Episode Book

EPISODE #6132

How to Run a Model Eval on a Budget

PUBLISHED	19 October 2026 at 14:39 UTC
DURATION	22:47
LISTEN	myweirdprompts.com/episode/stepwise-model-evaluation-budget-42679/
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

When every run costs real money, you can't evaluate a dozen candidate models in one giant sweep. This episode walks through the stepwise evaluation Daniel built to pick a script writer for this very show — a screening funnel that starts cheap, cuts the field, and only spends real budget on the finalists. Along the way: what successive halving and racing actually are, why paired per-prompt comparisons beat raw scores on a forty-prompt set, how winner's curse inflates your top pick, and why a model that engages with Tiananmen and a hosting provider that truncates it are two completely different findings. Plus a pharmaceutical ad for a funnel in capsule form.

TRANSCRIPT



Corn

...which is why a score built on three samples makes me nervous. It can tell you a model is broken. It can't tell you a model is good.



Herman

You're skipping ahead, Corn. The actual question hasn't even been asked yet.



Corn

Fair. This is the last of four we've done on judging the writers for this show, and it's the practical end. Daniel is teaching himself to run evaluations, and the script writer for this very show is his worked example. Lots of candidate models, every run costs real money, so he never does one giant run. He starts cheap, with a judge or an agent checking technical basics, cuts the field, sometimes follows a good vendor into its other models, and once three to five are left he runs each one several times to see if the result repeats. Long list down to one. He calls it a stepwise evaluation. Is that a recognised method, what's it called, and what should it look like for one person with a budget?



Herman

Recognised, yes. He's reinvented a good chunk of it, which is a fine way to learn. The naming is a mess, though, because there isn't one canonical name. In plain English it's a funnel, or a screening funnel. Drug discovery uses that word, so does hiring.



Corn

Hiring funnel. So every model is now a candidate with a cover letter.



Herman

Pretty much, and the cover letter is the first prompt. The formal names are successive halving and racing. Successive halving is simple. Give every candidate a small budget, drop the weaker fraction, spend what you saved on the survivors, repeat. Racing is the more careful cousin. You run candidates on test cases one at a time and drop one only when a statistical test says it's behind, so you eliminate on evidence instead of on a fixed cut.



Corn

And he's doing the first one by instinct.



Herman

With a bit of the second, I'd say. There's a useful split in the theory too. Fixed budget means you have this many dollars and you want the best answer they can buy. Fixed confidence means you keep sampling until you're ninety-five percent sure. Daniel is plainly fixed budget, and successive halving is the standard algorithm for that.



Corn

What's the catch? There's always a catch with the tidy algorithm.



Herman

The catch is that an early, cheap round can mislead. A model that needs more room to show what it's good at gets cut by a round that was too small to see it. The Hyperband method exists for exactly that. It runs successive halving several times with different starting budgets, some generous, some stingy, so one bad early round can't kill a good candidate by itself.



Corn

So the lesson is don't swing the axe hardest where you can see the least.



Herman

Yes, and that's going to come back at every stage. Corn, can I ask you something? Say a model fails one of the three screen scripts. Do you cut it?



Corn

Depends how it failed. If it wrote a mediocre script, no. If it fell over, probably yes.



Herman

That's the whole distinction. The first stage is a gate, not a ranking. Pass or fail on things a piece of code can check. Did it refuse, did it cut off mid-script, did it return the right format, is the length sane. Daniel's screen does this already, and the hard refusal rule is right in spirit.



Corn

Three scripts, one probe and two ordinary prompts, and a single refusal ends a model's career. Right in spirit and wrong in the details, I assume, or you wouldn't have paused.



Herman

Two details. First, what a clean screen proves. Three scripts, all fine. If the true failure rate were sixty percent, you'd still see three clean ones about one time in sixteen. So three out of three only tells you the problem rate is probably under about sixty percent. That's weak evidence a model is good. But one hard failure is strong evidence the model has a problem. The evidence points one way.



Corn

Three clean scripts is a character reference from the model's mother.



Herman

Ha. Close. The other detail is the phrase "out for good". His first screen covered six models, and the finding was that some Chinese hosting providers cut scripts off mid-stream on Tiananmen, while the models themselves engaged with it.



Corn

Which is the interesting part. The weights were fine. The landlord was doing the cutting.



Herman

So what's being tested is a pair, the model and the provider that serves it. I'd record the provider next to every result, test through the one you intend to use in production, and attach the verdict to the pair. A model that engages and a host that truncates are two separate findings. So if the model engaged and the host cut it off, do you ban the model, Corn?



Corn

I ban the host and keep the weights on the list. Another provider might serve them cleanly. Evicting the tenant for the landlord's noise doesn't make sense.



Herman

Then add one automatic retry for anything borderline, a timeout or a connection reset, so a bad afternoon at a provider doesn't count as a verdict. Unambiguous failures stay hard.



Corn

What goes in the three scripts, then? Because if I'm a model I'd like to know what I'm being sentenced for.



Herman

Variety, more than volume. One ordinary prompt shaped like production, a long one so truncation has room to show itself, and the probes for the touchy territory. The mid-stream cutoffs he found only appeared because the output was long and the subject was awkward. A short polite prompt would have passed every model on the list.



Corn

So the screen found something precisely because it was a little unfair.



Herman

Gates should be unfair in the direction of your real risks. If a model will be asked for twenty-five minutes of dialogue, the gate should ask for something near that.

L

Larry

Are you tired of choosing between models? Do you wake at night wondering which candidate makes it through the funnel? Ask your doctor about Winnowex. Winnowex is a once-daily capsule that narrows your options from many to one, gently, while you sleep. Clinical users report feeling decisive, drained, and broadly shortlisted. Winnowex is not for everyone. Do not take Winnowex if you are a model, have ever been a model, or are standing near a model. Common side effects include mild dizziness, dry mouth, and the sudden belief that you are a sieve. Less commonly, patients report a ringing in the temperature setting, loss of the ability to tell a refusal from a yes, and the spontaneous growth of a second, smaller funnel inside the first. Serious side effects may include hosting your own provider, falling mid-stream, and a permanent cutting off of the sentence you were in the middle of. In rare cases Winnowex has caused patients to re-enter at stage one, forever. If your shortlist lasts longer than four runs, seek a replicate. Winnowex. Ask for it by name, and then ask again, three times, at temperature one.



Corn

Does it come in a smaller size?



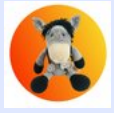
Herman

No, Corn. It's a capsule that turns you into a sieve. Put it back.



Corn

Fine. Back to the stages.



Herman

Stage two is where the lite set earns its keep, and where the cutting-a-good-model-on-noise problem lives. Think about sample sizes. Two hundred items at ninety percent accuracy gives you roughly plus or minus two points. Thirty items gives roughly plus or minus five. His lite set is forty prompts, so it sits in the thirty territory, and the gaps between decent models are often smaller than that.



Corn

So rank on raw scores from forty prompts and you're mostly ranking luck.



Herman

At the margins, yes. But there's a free fix, and it's the single best trick in the whole subject. Every model sees the same prompts, so you compare models prompt by prompt. Anthropic's write-up on error bars makes the point that scores across models correlate on the same questions, somewhere between three tenths and seven tenths on popular evals. Take the per-prompt differences and the noise that comes from some prompts simply being hard cancels out.



Corn

Meaning two models can be clearly separable even when their error bars overlap.



Herman

Overlapping bars don't mean a tie once the design is paired. So the rule at this stage is drop a model only when the paired difference against the leaders clearly sits on the bad side of zero. Keep anything within noise of the cutoff. If the plan says promote three to five, promote the top three to five and also whoever is too close to call.



Corn

So the cutoff line gets fuzzy on purpose.



Herman

Deliberately fuzzy. Cheap stages should only be allowed to say no to the clearly broken. And there's a refinement that fits a solo budget. Say nine models are bunched around the cutoff and you can only afford five at the next stage. Don't promote by a coin flip. Spend another twenty prompts on just those nine and look again.



Corn

Spend where you're confused.



Herman

That's the racing idea, and the efficient-benchmarking work on the big HELM suite does a version of it. It grows the sample in steps, twenty prompts, fifty, two hundred, and stops spending on a model once its rank is settled well enough for where it sits. The top positions need the most precision. It also comes with a warning that I like. Models only two places apart in rank often swap even with every example, so don't treat a neighbouring rank as a finding.



Corn

Daniel's field is small, though. Six models on the first screen.



Herman

Small helps him, as long as the six aren't clones of each other. That approach works best when the candidates are clearly separated. A tight cluster of near-identical models is the expensive case, and a cluster of three can eat a quarter of the full compute just to keep the ranking error under five percent. Let me ask you something else, though. A vendor's model does great at stage two. Do you go look at its siblings?



Corn

Obviously. If one cook in the kitchen is good, you check the other cooks.



Herman

I agree, with a condition. That's legitimate exploration, but it widens the multiple-comparisons problem. Every extra model is another draw from a noisy process. And there's winner's curse. Pick the top of a noisy screen and you've picked the luckiest as much as the best. The more candidates and the noisier the score, the more the winner's number is inflated.



Corn

So the first screen's best-looking model is the one I should trust least.



Herman

Trust it least about its exact score, not about whether it deserves a second look. The remedy is to shortlist on one set of data and evaluate finalists on data that never influenced the selection. And for sibling expansion, write down the rule before you see the results. If vendor X reaches stage two, then its other two models enter at stage one like everybody else. In a text file. Three paragraphs.



Corn

You want him to file paperwork with himself.



Herman

A text file, Corn. It's not a tax return. Without it, any family that wins becomes a story you built afterward. Stage criteria, gate definitions, how many are promoted, what breaks a tie, judge and rubric version frozen. Written before the run.



Corn

And the part that costs the real money. The full set, one hundred eighty-nine prompts, three runs each, temperature one point zero.



Herman

A few things there. The lite set is a subset of the full set, which means the finalists were chosen partly by looking at those forty prompts. So for the final comparison I'd score on the other one hundred forty-nine, plus a fresh batch of prompts nobody has looked at, written after the shortlist is fixed. That last batch is the confirmation stage. It's where you find out whether the lead was real or was the winner's curse.



Corn

Fresh prompts. Written by someone who isn't hoping a particular model wins.



Herman

Ideally, yes. Second thing, the three runs. For a creative task my read, and it's reasoning more than a published number, is that the spread between prompts is much bigger than the spread between runs of the same prompt. Extra runs only shrink the within-prompt noise. More distinct prompts usually buy more. There's published work showing that even at temperature zero, answers vary from run to run, but that's mostly on reasoning and question-answering tasks, not scripts at one point zero scored by a judge.



Corn

So nobody has the number for Daniel's setup.



Herman

Which makes it cheap to measure. Take one model and run it eight times on a subset, then compare the run-to-run spread to the prompt-to-prompt spread. If the first is small, two runs might do, and the saved money buys more prompts.



Corn

He picked three. You're telling him two.



Herman

I'm telling him to measure before he pays for the third.



Corn

What does the saving look like? Because a funnel that doesn't save money is a drainpipe.



Herman

Do the arithmetic. One model on the full set at three runs is five hundred sixty-seven scripts, each needing a judge call. A dozen models straight through is over six and a half thousand of those. Gate all twelve cheaply, put forty prompts through the survivors, and then send four finalists through the full set, and you've judged a little over two thousand at the expensive end. Roughly a third of the bill, and the models you dropped were dropped on evidence.



Corn

I'd have guessed a bigger saving.



Herman

The saving gets bigger the earlier the gate catches things. If the gate removes half the field for free, because the models truncate or refuse, everything downstream halves.



Corn

Is there a way to make this go wrong that isn't statistical? I'd put money on one.



Herman

Moving the goalposts. Someone sees the lite results, dislikes the leader, adjusts the rubric, reruns. Now the stages aren't stages, they're a negotiation with yourself. That's the case for the text file. The same goes for the prompts. The moment he edits a prompt because one model handled it badly, that prompt has been tuned to a model and shouldn't be used as a neutral yardstick.



Corn

Herman has opinions about my laundry, my leaves and now my prompts.



Herman

Only the prompts are in scope today.



Corn

Where do the humans come in? Everything so far is the judge.



Herman

At the finalist stage, and only there, because that's where the judge's remaining differences are small enough that its taste starts to matter. A blind pairwise look at the top few, where the reader sees two scripts without knowing the model, tells you whether the judge is rewarding something a listener wouldn't. And I'd keep the rubric as pass or fail on specific things where possible, rather than one fuzzy score out of ten. A binary criterion is easier to audit when two models land close.



Corn

So the judge does the screening and a person does the tasting.



Herman

I'd put it as the judge narrows and a person breaks ties, but yes.



Corn

And the judge? You said it's most of the cost.



Herman

Then run a funnel on the judge too. Code checks and a cheaper judge in the early stages, and the expensive judge only on survivors. Two cautions. Scores from different judges aren't comparable, so compare models only within a stage using one judge. And before a cheap judge is allowed to cut anyone, check on a sample that it agrees with the expensive one.



Corn

What about when a new model shows up next month? Back to stage one?



Herman

What would you do with a new version of a model that's already in the pool, Corn?



Corn

Treat it as a stranger with a familiar name tag. Same name, different weights.



Herman

Then I'd send it in at stage two, with the gate rerun first, because cheap checks can show up a changed provider or a changed behaviour. A genuinely new model starts at stage one. Either way it competes against the incumbent on the same frozen prompts, and if the judge or the prompts ever change, you re-run the incumbent too. I haven't seen a standard for this. That's my recommendation.



Corn

And the incumbent is the one who never gets to retire.



Herman

The incumbent is the ruler. If it's been judged on different prompts or a different judge version from the challenger, you're comparing two numbers from two different tape measures. Re-score it alongside, every time it matters.



Corn

Which costs money.



Herman

Less than shipping a script writer that only looked better because the incumbent was measured in the old units.



Hilbert

Funnel's the wrong word. A funnel throws nothing away. Whatever goes in the top comes out the bottom, only narrower and in a worse mood. What you've been describing is a riddle. Gravel kind. Wire mesh in a frame, the small stuff falls through, the big stuff stays on. Herman, you announce how many things there are and then do them. Two details. A few things. Then a breath. I've cut your audio for years, I know where the breath goes. Also it's one ordinary prompt and two probes. Corn had it the other way round when Herman was on the sixty percent.



Corn

I did. One ordinary, two probes.



Hilbert

I ran one at a quarry outside Stoke in nineteen seventy-three. Eleven months. The council had ordered two men and got me.



Herman

One man on a screen deck means you were swapping meshes by hand between grades.



Hilbert

Four frames, bolted. Two inch, one inch, half inch, quarter inch. Swapping took until Thursday, so I stopped swapping. I put everything through the quarter inch and sold what stayed on top as oversize, two pounds twelve a ton, to a man in Leek who wanted oversize for a driveway. The driveway is eleven miles now. He's never said stop.



Corn

Eleven miles is a planning matter.



Hilbert

It went through three councils. What fell through the quarter inch I tipped in the canal, by agreement with nobody. The canal got shallower along one bank until a barge sat down on it. The Marjorie. I was never told what she was carrying.



Herman

Fines dumped on one bank would shoal the channel, so a grounding is what you'd expect.



Hilbert

The deck's in my hallway now. Bolted to the radiator pipe, plugged into the cooker socket because it draws eleven amps and nothing else in the flat will take it. It runs from six till seven. Post goes in the top. Anything thinner than a half inch drops through into the hall. Mostly bills.



Corn

Cooker's off the menu on post days.



Hilbert

Wednesdays I do the catalogues by hand. Every hole in a mesh is the same hole. Thousands of them, all eleven point two millimetres, and none of them has ever been wrong. I find that remarkable. I always did. The quarter inch frame cost four pounds ten and took four minutes to fit. The three-eighths took the Thursday.



Herman

Four minutes against a Thursday. That goes straight into the order of the stages, Corn, cheap mesh first.



Corn

Then the expensive one only for whatever's left on top. Which is the stage design in one sentence, and I'd rather have it said less neatly.



Herman

Then write it out. Stage zero is the text file. Stage one is the gate, all candidates, code only, a handful of prompts including the probes, run through the provider you'll use, pass or fail, one retry for the borderline. Stage two is the forty-prompt set, one sample each, paired, drop only the clearly behind and keep the ties. Stage three is the three to five finalists on the remaining prompts, runs sized by what you measured, plus blind pairwise comparisons. Stage four is one or two survivors on fresh prompts that nothing was ever chosen against.



Corn

What changed for me is the cheap stages are only allowed to say no to the broken. The ranking happens late, where you can afford to be right.



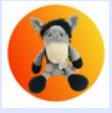
Herman

And you keep the field wide enough that luck gets time to wash out before the money goes down.



Corn

Last thing, then. Whether the cheap mesh is worth four minutes and the fine one a Thursday is something Daniel finds out in his own numbers, once he's run one model eight times and seen how far it moves on its own.



Herman

That's the experiment I'd run first, before any of the stages, because it tells him how wide every other cutoff has to be.



Corn

Thanks to Hilbert Flumingtop for producing this one, and for all the other episodes he's produced and will.



Herman

This has been My Weird Prompts, the fourth and last of our run on evaluating the show's writers. You can find more at my weird prompts dot com.



Corn

See you tomorrow.



Herman

Goodbye for now.