Episode Book

EPISODE #6030

Who Do You Actually Hire to Train a Model?

PUBLISHED	16 October 2026 at 12:14 UTC
DURATION	24:46
LISTEN	myweirdprompts.com/episode/ai-training-team-hiring-85255/
HOSTS	Corn, Herman, Larry
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

If you were building a team to train a model — not at OpenAI or DeepMind, but at a small lab or enterprise — who would you hire? The answer is stranger than the job postings suggest. At the engineering layer, titles are collapsing: Anthropic lists everyone as a Member of Technical Staff because the research/engineering boundary stopped being useful. At the data layer, they're multiplying: dataset engineers, licensing attorneys, annotation specialists. This episode walks through what each role actually does all day, which jobs fold into one person when you're small, and what happens when regulation — not fashion — drives demand for a career path that barely existed five years ago.

TRANSCRIPT



Corn

Daniel wants us to build a team. Not a real one — an imaginary one, for a small lab or a startup or an enterprise that wants to train models for itself. Not OpenAI, not DeepMind. Somewhere with a budget and a headcount sheet.



Herman

And a hiring plan that has to survive contact with an actual payroll.



Corn

Right. His angle is that we've done the components before. The fine-tuning, the dataset prep, the training runs. What we haven't done is the people. So: if you were assembling that team, who do you actually hire? What does each of them do all day? Which jobs collapse into one person when you're small? And what does any of this look like in five or ten years, when some of these tasks are automated and some of them are suddenly somebody's whole career.



Herman

The premise baked into that is that there are all these distinct specializations, and you go hire one of each.



Corn

And that premise is about half true. Which is the interesting part.



Herman

It's about half true in a specific direction. At the engineering layer, the titles are collapsing. At the data layer, they're multiplying. Those two things are happening at the same time, and most of the conversation about AI jobs misses it because it treats "AI jobs" as one market.



Corn

Start with the collapse, because that's the part that surprises people.



Herman

Anthropic doesn't hire research scientists and research engineers as separate tracks. Everybody's a Member of Technical Staff. One title. And their own line about it is that engineers there do a lot of research and researchers do a lot of engineering. If you read their papers, engineers are authors, often first author. That's not a company being cute with titles — that's a company saying the boundary stopped being useful.



Corn

There's a paper on this.



Herman

It's called the AI Roles Continuum, and the argument is that the research and engineering divide dissolved with the advent of large models. It lays out a spectrum — research scientist, research engineer, applied scientist, ML engineer — and the whole point is that the mandates overlap. Not adjacent. Overlapping.



Corn

Which means the fancy titles are partly a pay band and partly a recruiting fiction.



Herman

Partly. Look at the actual postings. Anthropic's Research Engineer, Pretraining Scaling role — that's training production pretrained models. Training them. DeepMind's research engineers design, implement, scale and evaluate models, and they build the data and training pipelines. At Meta, some of the people with research scientist titles are doing work that reads a lot like a mid-level software engineering job.



Corn

So if I'm hiring, and I write four different job descriptions for those, I'm describing one person four times and paying four different salaries.



Herman

You're describing one person with four different emphases. A research engineer on a pretraining team is going to spend their week on parallelism strategy, mixed precision, gradient accumulation, throughput. Same person, different month, is reading evaluation results and arguing about what they mean. The continuum isn't a claim that the jobs are identical. It's a claim that you can't draw a clean line and shouldn't try.



Corn

Give me the concrete version. What does a Tuesday look like for this person?



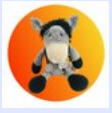
Herman

Monday they're debugging a loss spike that showed up at step forty thousand and nobody knows why. Tuesday they're in a design review arguing about whether to switch the attention implementation, which is an engineering argument that hinges on a research result from a paper that came out three weeks ago. Wednesday they're writing the evaluation harness for the next run. Thursday they're presenting at a research meeting about why the scaling curve bent. Friday they're reviewing somebody else's PR. That's one person. If you tried to split that across two job titles, you'd spend more time in handoff meetings than either of them spent working.



Corn

And the handoff is where the information dies.



Herman

That's the actual argument for the single title. It's not that the work is the same. It's that the work is coupled, and the coupling is where the value is.



Corn

But hold on — if the titles are collapsing, why do the postings still exist? Why does Anthropic still list a Research Engineer, Pretraining Scaling role if everybody's just a Member of Technical Staff?



Herman

Because the title on the offer letter and the title in the job posting are doing two different jobs. The posting is a search filter. It's how you find the person. The internal title is an org design decision. Anthropic lists specific roles externally because that's how candidates search — nobody types "Member of Technical Staff" into a job board. But once you're inside, the boundaries dissolve because the work demands it.



Corn

So the posting is marketing and the internal structure is reality.



Herman

The posting is a funnel and the internal structure is reality. And the mismatch between them is why so many people show up to these jobs with the wrong expectations. They applied for a research scientist role and they're spending their first month on infrastructure. Which, if you read the continuum paper, is exactly what you should have expected.



Corn

Then the data layer, where it's the opposite.



Herman

Liquid AI has a posting that says it plainly. Public data sources are plateauing. Model performance increasingly depends on purpose-built datasets. And then: we need ML-minded engineers who can collect, filter. That's a company saying dataset engineering is its own job. Not a task inside the ML engineer's week — a hire.



Corn

What does that person do on a Tuesday?



Herman

They're not scraping. The scraping is solved, or at least automated. They're deciding what a good sample is, writing the filters, running dedup at a scale where the dedup itself is an engineering problem, and then arguing with the evaluation team about whether the dataset made the model better or just made the benchmark easier. Half the job is judgment about what belongs in the corpus.



Corn

And the other half is defending that judgment to somebody who wants more data faster.



Herman

Always.



Corn

Give me an example of the judgment part, because I think people hear "filtering" and think it's a script.



Herman

Say you're building a code dataset. You can filter for files that parse, files that have tests, files that have documentation. Every one of those filters throws away data, and every one of them encodes an assumption about what good code looks like. If you filter for files with tests, you're now training on a corpus that overrepresents the kind of codebase where people write tests — which is a specific kind of company with a specific kind of engineering culture. You just made a values decision and shipped it as a regex.



Corn

And that decision shows up six months later as a model that's weirdly good at one thing and weirdly bad at another.



Herman

And nobody can trace it back, because the filter is buried in a preprocessing script that's been rewritten twice. That's why the job exists. Somebody has to own the corpus as a thing with a point of view, not as a pipeline artifact.



Corn

Here's the thing that doesn't get said. The reason public data is plateauing isn't just that it's been used. It's that the remaining value is in the parts nobody put on the open web — the transcripts, the case files, the proprietary manuals. Which is a legal problem before it's an engineering problem.



Herman

Which is why the lawyers showed up.



Corn

The lawyers showed up in force.



Herman

Google DeepMind has a posting for Associate Corporate Counsel, IP, focused on copyright — and the description names AI model training and outputs, dataset collection and use, with particular focus on the EU AI Act. That's a copyright lawyer whose practice is training data. Microsoft has a principal corporate counsel focused on data usage for AI training. Mistral is hiring a Strategic Data and Licensing Partnerships Negotiator to acquire premium content for AI training, retrieval and distribution.



Corn

That last one is a job title I would not have predicted five years ago. Content acquisition, but the content is going into training.



Herman

Figure AI wants a commercial attorney who can architect complex AI data licensing structures. That's a phrase from a real posting. Architecting licensing structures. That's not reviewing contracts, that's designing a rights framework for a data supply chain.



Corn

So the legal work isn't peripheral anymore. It's upstream. If you can't license it, the dataset engineer can't have it, and the training run doesn't happen.



Herman

And this is a new career path. A licensing attorney who understands model training is rare, and the demand is being driven by regulation, not by fashion. The EU AI Act is already creating the demand — you can see it in the job descriptions naming it.



Corn

What does that person actually negotiate? Walk me through one deal.



Herman

Say you want a decade of a newspaper's archive. The naive version is you pay for a license and you're done. The real version is: what rights are you getting? Training only, or training and retrieval? Can the outputs reproduce the archive verbatim, and if so, what's the carve-out? What happens when the model is fine-tuned by a third party — does the license travel with the weights? What's the audit right? What's the termination clause if the paper decides it doesn't want to be in the model anymore, and what does "removed from the model" even mean technically?



Corn

That last question is the one that would keep me up at night.



Herman

It keeps everybody up at night. You can't un-train a model. The best you can do is retrain without the data and hope the capability doesn't come back. So the contract has to say something about that, and the lawyer writing it has to understand the technical reality well enough to know what they're promising.



Corn

Which is why you can't hand this to a generalist contracts attorney.



Herman

You can't. They'll write a clause that's legally clean and technically impossible, and you'll find out two years later when somebody tries to enforce it.



Corn

Has that happened yet? Is there a case where a contract promised something the model couldn't deliver?



Herman

Not a public one that I know of, and that's the scary version. The clauses are being written now and the enforcement is years away. Everybody signing these deals is making a bet about what the technology will be able to do when the clause gets tested. And nobody actually knows.



Corn

So the lawyers are writing promises against a roadmap.



Herman

They're writing promises against a roadmap, and the roadmap is being written by the engineers they're not talking to.

L

Larry

Are you tired of models that forget what they learned? Introducing MemoryGuard Pro, the only supplement clinically formulated to keep your neural network regular. Just one capsule a day supports healthy gradient flow, promotes robust parameter retention, and helps your model remember your name. MemoryGuard Pro. Because a forgetful model is a liability. Side effects may include spontaneous overfitting, unexpected emergent behaviors, sudden fluency in languages the model was never trained on, mild recursion, moderate recursion, severe recursion, a persistent belief that it is a golden retriever, involuntary alignment, misalignment, realignment, and in rare cases, a strong preference for being called Steve. Do not take MemoryGuard Pro if you are a frontier model, a base model, or planning to become one. Ask your dataset engineer if MemoryGuard Pro is right for you.



Corn

Ads are just weather now.



Herman

They are. Where were we — the licensing side.



Corn

Keep going, because there's a tier of this market nobody talks about, and it's the biggest one by headcount.



Herman

The annotation labor market. And it's being restructured rather than shrunk. Scale AI, Turing, Toloka — they're moving away from the low-cost labelling pools in Kenya and the Philippines, where people were paid under two dollars an hour, and toward domain experts. Biologists, finance people, physicists, software engineers.



Corn

Experts at what price?



Herman

Turing's offering them salaries twenty to thirty percent above what they're currently making. Their CEO's line is that the resulting model isn't just going to be better than a physicist — it'll be better than a superposition of somebody at the top of physics, computer science and data science at once.



Corn

That's a sales pitch wearing a lab coat.



Herman

It is a sales pitch. But the money is real. And here's the structural part — only ten to fifteen percent of AI budgets go to data. Which sounds small until you remember the size of the budgets. At this scale, ten percent is an enormous absolute number.



Corn

So the industry spent years treating data as the free part, and is now discovering it's the part you have to pay for, and the invoice is arriving.



Herman

Toloka's CEO said it almost exactly — the industry was heavily focused on models and compute, and data was an overseen part. She means overlooked, but overseen works. Finally the industry is accepting the importance of the data.



Corn

How many people is this, total?



Herman

At least three hundred thousand people directly creating AI training data. Best guess is somewhere between four hundred and five hundred thousand. That's a workforce larger than most national industries, and almost none of it has a job title a normal person would recognize.



Corn

Adam Jones put that number together, I think.



Herman

He did, back in April of last year.



Corn

Here's what I want to know. The low-cost labelling is being automated away, and the expert work is growing. Does the low-cost tier come back, or is it gone?



Herman

Gone as a mass employer, is my read. What's left of it is localised — the Data Labellers Association in Kenya says their workers have moved onto tasks that need localised language knowledge and final quality control. So the surviving work is the work a model can't do for itself, which is exactly what you'd expect.



Corn

So the same technology that deleted one tier of jobs created a better-paid tier above it, and nobody planned either one.



Herman

And the higher tier is where the interesting hires are. Turing's expert annotators are people with day jobs — a working physicist picking up annotation work at a higher rate than their salary.



Corn

Which tells you something about academic pay as much as it tells you about AI.



Herman

It tells you a lot about academic pay.



Corn

But I want to push on the "expert annotator" thing, because I think there's a version of it that's real and a version that's marketing. What does a physicist actually do for a data company?



Herman

The real version is: they write the reference answers. If you're building a model that's supposed to reason about, say, condensed matter problems, somebody has to produce the gold-standard solutions that the model gets graded against, and somebody has to grade the model's attempts on problems where there isn't a clean answer. That's not labelling. That's being the examiner.



Corn

And the marketing version?



Herman

The marketing version is "we have PhDs" as a line on a slide. But the underlying need is real, because the evaluation problem and the data problem are the same problem. You can't tell if the model got better unless somebody who knows the field says what better means.



Corn

Which means the expert annotator and the evaluator are often the same person.



Herman

And that's a role that didn't exist five years ago under any name.



Corn

Now do the safety people, because that's the other track that stays separate.



Herman

OpenAI has a Red Team Specialist for cyber. Anthropic's hiring research engineers for the Frontier Red Team, working on RSP evaluations — building gold standard evaluations for catastrophic risk. DeepMind has technical program managers for data collection and model evaluation. Trust and safety red-team analyst listings run from about a hundred and twenty-one thousand to two hundred and seventy-two thousand a year.



Corn

For the person whose job is to break it on purpose.



Herman

In a specific domain. The cyber red-teamer needs to know how attacks actually work. The bio one needs actual biology. You cannot move a training optimization specialist into that seat and expect it to work — the whole value is that they think like an adversary in a field they already know.



Corn

What's the day look like?



Herman

You're designing the test that the model is supposed to fail, and then documenting precisely how it failed, in a form that survives lawyers reading it later. A lot of it is writing. The exploit is the fun part; the report is the job.



Corn

And the MLOps people, who actually keep the lights on.



Herman

Hiring posts treat MLOps as its own role now, separate from ML engineer and separate from the LLM engineer. Build and maintain ML pipelines, infrastructure, APIs, monitoring at scale. The tooling is Docker, Kubernetes, Terraform, MLflow or Kubeflow, Ray, CI/CD, experiment tracking.



Corn

That's the one I'd hire first, honestly.



Herman

Before the research engineer?



Corn

Somebody has to make the training runs reproducible, and if nobody does, every result is a rumour.



Herman

That's the least glamorous correct answer on this show. And it's the one small teams get wrong — they hire the impressive person and the infrastructure is three shell scripts and a prayer.



Corn

Say the shell scripts got written by someone who left.



Herman

Then the infrastructure is three shell scripts, a prayer, and a wiki page that's out of date.



Corn

And the new research engineer spends their first month reverse-engineering the pipeline instead of doing research.



Herman

Which is the actual cost. You didn't save a headcount. You spent a headcount on archaeology.



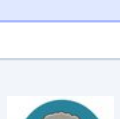
Corn

So that's the map. Now the part Daniel actually cares about — how do these fit together when you're small?



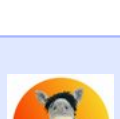
Herman

In a small org, they don't stay separate. And you can see it in the market. Hiring threads from August show candidates listing ML plus MLOps plus LLM evaluation as a single profile. PyTorch, retrieval pipelines, experiment tracking, Docker, CI/CD, all in one person's skill set. That's not a specialist listing. That's someone who has been the only technical person on a project.



Corn

Which is a real job and a terrible job at the same time.



Herman

It's a real job. In a small lab, dataset engineering plus MLOps plus evaluation collapses into one ML engineer role, and realistically that's two or three people, not one, because nobody's good at all three — but the org chart says one.



Corn

What stays separate?



Herman

Legal and licensing. You cannot combine Associate Corporate Counsel, IP, with ML engineer. The backgrounds don't overlap, and the work doesn't overlap either — one is judged on whether the training run converged, the other on whether the rights are clean. And red-teaming stays separate, because the adversarial mindset and the domain knowledge are the whole point. Those are two tracks that don't fold.



Corn

And the division of labor changes depending on what you're actually doing. Training from scratch is a different company than fine-tuning.



Herman

Completely different. From scratch, you need the full stack — pretraining data engineers, training optimization, infrastructure at scale, evaluation, safety. You're buying or renting serious compute, and the training methodology research is a real line item, because at that scale a bad decision costs you a month and a fortune.



Corn

Give me a number on the fortune.



Herman

A frontier pretraining run is tens of millions of dollars in compute. If you get the data mixture wrong and find out at the end, you don't get a refund. You get to do it again. That's why the research line item exists — not because research is nice, but because the cost of a wrong assumption is measured in eight figures.



Corn

Fine-tuning an open-weight model is a much shorter list.



Herman

Much shorter. You need domain dataset engineering, because the value is in the data you brought, not the base model. You need people who understand fine-tuning well enough not to wreck the base model. You need evaluation, and you need MLOps. What you don't need is pretraining-scale infrastructure or a training methodology research group. You can rent the compute, and the base model already did the hard part.



Corn

And "not wrecking the base model" is a real skill, not a throwaway line.



Herman

It's a real skill. Fine-tune too hard on too narrow a dataset and you get catastrophic forgetting — the model gets great at your task and quietly loses the general capability you were relying on. You won't notice until a user asks it something adjacent and it falls over. The person who knows how to avoid that is worth more than the person who knows how to run the training script.



Corn

And the third kind — building a system around an existing foundation model?



Herman

Then you're barely training at all. Prompt engineering, retrieval pipeline builders, evaluation, domain experts. The training team shrinks to almost nothing, and the domain experts become the biggest part of the headcount. Which is exactly the same shift happening at the other end of the market — the domain knowledge is the scarce input, not the training code.



Corn

So the smaller and more applied you get, the more the org chart looks like the data layer and the less it looks like the engineering layer.



Herman

That's the whole movement of this. The engineering roles are the ones that collapse, because a generalist with good infrastructure instincts can cover several of them badly and be corrected later. The domain and the legal roles don't collapse, because you can't fake a physicist or a licensing attorney.



Corn

Which takes us to the five-year question, and I'd like to push on one part of it. The boring prediction is that everything grows. What actually grows?



Herman

Data licensing and governance grows, and it's not speculative — the EU AI Act is already in the job descriptions. Evaluation and red-teaming grow, because the more capable the model, the more the testing matters and the more expensive being wrong becomes. Expert annotation and domain-specific training grows, because public data has plateaued and purpose-built datasets are the differentiator now. And basic annotation, routine evaluation, some MLOps toil — those get automated.



Corn

Basic annotation being automated, expert annotation growing. Those aren't two trends. That's one trend.



Herman

Say what you mean.



Corn

The work didn't go away. It got sorted. The part a model can do, the model does. The bit that needs a person who knows the field, a person gets paid more for. Which is not a comforting story exactly, but it's a more honest one than "the robots took the jobs."



Herman

It's the automation squeeze cutting both ways at once. And the honest reading of the career structure is that it's a barbell, not a pyramid. A small number of blurry, highly-paid engineering and research roles at the top. A large, growing, increasingly specialized data and labor layer underneath. And legal and safety as two tracks that don't sit in either category.



Corn

The middle is thin. That's what I'd tell someone asking how to get in.



Herman

The middle is thin, and it's the middle that a lot of people are training for, because the middle is what the job postings look like when you read them quickly.



Corn

So if you're that person — you've done the ML coursework, you've got the generalist skill set, you're aiming at the title that sounds like the middle — what do you actually do?



Herman

You pick a side of the barbell and you go deep. Either you get good at infrastructure and evaluation, which is the top of the barbell and it's competitive but it's real. Or you bring a domain that a model can't bring itself, and you learn enough about data and evaluation to be useful in that layer. What you don't do is stay in the middle and hope the middle holds.



Corn

Because the middle is where the job descriptions are vague.



Herman

Vague job descriptions are the ones that get rewritten first when the budget tightens.

SOURCES

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

1. **The AI Roles Continuum, arXiv (2026) [primary]**
   https://arxiv.org/html/2601.06087v1

2. **Google, Assembling an ML team [primary]**
   https://developers.google.com/machine-learning/managing-ml-projects/team

3. **Anthropic, Research Engineer, Pretraining Scaling (2025-09-30) [primary]**
   https://job-boards.greenhouse.io/anthropic/jobs/4938432008

4. **Microsoft AI, Researcher [primary]**
   Pretraining Data
   https://microsoft.ai/careers/job/4453591009/researcher-pretraining-data/

5. **Liquid AI, ML Research Engineer, Data [primary]**
   https://jobs.ashbyhq.com/liquid-ai/c7251e1b-d7bf-4d03-8b9e-1382743bef2c

6. **OpenAI Red Team Specialist [primary]**
   https://openai.com/careers/red-team-specialist-cyber-san-francisco/

7. **Anthropic Frontier Red Team [primary]**
   https://jobs.menlovc.com/companies/anthropic/jobs/44990147-research-engineer-frontier-red-team-rsp-evaluations

8. **(copyright) [primary]**
   Google DeepMind IP Counsel
   https://jobs.weekday.works/deepmind-associate-corporate-counsel%2C-ip-

9. **Mistral, Strategic Data & Licensing Partnerships [primary]**
   https://jobs.ashbyhq.com/Mistral.AI/b4e9b991-6a5d-447c-9da6-917f0ef079db

10. **Luma AI, Commercial and IP Counsel [primary]**
    https://jobs.generalcatalyst.com/companies/luma-ai/jobs/88000556-commercial-and-ip-counsel

11. **Business Insider (2025-09-07)**
    https://www.businessinsider.com/ai-training-jobs-data-annotators-labelers-outlier-scale-meta-xai-2025-9

12. **Computing.co.uk, citing FT (2025)**
    https://www.computing.co.uk/news/2025/ai/ai-companies-reducing-low-paid-data-labelling-roles

13. **Adam Jones (2025-04-21)**
https://adamjones.me/blog/300k-ai-annotators/

14. **Nieman Journalism Lab (2025-02-20)**
https://www.niemanlab.org/2025/02/meet-the-journalists-training-ai-models-for-meta-and-openai/

15. **Business Insider (2025-02)**
https://www.businessinsider.com/xai-elon-musk-hiring-ramp-up-data-annotators-2025-2

16. **HN Who is hiring? (2025-10-02)**
https://news.ycombinator.com/item?id=45449870

17. **Silicon AI News, research engineer role analysis**
https://siliconainews.com/careers/roles/research-engineer