan.



Herman

That spread doesn't establish whether detergent affected revenue. We'd still need the comparison and a defensible account of what changed together.



Hilbert

I accept that. The pointer is not a detergent verdict. It indicates takings. Bernard insists the uncertainty must move separately, so there will be a smaller needle operated by his downstairs neighbour through a bicycle brake cable. She has agreed to work Thursdays. We open Friday. I have put a grey cloth over that part.



Corn

A staffed uncertainty indicator. Coverage could become expensive.



Hilbert

Cut that little cable knock from my microphone, Herman. The cable for the dial is longer. Bernard has bought twelve metres because he wants room for promotion. I had a spreadsheet licence. I returned it. The paper stays flat on the table, and the broad ruler is easier with anteater claws. Friday's demonstration will use thirty-seven sealed envelopes. Each contains a different weekly taking and a drawing of the relevant spoon.



Herman

Randomizing the order at least prevents you selecting only the weeks that make the pointer behave nicely.



Hilbert

Bernard has appointed a tortoise to choose the envelopes. It belongs to the balcony association. They require a written guarantee that it will not be described as slow. Corn, you distinguished the spread of answers from the spread of averages earlier. I have used both phrases on the programme. Bernard has translated them into “large needle” and “other needle.” The programme is six pages.



Corn

Keep the tortoise off the timetable.



Hilbert

It has a substitute. A second tortoise, with a union card made by Bernard. The first one dislikes blue envelopes. I have changed the envelopes to cream. Bernard will sit beside the machine in his good cardigan. I've put a footrest under his chair. His ankles swell when he supervises a launch.



Herman

If customers buy extra washes to see the heron dispense coins, the demonstration itself could alter the takings.



Hilbert

Then I'll mark Friday separately.



Corn

A real evaluation would have to mark that kind of intervention separately too. Freeze the conditions where possible, record changes where not, and resist treating every new score as directly comparable.



Herman

Particularly when the evaluation has become part of development. Once teams optimize against it repeatedly, the uncertainty calculation for an untouched sample no longer describes the whole selection process.



Corn

The misconception, then. “The p-value tells me the probability this improvement happened by chance.”



Herman

It tells you how unusual results at least this extreme would be under the specified null and assumptions. Chance uncertainty belongs to that setup, including the statistic, sample size, and dependence. Standard deviation alone cannot supply it.



Corn

I'd like AI evaluations to make changed conditions as visible as changed scores. Otherwise better error bars will merely give us more precise arguments about comparisons we should never have made.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you want more of this, try episode nine, Benchmarking Custom ASR Tools - Beyond The WER; episode eleven, How Does Fine Tuning Work Anyway; and episode nineteen, AI Images. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy these conversations, leave us a review wherever you listen.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Anthropic [primary]**
   https://www.anthropic.com/research/statistical-approach-to-model-evals

2. **Adding Error Bars to Evals, arXiv [primary]**
   https://arxiv.org/abs/2411.00640

3. **Measuring all the noises of LLM Evals, arXiv [primary]**
   https://arxiv.org/abs/2512.21326

4. **Don't Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints, arXiv [primary]**
   https://arxiv.org/abs/2503.01747

5. **Quantifying Variance in Evaluation Benchmarks, arXiv [primary]**
   https://arxiv.org/abs/2406.10229

6. **Wikipedia, last edited 2026-10-17**
   https://en.wikipedia.org/wiki/Statistical_significance

7. **Wikipedia, last edited 2026-09-13**
   https://en.wikipedia.org/wiki/Standard_deviation

8. **Wikipedia, last edited 2026-10-13**
   https://en.wikipedia.org/wiki/Noise:_A_Flaw_in_Human_Judgment

9. **ASA 2016 Statement on P-Values**
   https://orthotw.com/2016/05/american-statistical-association-asa-issues-p-value-warning

10. **Moving to a World Beyond p 0.05, ASA 2019**
    https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913