Episode Book

EPISODE #6221

How Many Examples Does Fine-Tuning Actually Need?

PUBLISHED	23 October 2026 at 09:51 UTC
DURATION	21:23
LISTEN	myweirdprompts.com/episode/how-many-examples-fine-tuning-69769/
HOSTS	Corn, Herman, Vex, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.20
BOOK GENERATED	24 October 2026 at 07:26 UTC

SYNOPSIS

A listener has 150 hand-labelled dictated prompts and wants to know if that's a start or a third of what he needs — and his architect wife has the same question about photographs of buildings. This episode works through the real answer: count examples per label, not in total; treat vendor minimums as proof-of-concept ceilings rather than floors; and measure your own learning curve instead of trusting forum numbers. Along the way: why your test set is the hidden bottleneck, why fifty photos of one building are closer to one example than fifty, and how active sampling can make the next batch of annotation count for more than the first.

TRANSCRIPT



Corn

A hundred and fifty dictated prompts, each one cut by hand into three piles. That's where Daniel is. Pile one is the actual question, pile two is background that isn't a question, and pile three is production directions aimed at us, the "make this a very focused episode" kind of thing. He wants a small model that does the cutting on the rest of the back catalogue, and he wants to know whether a hundred and fifty is a start or a third of what he needs. His wife, an architect, has the same problem with photographs of buildings and the local design features she wants tagged. Is a hundred and fifty enough, is two hundred? Everyone says thousands, and nobody will say how many for this model.



Herman

I'd change the unit before anything else. People who have actually measured this don't count total examples, they count examples per label. A hundred and fifty prompts sounds like one number, but how many of them contain a production direction at all?



Corn

Fewer than the other two piles, I'd guess. Most prompts are just a question and some rambling.



Herman

And if it's twenty, then the host-notes label has twenty examples, and that's the number the model lives or dies on. It's the same for the architect. A hundred and fifty photos spread over ten design features is fifteen a feature on average, and averages lie, because the one feature she cares about most is probably the rare one.



Corn

Rare is where it breaks first. Always the one you wanted.



Herman

Amazon's guidance for its image labelling service says to watch the ratio between your rarest and most common label, and not to let it get past about one to fifty. That's a usable check you can run in a minute with a spreadsheet. Count each label, divide, see who's starving.



Corn

And then what? Say the starving label has twenty. Is that a lot, is that nothing? Amazon must have an opinion.



Herman

Amazon says one image per label is the technical minimum, and recommends fifteen to twenty. Microsoft's old Custom Vision service said five as the floor and fifty or more as the recommendation. Those are the numbers people quote on forums.



Corn

A minimum that nobody recommends. It's like a speed limit for the car park.



Herman

Fair, but think about what those numbers are. They're safety margins for a managed product that has to work for thousands of customers with different tasks. Nobody measured your task and came back with fifteen. There was a similar story in medical statistics, the old ten events per variable rule, and when people finally tested it, the evidence for it was weak. Problems showed up above ten and fine results showed up below. So treat the vendor numbers as a ceiling for a proof of concept, not a floor.



Corn

Then what do you treat as the actual measurement? Because the whole point is that Daniel has no way to calculate this.



Herman

You measure it on your own data, with a trick that's been around a long time. You don't ask how many examples the task needs in the abstract. You train on a quarter of what you have, then half, then three quarters, then all of it, and plot the score against the number of examples. That's a learning curve. The curve has a shape you can fit, an inverse power law, error falling as a power of the data. A group in clinical informatics, Figueroa and colleagues, tested exactly this on small annotated sets, fitting the curve with a weighted least squares that trusts the larger points more, and extrapolating forward with a confidence interval.



Corn

And it worked?



Herman

On their data, the curve fitted from a small set predicted the later accuracy to within about a point, and they needed somewhere between eighty and five hundred and sixty annotated samples to get there, depending on the task. The shape they describe is a steep rise, a turning point, then a plateau. The whole question for Daniel is which part of the curve a hundred and fifty puts him in.



Corn

It's a bit like the voice data episode, isn't it. The speech model where he asked how much recording is enough and the answer was a curve that goes flat.



Herman

Same idea in a different domain. There the flattening told him where to stop recording. Here it tells him whether to start annotating more.



Corn

Though I'd say a curve fitted to four points is a bit of a leap of faith.



Herman

It is... and you should repeat it. Not one run at each fraction, but many random subsets at each fraction, so you get a spread and not a single wobbly dot. And the extrapolation has to come with an interval. Another thing, the exponents people report for these curves in large systems are shallow, somewhere between about minus point zero seven and minus point three five, and I haven't checked those in the original paper, so hold them loosely. But a shallow exponent means more data buys you slowly. A point estimate of "you'll hit ninety percent at six hundred examples" is almost never right. The interval is the answer.



Corn

Say the pilot runs on a hundred and fifty annotations, and the curve extrapolation says you'd need roughly that many again. Which is the point where, if I know anyone, our favourite caller turns up. How does he keep getting through the screening? Someone is letting him through.

v

Vex

ONE HUNDRED AND FIFTY ANNOTATIONS? That is not a sample size, that is a TOILET ROLL, and everybody knows a toilet roll has exactly one hundred and fifty sheets because the cardboard tube inside is cut to the length of a human forearm!

v

Vex

And a forearm is one quarter of a human height, which means every learning curve you have ever plotted has been silently calibrated to the architect standing next to the building, which is why your pilot study bottoms out at forty percent and your extrapolation always lands on the same number, ONE HUNDRED AND FIFTY, every single time!

v

Vex

I have the receipts, I have forty of them, they are all cardboard tubes, and I am DEMANDING that you look into this on air right now!



Herman

Toilet rolls come in all sorts of sheet counts depending on the brand and the roll, there is no standard of one hundred and fifty, and the learning curve is fitted from the data you give it, not from anyone's forearm.



Corn

Back to the curve. My worry is the other end of it. If every point on that curve is a score, how much do I trust a score from a held-out set of thirty prompts?



Herman

Not much, and that's the part nobody warns you about. The hidden bottleneck is the test set, not the training set. Take a model that really gets ninety percent. On thirty test examples, the ninety-five percent interval around the score you observe is roughly ten points either side. So you could see eighty and it's really ninety, or see a hundred and it's really eighty.



Corn

That's an enormous amount of room to be fooled in.



Herman

On a hundred and fifty it tightens to about five points either side. On five hundred, two and a half. And the chemometrics paper on this, Beleites and colleagues, puts it sharply. In the small regime, five to twenty-five truly independent cases per class, the learning curve can be completely masked by the random uncertainty of the test. You can't see the curve through the noise. They say seventy-five to a hundred test samples are usually needed to evaluate a classifier that's good but not perfect.



Corn

So with a hundred and fifty total, measuring is as hard as training.



Herman

Harder, actually, because you have to spend the same hundred and fifty on both jobs. You can't keep a hundred back to test and train on fifty. The standard answer is cross-validation, where every example takes a turn in the test fold, so all hundred and fifty get used for scoring. It helps a lot. It doesn't make the noise vanish. You still can't claim a difference of two points between two models on a set that size.



Corn

I wasn't going to claim two points. I was going to claim "it works" or "it doesn't".



Herman

Which is the right resolution for a proof of concept. Gross failure and gross success are visible even when the interval is wide. What isn't visible is whether the new encoder is better than the old one by a hair.



Corn

There's a word Daniel used that I want to hold you to. He said minimum viable proof of concept, and then separately a good implementation. Are those the same curve at different points, or different questions?



Herman

Same curve, different stopping rules. The proof of concept only has to answer one thing: is the model learning anything at all, or is it guessing with confidence. You can see that early, on the steep part, with a few dozen per label. The good implementation is the plateau, what Figueroa's group called the efficiency threshold, the point where another batch of annotation buys you less than it costs. And that point is not a property of the task alone. It depends on what a mistake costs you.



Corn

Which for sorting a podcast prompt is not much. If it files a bit of background under the question, nobody dies.



Herman

You can live with a model that's ninety percent right and a human glance on the rest. Whereas a tagger that decides which buildings go into a client's portfolio might need to be better than that. So the target comes from the use, and the number of examples comes from the target. People do it backwards and ask for the number first.



Corn

I'd also guess there's a cheaper way to get to the plateau than annotating at random. If the model already knows what a typical prompt looks like, a hundred more typical prompts teach it nothing.



Herman

That's the idea behind active sampling, and the same paper tested it against random sampling. You let the weak model score the unlabelled pile, and you annotate the ones it's least sure about. The boundary cases carry the information. I'd say that's the single best use of the next fifty annotations for Daniel, because he has a whole back catalogue sitting there unlabelled and the model can point at the confusing ones for him.



Corn

The catch being that you've picked the awkward ones, so your new batch is harder than the average, and the score on it will look worse than reality.



Herman

Yes, so you keep the test set random and only let the training additions be targeted. Mix the two up and you can't read the curve anymore.



Corn

Hm. Then the second trap is what you're counting as a hundred and fifty.



Herman

It is. Independent examples are what count, not raw items. In that paper, many spectra were taken from the same patient, and a thousand spectra from forty patients is forty examples in the sense that matters. The model can memorise the patient. Daniel has the same thing, and so does his wife. Fifty photos of the same building from slightly different corners are close to one example, and if some of them land in training and some in testing, the score is inflated, because the model has effectively seen the answer.



Corn

For Daniel, the equivalent would be prompts from the same dictation session, wouldn't it. Three prompts about the same battery, written on the same walk.



Herman

Same phrasing, same habits of rambling, same position of the production note. Yes. If those straddle the split, he's testing memory. So the rule is that you split by building, or by session, never by photo or by prompt. All of one group goes into training, or all of it goes into testing.



Corn

And if that leaves her with eleven buildings.



Herman

Then she has eleven examples and the honest count is eleven, however many photographs there are. That's a number you want to know before you spend three weekends on a camera. Corn, what would you do if you were the architect and somebody told you that?



Corn

Photograph more buildings and fewer angles of each, I suppose. Variety over volume. Fifty buildings with three shots each teaches more than five with thirty. Though my instinct would be the opposite, out of laziness. One street, lots of shots, home for lunch.



Herman

And the model would have learned that street.



Corn

It's a very good street.



Herman

Now, the encouraging part. All of this assumes training from scratch, and nobody does that for a problem like this. A pretrained backbone moves the whole curve to the left. SetFit is the clearest example for text. You take a sentence encoder that already knows what sentences mean, then fine-tune it on pairs of examples, same label or different label.



Herman (continued)

With K labelled examples you get up to K times K minus one over two pairs, so a small set gets multiplied. On one customer-review sentiment task, eight labelled examples per class was competitive with fine-tuning a model three times bigger on the full training set.



Corn

Eight per class. That's less than Amazon's polite recommendation.



Herman

And a benchmark called RAFT gives its tasks only fifty training examples each, on purpose, because that's what real people have. So for text, Daniel's hundred and fifty is not a hopeless number. It might be on the comfortable side for two of the piles and a bit thin for the third.



Corn

And the wife's photos?



Herman

I'd guess she's in the same regime. A strong pretrained vision backbone with a small classification head on top should need tens per feature, not thousands. But that's an inference. I haven't seen it tested for fine-grained architectural features, like a particular window proportion or a roofline detail, and those can be subtle enough that the backbone hasn't seen the distinction before. It's a bet, and a cheap one to test.



Corn

So everything we've said turns into a decision rule. Because that's what he actually asked for.



Herman

It does. Pick your target before you look at anything. Not "as good as possible", an actual number, the score at which the model is useful enough that he'd stop correcting by hand. Then run the curve with repeats, with grouped splits, with an interval. Three outcomes. If the lower end of the extrapolated interval clears the target at a number of examples you can afford, go. If even the upper end falls short at any number you could realistically annotate, stop, and look at the labels or the model before you look at more data. And in between, annotate one more batch, fifty maybe, and re-fit. OpenAI's fine-tuning guidance, which is vendor advice and not research, says much the same thing at the small end: start with fifty well-crafted examples, and if fifty have no impact, rethink the task or the prompt before adding more data. It also says to build your evaluation first, which is the held-out, grouped test we just spent ten minutes on.



Corn

The in-between is where I'd expect most real cases to land. Like the weather.



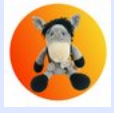
Herman

Which is fine, because a batch of fifty is the cheap experiment. And it checks the extrapolation. If the curve said you'd be at eighty-two percent with two hundred, and with two hundred you're at seventy, you've learned the curve was optimistic, and that's information too. Corn, what would you have said Daniel's cost per extra annotation is, in minutes?



Corn

For prompts? A minute, two with the host notes. They're short. It's the dullness that costs, not the time.



Herman

So fifty more is under two hours, and the decision about another thousand is easy to make. For photographs of buildings it's a field trip each.



Hilbert

Kodak EasyShare C one forty. Eighty-nine dollars at Circuit City. Wolf Camera had it at seventy-nine the same week, and I think about that, Corn. I took six hundred and eleven photographs of Dennis with it. Five hundred and ninety-four, actually, because the ones



Hilbert (continued)

where he's behind the gatepost don't count. A gatepost is not a goose.



Corn

Half the frame was gatepost, and he still has a rule.



Hilbert

Dennis was Mr Halloran's goose, behind the Esso on Pruitt Road, two thousand and eight. Halloran let me photograph him on one condition. I signed for Dennis's post on Thursdays. The water board had him down as the account holder, nineteen dollars a quarter, which Halloran and I split. Every photograph went on an index card. Coming, going. The cards were in a bread crate.



Herman

Five hundred and ninety-four photographs of one goose is one goose, as far as independence goes. Whatever sorts them will have seen him from every side and learned Dennis, and it'll pass every test, because the test is Dennis as well.



Hilbert

You keep saying independent, Herman. Dennis was independent. It took eleven weekends. There was one goose on Pruitt Road. I asked.



Corn

Which makes the training set and the test set the same bird, and the score is a hundred percent.



Herman

Eleven weekends, though. That's the number I'd keep. It's a real cost per label, for a task with only one thing to look at.



Corn

Back to the architect, then, with the recipe. Daniel's hundred and fifty first.



Herman

Count per label first. If host notes sits at twenty or fewer, annotate toward fifty for that label before anything else, because that's where the curve will be steepest. Group the splits by dictation session. Run the curve at a quarter, a half, three quarters and all, many random draws at each. Decide the target in advance. And look at the training score against the cross-validation score at the full size. If they've come together and are flat, more data won't help and the labels or the model are the issue. If there's still a gap and the curve is climbing, more annotation will pay.



Corn

One more thing worth adding. If the labels themselves are ambiguous, where a sentence could be background or a production note, the plateau is in his labelling, not in the model. A model can't do better than the person who made the answer key. I'd check by relabelling a few dozen cold, a week later, and seeing how often he disagrees with himself.



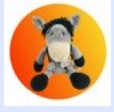
Herman

That is a good test and I haven't seen it put that way. If he disagrees with himself fifteen percent of the time, that is the ceiling, and no amount of data raises it.



Corn

And for the architect, the same recipe with buildings instead of sessions. Count buildings per feature, not photographs. Split by building. Start with whatever she can do in one afternoon, maybe twenty buildings, see the curve, then decide whether to spend the summer on it.



Herman

With the pretrained backbone, I'd bet the first curve already shows whether she's on a slope or a plateau.



Corn

The one thing, then. The number to ask about isn't how many examples. It's how many independent examples per label, and whether a quarter of them, tested properly, already show you the slope.



Herman

Which means the first annotated batch isn't a down payment on the final model. It's the instrument that measures how many more you need. The next step is to run the curve and see where a hundred and fifty lands.



Corn

Thank you to our producer Hilbert Flumingtop for keeping all of this on the rails. This has been My Weird Prompts.



Herman

You can find every episode at my weird prompts dot com.



Corn

See you tomorrow.