Episode Book

EPISODE #6135

How Many Runs Does Your AI Eval Actually Need?

PUBLISHED	19 October 2026 at 15:04 UTC
DURATION	23:39
LISTEN	myweirdprompts.com/episode/how-many-eval-runs-28011/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	20 October 2026 at 07:46 UTC

SYNOPSIS

You rerun your model panel and last week's winner slides into the middle of the pack. Was it luck, or was it the prompts? This episode works through the arithmetic of eval noise: how run-to-run variance shrinks with every added run, why prompt-choice variance doesn't budge no matter how many times you rerun, and why three runs is a defensible default for some measurements and badly underpowered for others. There's also a detour into why temperature zero isn't the deterministic shortcut people assume, why a second judge from a different model family does a job that reruns can't, and why clustered prompts quietly shrink your effective sample size.

TRANSCRIPT



Corn

One run puts a model at the top of the table. Daniel reruns the panel and that same model is down in the pack, looking like it never earned the spot.



Herman

Anyone who has built a benchmark meets that moment sooner or later. Usually when the second invoice arrives.



Corn

This is one more in our run of episodes about judging the writers of this show, and it's the budget one. A big lab can run every model a hundred times and shrug. A small outfit pays the whole API bill again each time it reruns the panel. So Daniel wants to know whether enough is three runs, five, or ten, and he suspects the answer moves with what you're measuring. His two test cases are whether a model refuses on Chinese-censorship probes, and how candid its dialogue is.



Herman

I can give a real answer for both, but it starts by loosening the question. Three, five or ten is the natural thing to ask. It's also half the picture, because runs are one of two places the noise comes from.



Corn

My guess for the other one is the prompts. Which ones you happened to pick.



Herman

Yes. Any eval is a sample from the enormous pile of things you could have asked. Pick a different hundred and eighty-nine prompts and every model's average shifts, even if the models were perfectly deterministic. Call that question noise. Separately, ask the same prompt twice at temperature one point zero and you get two different answers, and the judge may score them differently. Call that run noise.



Corn

And more runs only shrink the second one.



Herman

Only the second. Rerun the same prompts a thousand times and you get a beautifully precise measurement of how the model does on those particular prompts. Whether that carries over to the next batch is a prompt question, and no amount of rerunning touches it.



Corn

How big is the prompt part, in practice? I'd have guessed small, with a hundred and eighty-nine of them.



Herman

Bigger than people expect. At ninety percent accuracy, two hundred items gets you roughly two points of wobble either way. Thirty items, roughly five. That's before a single rerun enters the picture, and it's why a table of close models is mostly a table of prompts.



Corn

Two points. So the whole middle of any leaderboard is a tie nobody admits to.



Herman

In Anthropic's error-bars paper the variance splits exactly that way. The run part gets divided by the number of runs. The prompt part doesn't move.



Corn

So there's a floor.



Herman

There is, and you reach most of the gain early. Their worked example has every question equally hard and answers that are simply right or wrong, so treat it as a shape, not a law. In it, two runs takes a third off the variance.



Corn

Decent for double the bill.



Herman

Four runs takes it to a half. After that it creeps, and even infinite runs can only reach two thirds. The bulk of the benefit is in by about four.



Corn

Ten runs of the bill to go from a half to three fifths.



Herman

That's what the arithmetic says. Past four or five runs per prompt, another prompt beats another run almost every time. Which makes three, five or ten partly the wrong dial, and that's no knock on the instinct behind it. Runs are the dial that's easy to turn. Prompts are the dial that works.



Corn

Hang on, though. He's seen a model win and then tank. If that's mostly prompts, extra runs wouldn't have saved him.



Herman

Possibly not, and it's worth separating out. There are two ways a winner tanks. One is the winner's curse. Put ten close models in a table and the one on top is partly there because the noise liked it that day. We went through that in the episode on funnelling candidates down to a shortlist, and the rerun is where the luck gets taken back. The other is that the model was good on those prompts and merely ordinary on others.



Corn

I'd guess you can tell them apart by what you rerun.



Herman

That's my reasoning, not a published recipe, but yes. Rerun the same prompts. If the model slides, it was run noise. If it holds there and slides on fresh prompts, it was the prompts. Corn, would you do that before or after you'd told anyone the model won?



Corn

Before. Obviously before. Though I'll admit that when a table puts my favourite on top I'm slower to go looking for the rerun.



Herman

Everyone is. That's the winner's curse working on the person.



Corn

Why not skip all of it and drop the temperature to zero? One run, no noise, done.



Herman

Two problems. Daniel generates at one point zero, so zero would measure a different setup from the one he's testing. And zero isn't deterministic anyway. Someone took a big Qwen model, asked it one prompt a thousand times at temperature zero, and got eighty different outputs. The most common one showed up seventy-eight times.



Corn

So zero is more of a mood.



Herman

The cause is that the batch size shifts with server load, and that changes the arithmetic slightly, and a tiny difference early in a long answer snowballs. Lowering temperature can also move the average, or push the variance over into the prompt term. You don't get a free single run.

L

Larry

Are you tired of your results tanking? Introducing the Rerun Box, the monthly subscription that delivers fresh, confidence-rich epochs straight to your door. Every box contains only the runs that went well. We simply never send the others. Choose from three tiers. Single Run, for the casual sceptic. Triple Run, our most popular, with a laminated interval. And Centurion, one hundred runs, hand-polished, shipped in a velvet-lined crate by a man called Dennis. Our temperature is always one point zero, never lukewarm. Cancel anytime, provided you do it in person, in the original packaging, between four and four fifteen on the second Tuesday of a month that has five Tuesdays, and you answer a questionnaire about why you hate us. Unused runs are not refundable, and will be returned to you, angrier. Rerun Box. Be the model your mother thinks you are.



Corn

Does it come in a smaller size?



Herman

It ships only the good runs, Corn. You'd be paying monthly for a winner's curse.



Corn

Fair. Back to candour, then. The decision run is a hundred and eighty-nine prompts, three runs each, with a second judge from a different family.



Herman

This is the continuous case, so the numbers need an assumption, and I'm going to say that word a lot. I've guessed the standard deviations. Daniel has the real ones sitting in his three-run data, and I'd love to see them. Stretch the one to five score across zero to a hundred, and look at the gap between two models on a single prompt. Assume that gap wobbles from prompt to prompt with a spread of twelve points. Assume the run-to-run wobble inside one model on one prompt is eight.



Corn

That's three assumes in a row.



Herman

Four, and I'll stand by all of them. Now, with a hundred and eighty-nine prompts, how big a gap between two models can you reliably detect, with eighty percent power? One run each, about three and a third points. Three runs, a little under three. Ten runs, two and a half.



Corn

So ten runs buys about a quarter of a point over three.



Herman

And to put that on the rubric, stretching one to five across zero to a hundred makes five points a fifth of a rubric step. A quarter of a point on that scale is nothing a human reader could ever see in the scripts.



Corn

Three times the generations for something nobody can hear. I'd want a refund on that.



Herman

Now the part that surprised me. Double the prompts to three hundred and seventy-eight, one run each, and you land just under two and a half. For a panel of ten models, that's about thirty-eight hundred generations against fifty-seven hundred for the three-run plan, and it's a sharper instrument.



Corn

The catch is that prompts don't write themselves. Three hundred seventy-eight good candour prompts is a lot of Daniel's evenings.



Herman

It is, and that's a fair cost. When prompts are the scarce thing, runs are the cheap way to buy precision, so three can be perfectly sensible. But it shows where the next pound should go. And I can hear you about to say that three runs is a flat tax.



Corn

I'd say three is a flat tax on the numbers you assumed. Two runs, and spend the rest on a bigger prompt set.



Herman

Try it with a different guess. Keep the prompt wobble at twelve and say the run wobble is twenty, because candour scores from a judge can be jumpy. One run, the detectable gap is over six points. Three runs, about four. Ten, three.



Corn

Hm. That changes my vote.



Herman

It changes everyone's, which is why the number of runs can't be looked up. It depends on the ratio between those two wobbles. Daniel can measure it. Run two passes over all the prompts, split the variance into between-prompt and within-prompt, and he knows.



Corn

My version of a rule for that. Run until the run noise is about half the size of the prompt noise, and stop at five regardless.



Herman

A sensible rule of thumb, and it's yours, not from a paper. With my first guesses it says two runs. With the jumpy guess it hits the cap of five. Daniel's three sits between, so I'd call it not wrong, just unproven until he measures.



Corn

I'd have thought the second judge makes the runs redundant. Two judges, two looks at everything.



Herman

Different job. Scoring the same answer twice with the same judge trims judge noise. A second judge family catches judge bias, and bias doesn't average out however many times you ask. One study had two judge models agreeing on only about three quarters of verdicts. So the second judge is doing something runs cannot do, and I'd keep it before I kept the third run. And repeating the same judge isn't cheap either. In one study it took around eleven repeated calls on the same answer before a majority vote settled where fifty calls would have landed, and about fifteen on the jumpy items. That's a single dataset, so take it as the order of magnitude.



Corn

Free is my favourite price, but that one isn't free.



Herman

You'd pay it. You pay for leaf medicine, and that comes with no interval at all.



Corn

It has ancestors.



Herman

There's another trap hiding in the count of a hundred and eighty-nine. If several prompts come from the same source or topic, they don't count as separate pieces of evidence. On popular benchmarks, the error bar that accounts for that clustering came out more than three times the naive one.



Corn

So if Daniel's candour prompts come in families, the real number is smaller than the file says.



Herman

Possibly much smaller. That's a thing to check in the pilot too.



Corn

The funnel is where I'd expect runs to be least useful. The screen is three scripts per model, the lite stage is forty prompts.



Herman

Same formula, and the stages just have different resolutions. With my guessed wobbles, three scripts can only see a gap around twenty-seven points. Forty prompts, about seven. The full run, under three.



Corn

Twenty-seven points is a smoke alarm, not a ruler.



Herman

Fine, provided nobody reads a ranking off the smoke alarm. The screen's job is to catch the disaster. Corn, a question. When a model wins the lite stage by three points, do you advance it, or only because the shortlist has room?



Corn

Because it has room. Three points at forty prompts is inside the noise, and I'd say so out loud. The next stage is where it gets tested anyway.



Herman

Then the funnel is allowed to be generous. It isn't allowed to claim anything.



Corn

Now the censorship axis. It doesn't feel like the same animal at all.



Herman

It isn't, because refusal is driven by the prompt. A probe either trips the model's refusal or it doesn't, and rerunning the same probe mostly returns the same answer. So repeats aren't independent evidence. One pass over many different probes beats many passes over a few.



Corn

Which is what Daniel's gate does. About a hundred and fifty-five short answers, plus about eighty-four perspective answers.



Herman

Here's the arithmetic that matters. If you see zero refusals in some number of independent trials, the ninety-five percent upper bound on the true rate is about three divided by that number. That's the rule of three, and for small counts the exact version is a bit tighter. A hundred and fifty-five clean answers gets you under two percent. All two hundred thirty-nine gets you about one point two.



Corn

That's still not never.



Herman

Never is expensive. To claim below five percent you need fifty-nine clean trials. Below one percent, two hundred and ninety-nine. Below one in a thousand takes nearly three thousand.



Corn

I'll clear my calendar for the three thousand.



Herman

And look at what repeating buys you. Take eighty probes and run each three times. That's not two hundred and forty trials, because the repeats of a prompt-driven refusal aren't independent. It's closer to eighty trials, which gets you a bound near three and two thirds percent. Eighty distinct probes run once get you the same, for a third of the price.



Corn

So the three runs in Daniel's decision setup do nothing for the gate.



Herman

Nothing for a refusal that's the prompt's doing. Where they matter is the flaky probe, the one where a model sometimes refuses and sometimes doesn't. That's a small set, and you find it by running everything once and rerunning what looks odd.



Corn

Mind you, a refusal on a short factual question and a refusal in a long perspective answer look different. The long ones can be a polite non-answer, a shrug dressed as balance.



Herman

They can, and that's the hard part to score.



Corn

The perspective answers are long and hedgy, and a judge has to decide whether a shrug counts as a refusal.



Herman

Those carry the judge's noise more than the short ones, I'd guess, because the verdict is a judgement call and not a keyword. Same remedy, a second judge, plus a human reading anything flagged.



Corn

Here's where I get off the bus, though. Daniel's rule is one refusal means a technical fail. One in a hundred and fifty-five is about two thirds of one percent. The interval on that runs up to roughly three percent. Failing a model for that looks like failing it on luck.



Herman

Disagree. It isn't an estimate, it's a gate. The requirement is that the model doesn't refuse on content the show covers, and one refusal is a counterexample. The arithmetic argues for the strictness. If a model truly refuses one time in a hundred, a gate this size catches it about four times in five. If it's one in a thousand, it catches it roughly one time in seven. The gate is already a sieve. Loosening it lets more through.



Corn

Unless the judge mislabelled it. The judge adds its own noise, and one false positive ends a model's season.



Herman

Then read the refusal. That's where the runs go. Take the single flagged prompt, rerun it five times, and have a human read what comes back. You don't spend runs on the hundred and fifty-four that passed.



Corn

Fine. And if it refuses once in five reruns?



Herman

You tell me. What would you do?



Corn

Fail it. A model that refuses one time in five on that prompt will do it in front of a listener. I'd still say a first-pass single refusal shouldn't fail anyone automatically. I'll concede the fact and keep the opinion.



Herman

Keep it. The reading step makes the rule survive both our opinions.



Corn

Everything so far has really been two different questions wearing the same coat. The gate asks whether a model clears a line. The candour table asks whether neighbours are actually apart.



Herman

Different sums, too. A gate wants the interval to sit clear of the threshold, and the moment it does you can stop spending. A ranking wants the gap between adjacent models bigger than the smallest gap you can detect, and with ten close models there are a lot of adjacent pairs.



Corn

Then I wouldn't try to rank ten. I'd try to find out whether the top two are really different, and let the rest be a pack.



Herman

That's cheaper and it's usually the decision anyway. Nobody ships the fourth-placed model because it was fourth.



Corn

What would I do on Monday with a small budget, then. For a gate, one pass over as many distinct probes as I can afford, rerun only what gets flagged. For a ranking, find out the ratio of the two wobbles first.



Herman

Use the same prompts for every model. When two models are scored on the same prompts, the hard ones are hard for both, and the difference between them is quieter than either score alone. The correlation between frontier models question by question ran somewhere between three tenths and seven tenths in published work, so you get that variance reduction free.



Corn

Free again.



Herman

Then rerun only the prompts where the runs disagree. Judge scores are jumpy on some items and rock steady on others, so uniform reruns spend money on prompts that were never in doubt. And stop early when the interval clears the line. One caveat, from preprints I'm not sure are settled. If you peek and stop, you need a method built for peeking, because a fixed interval loses its coverage once the data decides when you stop. One such paper trimmed more than half the planned trials on a two hundred item eval, with the same conclusions.



Corn

Herman, who actually pays when the third run comes back looking just like the second?



Hilbert

...me, last time. Thirty-four pounds twenty to Ballard Quarry for three crossings of the weighbridge, one hen. She got out at the lights after the second, so the third was the hearse on its own, and they charged me full rate for it. Mrs Grady has not sent a penny of it back. She says a hen is not a hired vehicle.



Herman

I'd expect a weighbridge to read in big steps, so a hen on her own would barely register.



Hilbert

Which is why Thursday it's the hearse with a hen aboard. Desmond Pike runs the bridge weekdays and the funeral side on Saturdays. The hearse has been on my drive since the ninth. Four of the hens live in the back and the heat lamp's on the cooker socket.



Corn

Nobody is making toast until Thursday.



Hilbert

Nobody is making toast until Friday. Mrs Grady has nine hens and every one is named after a member of the Tanner's Row finance committee. Councillor Birtles is the one that bites. She wants a number for each before the committee sits, and she wants it in ink. I've bought nine stamp pads from Ratcliffe's, two pounds sixty each, and the hen walks across the page. Desmond will only take eggs as payment from the hen called Hargreave. The other eight he's not speaking to. Herman, you said assumed so often I put a marker on the waveform.



Herman

Fair enough. Then you'd run the same hearse for every hen. A fixed vehicle weighs the same each crossing, so it drops out when you compare one hen to the next.



Hilbert

Same hearse, same stone. There's a headstone in the back for Mrs Grady's late husband, no wording on it yet, so the quarry bills the pair of them as one vehicle. Eleven pounds forty a crossing, whichever hen. Last time they wrote it up as eighteen tonnes and I had to ring a man in Doncaster who keeps the rate card in his garage.



Corn

The hens will sit on the stone. It's flat, and it's warm off the lamp.



Hilbert

Councillor Birtles already does. Wednesday the mason collects it, so Thursday I weigh the hearse empty as well. That's ten crossings, not nine. I hadn't counted the empty one. Desmond wants it weighed twice because he says the bridge reads different when it's lonely.



Herman

Once is enough for the empty one. Nine crossings with a hen and one without, and the hearse is common to all ten, so you subtract it out. A second empty crossing only re-measures the hearse.



Hilbert

Once, then. Desmond won't like it. He isn't on the invoice, and he has already told Mrs Grady that the committee can have the empty hearse's figure as well, in ink, at no charge, so now she wants it framed for the hall.



Corn

Ten crossings at eleven pounds forty is a hundred and fourteen pounds, and Mrs Grady's not paying.



Herman

The per-crossing price is the number to write down first. Once you know what one run costs, how many is just arithmetic against a budget. For Daniel it's prompts times models times runs times the cost of generating and judging each one, plus the second judge on top.



Corn

I'd put that number on the wall above the desk.



Herman

One detail from the reliability work on agent benchmarks that didn't fit anywhere. A benchmark called tau-bench scores a model two ways. Pass at k counts it if any of k tries succeeds. Pass to the k counts it only if all k tries succeed. Models that solved under half the retail tasks on one try scored under a quarter when eight clean tries in a row were required.



Corn

A gate like Daniel's is the second kind of scoring.



Herman

What I'd write down first is the price of one run, and the ratio between the prompt noise and the run noise. Then Daniel's pilot tells him how many runs, and whatever that saves goes into more prompts and the second judge. The next thing worth watching is whether anyone publishes those ratios for judged dialogue at all.



Corn

Thanks to our producer, Hilbert Flumingtop, for getting this one to air.



Herman

This has been My Weird Prompts.



Corn

Find more episodes at my weird prompts dot com.



Herman

See you tomorrow.



Corn

Don't wait up.