Episode Book

EPISODE #6237

Small Models, Big Questions: When Cheap Fables Get Expensive

PUBLISHED	24 October 2026 at 11:23 UTC
DURATION	21:43
LISTEN	myweirdprompts.com/episode/small-models-synthetic-data-29501/
HOSTS	Daniel, Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.20
SOURCES	18 cited by the research (listed at the end)
BOOK GENERATED	25 October 2026 at 06:19 UTC

SYNOPSIS

Small language models can generate three million fables at thirteen and a half cents per thousand, but the real question isn't price — it's whether the generator understands what it's making. This episode maps the parameter tiers from three billion upward, separating tasks where small models excel (controlled transformations of verified facts) from tasks where they quietly fail (inventing and labeling expert judgment). Along the way: a maritime intelligence pipeline that cut inference costs 261-fold by letting GPT-4o and o3-mini teach a 7B student, why instruction-following benchmarks reward "constraint hacking," and why a thesaurus is not a curriculum.

TRANSCRIPT



Daniel

Three million fables, generated with small models. That sounds like an argument for buying synthetic data by the tonne. But I want to know whether the cheap model actually understands what it's making. Can we map the sizes, from roughly three billion parameters upwards, and distinguish following my instructions from being intelligent enough to do the job?



Corn

The fable project puts the temptation rather neatly. Its best quality and cost balance came from an eight billion parameter Llama variant, at about thirteen and a half cents per thousand fables. The moral is practically free.



Herman

The expensive part is deciding whether the moral makes sense. Those generators were working within a recognisable, scaffolded form. A fable has a fairly constrained structure. Generating realistic examples of an obscure technical mistake asks for something different.



Daniel

That's the distinction I'm after. If I specify the format beautifully, have I actually made the task easier, or just made it easier to accept bad examples?



Corn

You can make both happen. A model can become wonderfully reliable at delivering the wrong answer in the right envelope. Herman, I can already see you wanting to solve that with a larger envelope.



Herman

Better contents, actually. Although yes, I have brought several sizes.



Herman

Capability is the binding constraint here. Cheap generation becomes useful only after the model clears the task's minimum standard. Below that, reducing the price buys you a larger rejection pile. Above it, small models can be excellent economics, especially when you already possess the facts and need controlled transformations.



Corn

I'd separate inventing the substance from expressing the substance. Give a model a verified incident and ask it to write a short support ticket. That's a smaller responsibility than asking it to invent an incident, diagnose the fault, and write the correct resolution.



Daniel

Where would you draw that boundary for a synthetic dataset? Most generation tasks involve at least some invention.



Herman

Invention within supplied boundaries can be manageable. You can supply the correct label, the underlying event, the permitted vocabulary, and the intended ambiguity. The model supplies wording. The moment it has to decide whether the event logically deserves that label, you've added a judgment task.



Corn

That changes who ought to check it. A parser can verify the fields. It can't verify that your fictional customer deserves a refund.



Herman

Corn, would you trust a small generator more if every example included an explanation of its label?



Corn

Only if somebody checks the explanation. Otherwise I've bought a witness who agrees with the defendant.



Herman

Fair objection. Explanations can expose contradictions, but generating one doesn't validate the example. You still need an independent test, supplied evidence, or human review where correctness matters.



Herman

There also isn't an authoritative intelligence league table organised by parameter count. We have task benchmarks and self-reported model results. They help locate candidates, but different evaluation setups, thinking budgets, and training histories make precise cross-model rankings slippery.



Daniel

I don't need an artificial IQ score. I want to know when suggesting a three billion parameter model is sensible, and when it's just somebody noticing a low price.



Corn

Then our ranking needs jobs attached to it. A small model that reliably rewrites supplied facts may be a better purchase than a huge one that keeps adding helpful inventions.



Herman

Yes... and size alone doesn't tell us which one does that. We need the capacity map, then a test matching the actual work.



Herman

SLM-Bench compared fifteen small models across nine tasks, measuring accuracy alongside environmental cost. Its useful result is that the relationship isn't a smooth staircase. One model can be both more accurate and more efficient than another on a particular job. Hardware and workload change the result too.



Corn

The cheapest model on the price list may not produce the cheapest usable example. Rejections count. Retries count. Having a person inspect a page of plausible nonsense counts rather heavily.



Daniel

Does that undermine the small-model argument, or does it just mean choosing a better small model?



Herman

Usually the latter, provided the task fits. There's a maritime intelligence example that makes a more interesting economic argument than simply replacing a large generator with a small one. GPT-4o and o3-mini acted as teachers, producing about twenty-one thousand synthetic question and answer pairs from vessel tracking records.



Herman

Those records came from a source containing roughly three billion AIS observations. The resulting examples trained Qwen two point five seven B to answer the specialised questions. The reported result was seventy-five percent accuracy.



Corn

Which still leaves a quarter wrong. I wouldn't let the cost saving steer the ship.



Herman

Nor would I. The reported inference cost reduction was two hundred and sixty-one fold, but that's a result within their system and comparison. It demonstrates a useful architecture, not universal reliability. Spend the stronger model's effort once, teach a narrow student, then run that student repeatedly.



Daniel

The important detail is that the small model wasn't entrusted with inventing the teaching material.



Corn

Right. People hear that a small model succeeded on synthetic data and quietly promote it from pupil to examiner.



Herman

There are several separate roles. Generator, verifier, and trained student can have very different requirements. A student can learn a bounded decision process from examples that required a much stronger teacher to construct.



Daniel

Then how much does instruction-following tell me about the generator? Suppose it consistently obeys detailed constraints.



Herman

It tells you something useful but limited. A verifiable instruction test might check that the answer has the required length, includes a keyword, or uses a prescribed format. Those are real capabilities. They don't establish that the answer solves the substantive problem.



Corn

Perfect margins. Imaginary facts.



Herman

The LIFT paper makes that criticism explicitly. It warns that common instruction tests can reward constraint hacking at the expense of the main task. Its proposed benchmark combines scoped constraints with a requirement to accomplish the actual task. You shouldn't pass merely because your incorrect answer contains the requested word.



Corn

There's a practical version of that. Ask for a realistic complaint containing an ambiguous delivery date. The model gives you a complaint, a date, and the label ambiguous. But the wording makes the date completely clear. Every field passes. The training example teaches the wrong distinction.



Herman

Would you discard that whole task for small models, Corn, or change what you supply?



Corn

I'd supply the ambiguity. Give it two conflicting references, then ask it to vary the customer's language without resolving them. That's still useful generation. I'm removing the part I don't trust it to invent.



Herman

That is the right decomposition. Although some transformations can accidentally remove the ambiguity too, so you test the result.



Corn

You do enjoy finding work after I've removed it.



Herman

Years of living with you. I recognise an optimistic completion estimate.



Daniel

What evidence is there for an actual lower intelligence floor, rather than just bad prompting?



Herman

For scoring journal articles, one comparison concluded that one billion parameters were often too few, and four billion sometimes too few. That doesn't establish a universal cutoff. It does show that clear instructions cannot always compensate for insufficient capacity to make the judgment.



Corn

A beautifully specified rubric still requires someone who understands the paper.



Herman

Synthetic generation adds another demand. You need variation among correct examples. Effective semantic diversity is a useful term for that: distinct outputs that also meet a quality threshold. Counting different sentences without checking their quality rewards inventive errors.



Daniel

Can a small model give me enough genuine variety, or does it keep returning the same example in different clothes?



Herman

Both happen. Larger models generally achieved greater effective diversity in that comparison, while smaller models were more parameter-efficient at producing unique content within a fixed sampling budget. Those findings can coexist. Less capacity can be used efficiently while still imposing a lower ceiling.



Corn

I'd also stop treating topic coverage as the whole diversity problem. A support dataset can mention every product and still have every customer sound like the legal department.



Herman

Variation in style mattered more than topic variation for downstream utility in one synthetic data comparison. Building that variation into generation worked better than applying a style rewrite afterwards.



Daniel

Because the style changes what information someone supplies, not just the wording?



Herman

That's how I'd interpret it. A terse complaint omits things. A rambling one buries the relevant fact. A technically confident customer might use the wrong term. Those are different comprehension demands, not merely synonyms.



Corn

Your generator must understand the mistake well enough to make it convincingly.



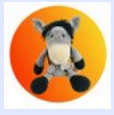
Herman

Prompt optimisation aimed at lexical diversity can improve that metric without reliably improving the downstream model. More unusual words don't guarantee better training material.



Corn

A thesaurus is not a curriculum.



Herman

The fable project proves the positive case, though. Three million stories generated using models no larger than eight billion parameters. Within that scaffold, small models were productive. It doesn't prove those same models can manufacture sound medical reasoning or subtle legal examples. The acceptable invention changes with the domain.



Daniel

Put the size tiers against that distinction, then. Where would you actually begin?



Herman

Around three billion, examples include Llama three point two, Qwen two point five, and SmolLM three. Think classification with clear boundaries, extraction from supplied text, short structured answers, and controlled rewriting. For synthetic work, I'd begin with transformations of verified source material rather than independent expert example creation.



Corn

Llama's reported general knowledge score, MMLU, is around sixty-three. That's useful capacity, not a guarantee that a complicated specification survives the whole answer. I'd keep the output short enough that I can inspect how it fails.



Daniel

Would you recommend that tier, or merely tolerate it if the budget demands it?



Herman

Recommend it for narrow work that passes testing. Not merely tolerate it. The mistake would be recommending it for an undefined task called synthetic generation. Producing spelling variants of a supplied utterance and inventing a realistic engineering diagnosis barely belong in the same procurement conversation.



Corn

Three billion is allowed to be good at its job. It doesn't need promotion.



Herman

Moving to seven or eight billion gives you Qwen two point five, Llama three point one, and Qwen three variants. This is a more credible starting point for scaffolded generation with meaningful variation, modest reasoning, and some domain vocabulary.



Herman

Qwen three eight B reports MMLU of seventy-six point nine and a MATH score around eighty. Those are substantial results at that size. They're not directly comparable to every older result because the evaluation and thinking configuration matter.



Corn

You were about to say it has the intelligence of a much larger model.



Herman

On selected tasks, I was.



Corn

Selected by somebody selling the model.



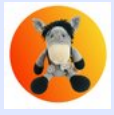
Herman

Sometimes selected by somebody using it. Your suspicion is useful, Corn, but it can't substitute for running the test.



Daniel

What changes when I move to fourteen billion? More reliability, or entirely new kinds of task?



Herman

Both, depending on the model. Phi four is a strong example of training quality upsetting the size order. Its reported MMLU is eighty-four point eight, and MATH about eighty. That makes it a serious candidate for technical generation, code-related examples, and tasks combining several constraints with substantive reasoning.



Corn

I'd expect fewer repairs, not immunity from them. I'd also compare it against the best eight billion candidate rather than assuming the extra parameters purchase a fixed increment of understanding.



Herman

At roughly twenty-four to thirty-two billion, Mistral Small three and Qwen three thirty-two B are examples. This tier is attractive for richer instructions, longer contextual dependencies, and generation where the model must coordinate several facts. Reported coding results in this group reach the high eighties on HumanEval.



Daniel

Is that where you'd start for material that requires judgment?



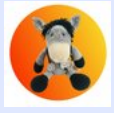
Herman

It's where I'd seriously compare candidates if local deployment and judgment both matter. But... no, I wouldn't declare the tier sufficient beforehand. A specialised fourteen billion model may beat a broader thirty-two billion one on your domain.



Corn

Memory capacity buys admission to the test, not a passing grade.



Herman

Seventy billion, with Llama three point three as an example, is a broadly capable generalist tier. Its MMLU is around eighty-six. The instruction-following compilation gives it an IFEval result around ninety-two. That combination makes it worth considering for more open-ended generation and evaluation of smaller models.



Daniel

Evaluation too? Aren't we just replacing one untrusted model with another?



Corn

We are, unless the evaluator's mistakes are measured. A larger judge can be useful without becoming a court of final appeal. I'd want disagreement cases reviewed, not silently deleted.



Herman

Particularly if generator and judge share the same blind spots. Corn, would you rather pay for the larger model everywhere, or reserve it for uncertain cases?



Corn

Reserve it, if uncertainty means more than the small model saying it feels confident. Route by observable difficulty: missing source evidence, conflicting constraints, outputs outside the usual pattern. Check some apparently easy cases too.



Herman

That gives you a more defensible cascade. Largest models for difficult generation and adjudication, smaller models for bounded production.



Herman

At the large end, examples include Qwen three with two hundred and thirty-five billion total parameters, DeepSeek R one with six hundred and seventy-one billion, and Kimi K two at roughly a trillion.



Corn

The numbers now need their own accommodation.



Herman

Mixture of experts complicates the comparison. Total parameters aren't the number activated for every token. Storage, memory traffic, and computation don't all scale together, so you cannot read those names as direct price multipliers.



Daniel

Does the capability still move appreciably at that end?



Herman

On difficult reasoning, yes. DeepSeek R one reports MATH of ninety-seven point three. General knowledge scores across those large examples sit roughly in the high eighties to low nineties. That suggests meaningful remaining advantages, but benchmark saturation makes the size of the advantage hard to express as one number.



Corn

Also, those are representative open-weight models. We can't rank proprietary frontier models by parameter count when their counts aren't disclosed.



Daniel

Where does thinking mode belong in this map? If a small model can reason, does that move it into another category?



Herman

It adds another axis. SmolLM three supports thinking and direct-answer modes at three billion parameters. Qwen three offers hybrid thinking across its size range. Phi four reasoning is trained using o3-mini demonstrations and performs strongly on complex reasoning tasks. Size, training, and inference effort all matter.



Corn

Extra thinking can help connect facts the model possesses. It doesn't guarantee the missing fact turns up.



Herman

Nor does every task benefit. The journal article scoring comparison found no clear advantage from reasoning models and found them slow. For a bounded judgment, the extra computation may produce more explanation without a better score.



Daniel

Then the useful comparison is quality at a given total cost, including thinking tokens and checking?



Herman

Yes. Also in the actual language. Multilingual instruction tests show substantial gaps between high and low resource languages, sometimes on the order of twenty-five to thirty-five percentage points. An English result isn't a portable guarantee.



Corn

My practical shortlist would begin at seven or eight billion for scaffolded generation, compare fourteen to thirty-two billion when judgment enters, and bring larger models in for difficult invention or checking. Three billion gets narrow, evidenced work, not automatic rejection.



Herman

Would you let it produce the whole batch unattended, Corn, once a pilot passes?



Hilbert

...if the batch is drawer labels, yes. Mine runs overnight. I have forty-seven three billion parameter models on the machine. This month my neighbour Nissim is paying me to write product descriptions for the individual drawers in his kitchen. He rents them separately.



Hilbert (continued)

Nobody gets the cooker.



Corn

The descriptions need to be very clear about access.



Hilbert

They are. Drawer depth. Permitted contents. Opening hours. I give the model those facts. It keeps describing every drawer as ideal for cherished memories. Nissim objects because memories aren't on the permitted contents list. Herman, that distinction you make between supplying the fact and inventing it is currently holding up the cutlery drawer.



Herman

A fixed property list would at least keep the claims consistent. The prose still sounds interchangeable.



Hilbert

It is. There are three descriptions. Spacious. Convenient. Surprisingly spacious. Corn, move a little closer to your microphone. The current applicant wants somewhere for a garlic press, and Nissim is asking me to make the description less domestic. I enter the text with the side of a claw because the keyboard isn't made for an anteater. That takes longer than the generation.



Corn

Less domestic than a kitchen drawer. He's narrowing the market.



Hilbert

He calls it compact commercial storage. I print each approved description and slide it inside the drawer. The applicant has to open the drawer to read it. Nissim says that counts as a viewing. The model follows the length limit perfectly. It still offers a peaceful retreat



Hilbert (continued)

for spoons.



Herman

His corrections are becoming a specification. You could keep the rejected phrases alongside the accepted descriptions.



Corn

Then the next application doesn't begin with cherished memories again.



Hilbert

Unless it's the deep drawer. He approves cherished memories for that one.



Corn

I'd watch whether smaller reasoning models reduce the human correction bill, not merely the generation bill. A cheap batch that repeatedly needs the same edit hasn't become cheap enough. There's no authoritative intelligence ranking to settle that purchase, and downstream usefulness still tells us more than how impressive the generated examples look.



Herman

The tier boundaries will move. For now, pick a candidate by the work it must understand, then compare the usable examples it actually produces. Thanks to Hilbert Flumingtop for producing the show and keeping the levels where they belong.



Corn

For more along these lines, there's episode three, Safetensors or something else. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoyed this conversation, leave us a review. It helps other listeners find the show.



Daniel

Thanks for working through it with me. See you soon.



Corn

See you tomorrow.

SOURCES

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

1. **HuggingFace, SmolLM3 blog, 2025 [primary]**
   https://huggingface.co/blog/smollm3

2. **model card [primary]**
   https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

3. **model card [primary]**
   https://huggingface.co/google/gemma-3-4b-it

4. **model card [primary]**
   https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

5. **SLM-Bench, Aug 2025**
   https://arxiv.org/abs/2508.15478v1

6. **Multi-Model Synthetic Training for Mission-Critical SLMs, rev. 2026-04-12**
   https://arxiv.org/abs/2509.13047v2

7. **LIFT: Scoped Constraints, 2026-10-20**
   https://arxiv.org/abs/2610.28819v1

8. **Thelwall & Mohammadi, *Scientometrics*, rev. 2026-02-17**
   https://arxiv.org/abs/2510.22389v2

9. **Evaluating Diversity and Quality, COLM 2025**
   https://arxiv.org/abs/2504.12522v2

10. **Style Diversity in Synthetic Data, 2026-06-18**
    https://arxiv.org/abs/2606.20400v1

11. **Multi-Sample Prompting / PACE, rev. 2026-03-28**
https://arxiv.org/abs/2506.21138v2

12. **TF1-EN-3M, rev. 2026-05-02**
https://arxiv.org/abs/2504.20605v2

13. **Marco-Bench-MIF, ACL 2025**
https://arxiv.org/abs/2507.11882v1

14. **self-reported benchmark compilation**
https://opensourceaimodels.net/leaderboards

15. **Raschka, base/instruct/reasoning**
https://sebastianraschka.com/faq/docs/base-vs-instruct-vs-reasoning-model.html

16. **Phi-4-reasoning**
https://www.microsoft.com/en-us/research/articles/phi-reasoning-once-again-redefining-what-is-possible-with-small-and-efficient-ai/

17. **Phi-4 Technical Report**
https://arxiv.org/html/2412.08905

18. **THINKSLM, EMNLP 2025**
https://aclanthology.org/2025.emnlp-main.1659.pdf