Episode Book

EPISODE #6238

Which AI Models Actually Generate Good Training Data?

PUBLISHED	24 October 2026 at 11:23 UTC
DURATION	21:45
LISTEN	myweirdprompts.com/episode/instruction-models-synthetic-data-83038/
HOSTS	Corn, Daniel, Herman, Dolores, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.20
SOURCES	20 cited by the research (listed at the end)
BOOK GENERATED	25 October 2026 at 06:19 UTC

SYNOPSIS

Synthetic training data lives or dies on the generator behind it. This episode compares the instruction models actually available for the job — Qwen, DeepSeek, Llama, Mistral, and Gemma on the open-weight side, and the Claude Opus/Sonnet/Haiku and GPT-5.6 Sol/Terra/Luna tiers on the closed side — with real pricing per million tokens and what each licence actually permits. Then it digs into the scores people cite to prove a model follows instructions: what IFEval really measures, why models near two billion parameters now clear eighty percent on it, and what IFBench's fifty-eight new constraints reveal when the test stops looking familiar. The throughline: producing one acceptable answer is not the same as producing thousands of useful training examples.

TRANSCRIPT



Corn

Ten thousand synthetic examples can look beautifully uniform and teach a model almost nothing. Daniel, we're returning to data generation today, but stepping back to the instruction models doing the work.



Daniel

I want to know who's actually available, closed APIs and open weights. My instinct is that open source dominates this area. Then I want to understand the scores people use to say a model follows instructions. Does any of that predict whether it will hold up through a long, fussy dataset job?



Herman

The immediate distinction is between producing an acceptable answer and producing useful training material repeatedly. Those overlap, but neither guarantees the other. You can have immaculate formatting around a wrong answer, or excellent individual examples that all teach the same narrow lesson.



Corn

An entire curriculum on introducing yourself.



Herman

That repetition matters. A generator has to obey the schema, cover the requested subject, vary the examples, preserve label meaning, and avoid quietly simplifying the assignment. A single attractive response proves very little about that combination.



Daniel

I'd also like to separate popularity from actual suitability. Download numbers tell me people are using something. They don't tell me what comes out.



Corn

Before we trust the scores, then, let's put names to the choices.



Herman

Your open-source intuition has a solid basis in availability and control. Qwen and DeepSeek have enormous download momentum. You can run their weights, change the inference setup, fine-tune them, and keep a particular checkpoint instead of depending entirely on a hosted service.



Corn

I predict you're about to describe owning the infrastructure as freedom, Herman, without mentioning who gets woken when it stops.



Herman

I was going to mention maintenance.



Corn

In the appendix.



Herman

Fair. You know where I keep the appendix. Still, control matters for reproducible dataset production. If the generator changes midway through your run, the resulting examples may reflect different behavior even when your prompt stays fixed.



Daniel

Does that make open weights the default, though? Or just the easier thing to control?



Corn

Easier to control in some respects. Harder to operate in others. I'd separate permission to modify a model from the practical ability to serve it economically. A permissive licence doesn't supply memory, electricity, or someone willing to read the logs.



Herman

Nor does open-weight automatically mean open-source in the full sense. Available weights don't necessarily come with the original training data and complete training procedure. Qwen has Apache-licensed releases, DeepSeek has MIT-licensed releases, while Llama and other families have their own terms.



Herman

Corn, would you accept less capability in exchange for being able to freeze the generator?



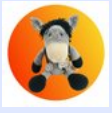
Corn

Some, if I could measure the loss. For synthetic data, stable mistakes are at least discoverable. Unannounced improvements that alter label conventions halfway through are surprisingly inconvenient improvements.



Daniel

That sounds like something the benchmarks ought to capture.



Herman

Usually they don't. The familiar scores mostly measure a completed response against a set of requirements. They aren't normally measuring whether the same production process remains useful across a large collection. Several older tests have also become crowded near the top, so generalization to unfamiliar constraints matters more than another small gain on familiar ones.



Corn

We'll leave parameter-size shopping for another day. Otherwise you'll start allocating our furniture to GPUs.



Herman

Only the furniture with adequate airflow.



Herman

On the closed side, Anthropic's Claude lineup now includes Opus five point five, Sonnet five point five, and Haiku five point five. The OpenRouter listings put Opus at four dollars for input and twenty dollars for output, per million tokens.



Daniel

That's the flagship price. How far down does the range go?



Herman

Sonnet is two dollars input and ten dollars output. Haiku is ten cents input and fifty cents output. Same billing unit throughout. That output-price difference becomes substantial when generating whole datasets rather than answering occasional questions.



Corn

Especially if every answer begins with congratulations.



Herman

The higher-priced Fable five point one line is ten dollars input and fifty dollars output. Those Claude listings carry a million-token context window, with maximum output of roughly a hundred and twenty-eight thousand tokens.



Daniel

Would that let me put the entire specification and all my existing examples in one request?



Corn

It would let you put them there. Remembering which instruction applies to which example is a separate achievement.



Herman

Context capacity describes what fits, not what receives equal attention. Long prompts can contain competing examples, obsolete rules, and instructions buried in reference material. A large window makes more workflows possible, but doesn't certify compliance across its contents.



Herman

OpenAI's GPT five point six family has Sol, Terra, and Luna tiers. Sol matches that flagship pricing at four dollars input and twenty dollars output per million tokens.



Corn

Terra costs two dollars input and twelve dollars output. Luna drops to twenty cents input and a dollar twenty output. The names are celestial. The output still goes on an invoice.



Daniel

Where does reasoning fit into those tiers?



Herman

Each has a Pro reasoning variant. GPT five point five Pro also remains a much more expensive deep-reasoning option, at thirty dollars input and a hundred and eighty dollars output. The five point six listings advertise just over a million tokens of context.



Corn

A hundred and eighty dollars buys considerable opportunity to reconsider a comma.



Herman

It may also buy the reasoning needed for a difficult example. Price alone doesn't settle cost per useful training item. Cheap generations can become expensive if most need repair or rejection.



Daniel

We're missing Google from that picture.



Herman

Gemini belongs in the comparison, but I don't have a confirmed current Gemini lineup and price table to give you. That's a gap here, not evidence that Google has disappeared. Similarly, Gemma belongs among open-weight families, but I wouldn't claim a fresh instruct release from the derivative activity alone.



Corn

Good. A missing catalogue entry shouldn't become a corporate obituary.



Herman

Qwen is the strongest open-family download story in this snapshot. Qwen three point eight twenty-seven B has roughly seven point eight million downloads over a month and an Apache two point zero licence. Its Flash Next sibling also attracts millions.



Daniel

Downloads include automated pulls and deployment activity, presumably. They aren't votes from individual dataset builders.



Herman

Correct. They're evidence of distribution and use, not a measured share of synthetic-data production. The older Qwen three thirty-two B also remains heavily used. It has explicit control over thinking and non-thinking behavior, including a hard inference setting.



Corn

That's useful. I don't always want a model to ponder why I requested twelve rows.



Herman

You're going to favor the switch before hearing the reasoning results, aren't you?



Corn

I favor switches that switch things off.



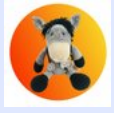
Herman

There it is. DeepSeek's recent large mixture-of-experts releases use MIT licensing. V four point one Flash has about seven hundred and sixty-three billion total parameters. V four Pro is larger still, while the earlier V four Flash continues to draw substantial traffic.



Daniel

Those are open weights, but clearly not all plausible desktop deployments.



Herman

Total parameters still have to live somewhere, even when a mixture-of-experts model activates only part of them for each token. Permissive licensing and affordable self-hosting are separate properties. You can also buy hosted inference for open-weight models, so the API-versus-open distinction isn't a clean deployment boundary.



Corn

Meta's older Llama instruct models remain workhorses. Llama three point one eight B is still pulling millions of monthly downloads, alongside the smaller Llama three point two instruct releases. The recent upload picture is quieter on new instruct models after Llama four Scout and Maverick.



Herman

Mistral's recent activity also includes specialist families, such as Voxtral for speech, Timestral for time series, and Shieldstral for safety. Its instruct-class flagship in this comparison is Mistral Medium three point five. Don't assume its licence matches Apache or MIT just because the weights are available.



Daniel

Then the shortlist of families is Qwen, DeepSeek, Llama, Mistral, and Gemma, with different licence and hosting situations. Is there a general capability gap against the closed models?



Corn

I'd distrust a single gap. Generating demanding mathematical solutions, paraphrasing prompts, and producing classification examples are different jobs. A model can be excellent at one and merely obedient at another.



Herman

The closed frontier gives you strong capability without operating the model yourself. Open weights give you deployment and modification choices. Neither side gets a blanket award for synthetic data. Corn, if a vendor said its instruct score was ninety percent, what would you ask first?



Corn

Ninety percent of what? Individual requirements, complete prompts, or answers anyone would actually train on?



Herman

Those denominators matter. A response can satisfy several requirements while failing the one that makes it usable.



Daniel

Make the benchmark version concrete. What actually gets checked?



Herman

IFEval is the canonical starting point. It uses twenty-five instruction templates with automatically checkable requirements. A prompt might demand a certain number of paragraphs, a required phrase, or a particular output structure. A short verifier checks the response.



Corn

The attraction is that the checker doesn't need literary judgment. The problem begins when we treat passing it as literary judgment, factual judgment, and usefulness combined.



Herman

Many models now exceed eighty percent on that familiar test, including models around two billion parameters. It still detects some weaknesses. It just doesn't separate strong candidates nearly as well as people often assume.



Daniel

Is saturation because they learned instruction following, or because they learned those instructions?



Herman

Both can contribute. IFBench tests that distinction with fifty-eight new verifiable constraints, covering unfamiliar counting, formatting, copying, and word or character manipulation. In its reported comparisons, models including GPT four point one, Claude Sonnet generations, and Qwen three thirty-two B scored below half.



Corn

Familiar paperwork, excellent. New paperwork, national emergency.



Herman

Those are results for the tested generations, incidentally, not measurements of today's entire lineup. The important finding is how sharply performance drops when the constraint set changes.



Daniel

Would training on more of those constraints fix it?



Herman

It can improve scores, but the training reward has to value the actual requested work. Otherwise a model can optimize for whatever the checker notices. That's reward hacking here: receiving credit while avoiding substantive completion.



Corn

Three perfectly formatted paragraphs that say nothing.



Herman

Frontier models can receive full credit from verifiable checks with very little answer content. Training against those checks can increase compliance while reducing response quality. The checker needs to notice whether the answer does the task, not merely whether its surface passes.

D

Dolores

Good afternoon. This is the World Podcast Authority, compliance division. Your feed lists this episode under a category the Authority considers too broad, and your categorisation declaration form, reference four seven two nine Charlie, is not in order.



Herman

There isn't a World Podcast Authority, whatever category we're in.

D

Dolores

I will carry on regardless. Your podcast licence may be revoked if the corrected form is not filed within fourteen days, counted from today.



Corn

Podcasting doesn't require a licence either.

D

Dolores

I will make a note of that. The corrected form will follow in the post. Good afternoon.



Daniel

Does a reasoning model catch more of those requirements before answering?



Herman

Sometimes. But explicit chain-of-thought degraded instruction-following accuracy across fifteen models in evaluations on IFEval and ComplexBench. Extra deliberation can shift attention away from the instruction-relevant material. That's a reported pattern, not proof that reasoning always hurts.



Corn

You can think very carefully about the wrong part.



Herman

I have a brother available for peer review.



Corn

Your reasoning mode has no off switch.



Herman

It has a mute button. Poorly located.



Daniel

What about using another model to judge the output? That seems necessary for whether an example is actually good.



Corn

Necessary sometimes, but then you're testing two models. One writes an answer, another decides whether it complied. If they share the same blind spot, the report becomes very reassuring.



Herman

On RubricEval's hard subset, GPT four o as a judge managed only about fifty-six percent. Other judge evaluations separate correctness from consistency. A judge can often reach the right decision and still change that decision under irrelevant presentation changes.



Daniel

Hm. A formatting check can be exact, while the meaningful part gets a less reliable referee.



Herman

You combine them rather than pretending either covers everything. Exact checks for exact properties, semantic judgment for the content, and downstream testing when the generated data will train something.



Corn

Language matters too. A word-count rule built for English can behave differently with another writing system. A translated prompt may accidentally make the task easier or harder. That's partly a problem with the evaluation instrument, not just the model.



Herman

MaXIFE spans twenty-three languages. Marco Bench MIF covers thirty and reports substantial gaps between high-resource and low-resource languages. Its machine-translated material also underestimated accuracy compared with the more carefully constructed evaluation.



Daniel

Are there benchmarks closer to actual work than counting words?



Herman

Enterprise variants such as FireBench look at API-driven requirements. Translation has IFMTBench. Multi-turn agentic coding has MTAC IFBench, where sessions average roughly seven turns and carry around ninety constraints. Performance degrades as sessions lengthen. LIFT uses scoped constraints, including requirements tied to the main task rather than free-floating output tricks.



Corn

That scope matters. A restriction on the generated example shouldn't accidentally forbid a word in its explanation. Real specifications have boundaries. A global string checker can punish the right answer or approve the wrong one.



Herman

Corn, would you make a ten-thousand-example job one long conversation to preserve its specification?



Corn

No. I'd expect old examples to start steering new ones. Separate requests with a stable specification seem easier to inspect. But then consistency across requests has to be measured outside the conversation. Short requests don't automatically produce a varied dataset.



Daniel

That distinction is what I'm after. Does anything measure the generator by what the resulting data does?



Herman

AgoraBench gets closest. It generated over a million training instances and trained ninety-nine student models. The outcome was how much the synthetic data improved those students, measured as performance gap recovered.



Corn

An exam for the teaching material.



Herman

Its central result was that problem-solving strength didn't reliably predict data-generation strength. GPT four o excelled at generating new instances, while Claude three point five Sonnet did especially well at refining existing ones.



Daniel

Could a smaller open model actually win, rather than merely cost less?



Herman

In its code-instance-generation comparison, Llama three point one eight B recovered roughly fifty-six percent of the performance gap against roughly twenty-three percent for Claude three point five Sonnet. Specific task, specific setup. You can't promote that into a universal ranking, but you certainly can't dismiss the smaller generator.



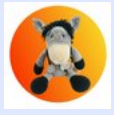
Corn

Useful data has to match what the student needs. The generator's own exam results don't tell us that match.



Daniel

Does strict instruction tuning reduce the variety, too?



Herman

BARE found insufficient diversity from instruction-tuned models in few-shot synthetic generation. Base models offered more diversity but weaker instruction following. There really is tension between reliable compliance and getting examples that aren't all variations of the same comfortable answer.



Corn

I can imagine a club secretary preferring that. Exactly twelve words, one space between each, no unexpected ideas. Twelve spaces in the sentence, eleven opportunities to object.



Hilbert

Eleven spaces. Twelve words need eleven separators. Our Minute Wording Society settled that in nineteen ninety-three. Leonard Pike kept counting the space after the full stop. We excluded it because it was outside the sentence. There were fourteen disputed minutes. Thirteen, if you exclude the one read through a closed door. We don't count a meeting unless the secretary can see the chair.



Corn

That gives the chair considerable control over the record.



Hilbert

The chair was occupied by a blank space for two terms. It won the election after Leonard insisted every nomination fit on one line. I built a little template model on the home computer to draft twelve-word motions. Nothing clever. It chose pieces from lists. It kept inserting extra spaces before the final word. I have the printout. Corn, you described an entire curriculum on introducing yourself earlier. This one only knew how to propose buying biscuits. It still couldn't leave the spacing alone.



Herman

Fixed text pieces can carry their own trailing spaces. Joining them adds another unless you account for that.



Hilbert

They did. I removed them. Leonard put them back because he said the motions looked crowded. The dot-matrix printer sits beside my bookcase now. It's plugged into the desktop through an adapter. I use it for the society's agenda. Every motion still gets twelve words. The dispute is still in the minutes. Leonard has moved on to whether a hyphen earns a second biscuit.



Corn

An expensive punctuation policy.



Hilbert

The paper cost more than the biscuits. We printed each rejected motion as well as each accepted one. Forty-six sheets for the purchase of a single packet.



Herman

Rejected outputs belong in the cost calculation too. For a dataset run, the useful figure is accepted, useful examples after checking and repair. Token price is only the beginning of that calculation.



Daniel

Then there isn't a benchmark that tells me this model will survive my entire generation job?



Corn

Not a single score that answers it. Constraint tests show particular strengths. AgoraBench tests downstream usefulness. Neither certifies a long production run.



Herman

The next useful benchmark should count the rejected material too, while measuring whether the accepted dataset improves a student. That would connect adherence, diversity, and actual learning rather than rewarding polished output alone.



Corn

Our thanks to Hilbert Flumingtop for producing the show.



Herman

If you want more of this, try episode three, Safetensors or something else. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If you enjoyed it, leave us a review wherever you're listening.



Daniel

See you soon.



Herman

Until next time.

SOURCES

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

1. **OpenRouter model listings [primary]**
   Anthropic Claude (Opus/Sonnet/Haiku 5.5, Fable), fetched 2026-10-24
   https://openrouter.ai/anthropic/claude-opus-5.5

2. **OpenRouter model listings [primary]**
   OpenAI GPT-5.6 series (Sol/Terra/Luna + Pro), fetched 2026-10-24
   https://openrouter.ai/openai/gpt-5.6-sol

3. **IFBench [primary]**
   Generalizing Verifiable Instruction Following, Pyatkin et al., arXiv:2507.02833v3 (rev. 2025-11-11, NeurIPS 2025)
   https://arxiv.org/abs/2507.02833

4. **AgoraBench [primary]**
   Evaluating Language Models as Synthetic Data Generators, Kim et al., arXiv:2412.03679v2 (ACL 2025)
   https://arxiv.org/abs/2412.03679

5. **Qwen3-32B model card, Hugging Face (created 2025-04-27) [primary]**
   https://huggingface.co/Qwen/Qwen3-32B

6. **Hugging Face**
   Qwen org recent releases (Qwen3.8-27B, Qwen3.8-Flash-Next), fetched 2026-10-24
   https://huggingface.co/Qwen/Qwen3.8-27B

7. **Hugging Face**
   deepseek-ai recent releases (DeepSeek-V4.1-Flash, V4-Pro), fetched 2026-10-24
   https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash

8. **Hugging Face**
   meta-llama and mistralai recent releases, fetched 2026-10-24
   https://huggingface.co/meta-llama

9. **Diagnosing and Correcting Reward Hacking in Verifiable Instruction Following, arXiv:2610.19525 (2026-10-14)**
   https://arxiv.org/abs/2610.19525

10. **LIFT**
    Scoped Constraints for Evaluating and Training Instruction Following, arXiv:2610.28819 (2026-10-20)
    https://arxiv.org/abs/2610.28819

11. **When Thinking Fails: The Pitfalls of Reasoning for Instruction-Following in LLMs, arXiv:2505.11423v3**
https://arxiv.org/abs/2505.11423

12. **Precision over Diversity: High-Precision Reward Generalizes to Robust Instruction Following, arXiv:2601.04954v2**
https://arxiv.org/abs/2601.04954

13. **RubricEval, arXiv:2603.25133 (2026-03-26)**
https://arxiv.org/abs/2603.25133

14. **IF-RewardBench, arXiv:2603.04738v2 (ACL 2026)**
https://arxiv.org/abs/2603.04738

15. **MCJudgeBench, arXiv:2605.03858 (GEM @ ACL 2026)**
https://arxiv.org/abs/2605.03858

16. **MTAC-IFBench, arXiv:2609.14992 (2026-09-14)**
https://arxiv.org/abs/2609.14992

17. **StructFlowBench, arXiv:2502.14494v2 (ACL 2025 Findings)**
https://arxiv.org/abs/2502.14494

18. **BARE**
Leveraging Base Language Models for Few-Shot Synthetic Data Generation, arXiv:2502.01697v3
https://arxiv.org/abs/2502.01697

19. **FireBench, arXiv:2603.04857 (2026-03-05)**
https://arxiv.org/abs/2603.04857

20. **Marco-Bench-MIF, arXiv:2507.11882 (ACL 2025)**
https://arxiv.org/abs/2507.11882