Episode Book

EPISODE #6154

Who Checks the Labels? Weak vs Strong Supervision

PUBLISHED	20 October 2026 at 09:56 UTC
DURATION	31:06
LISTEN	myweirdprompts.com/episode/weak-strong-supervision-explained-27173/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
SOURCES	19 cited by the research (listed at the end)
BOOK GENERATED	21 October 2026 at 07:50 UTC

SYNOPSIS

Paying someone to label every example feels reassuring until you ask who checked the labels. This episode untangles the mess of strong supervision, weak supervision, self-supervision, and distillation — separating how a training target is produced from how trustworthy it actually is. From labeling functions and data programming to pseudo-labeling and the Llama 3.1 distillation work, we dig into where the signal really comes from, why specificity matters more than people admit, and how a plausible mistake can quietly become the next model's education.

TRANSCRIPT



Corn

Paying someone to label every example sounds reassuring until you ask who checked the labels. Daniel's prompt gets into that gap. He wants to know what strong and weak supervision actually mean when the teacher might be a human, a rule, or another model generating the entire exercise.



Herman

The invoice tells you who supplied the label. It doesn't tell you whether the label is correct.



Corn

Particularly when the supplier also writes the exam. Daniel's interested in synthetic training data, especially a powerful teacher model teaching a smaller student. Does that count as weak supervision, strong supervision, or distillation? Then there's the practical part. How do you stop a plausible mistake becoming the next model's education?



Herman

I'd separate how the target is produced from how trustworthy it is. Those get mashed together constantly. Human annotation can be inconsistent. Automatic labels can be excellent on a narrow task. A synthetic example can have an objectively checkable answer.



Corn

The human does at least come with a complaints department.



Herman

Sometimes it's the same human, looking more tired. There isn't a universal supervision strength score that settles this. We describe coverage, noise, specificity, and where the signal comes from. Already that's more than one axis.



Corn

Specificity matters more than people admit. If I label an entire recording as containing speech, I've supplied a correct label. I haven't told a transcription model which words occur where. Correct and useful aren't interchangeable.



Herman

That's a good distinction for strong supervision. Conventionally, you have explicit targets for the task on your training examples. For transcription, audio paired with an accurate transcript. For segmentation, pixels marked with their classes. The canonical expensive workflow pays annotators, defines the labeling instructions, resolves disagreements, and audits the results.



Corn

The instructions are doing some of the teaching before the model sees anything.



Herman

Especially for ambiguous categories. Is a sarcastic customer review positive or negative? Are we labeling literal wording or intended sentiment? A perfectly consistent annotation team can still train the wrong behavior if the specification is wrong.



Herman

Corn, suppose you have a fixed budget for customer support classification. Would you buy more labels or spend it fixing the category definitions?



Corn

Definitions first, then enough labels to find out whether the definitions survive contact with customers. Otherwise you pay people to disagree at industrial scale. I'd also keep an ambiguous category instead of forcing every message into a drawer it doesn't fit.



Herman

That choice changes the learning problem. A model forced to pick a category learns a different boundary from one allowed to abstain. Supervision includes what answers are permitted, not merely who types them.



Corn

Already your tidy spectrum has acquired drawers.



Herman

You predicted I'd draw a straight line and put humans at the respectable end.



Corn

You have the voice you use for diagrams.



Herman

I'm resisting the diagram. Self-supervision and synthetic supervision don't even belong at successive points on one line. One describes how the training target comes from the data. The other describes where examples or labels originate.



Corn

Then keep the distinctions attached to actual work. A stack of annotated recordings, a pool of unmarked messages, a teacher making exercise sheets. What changes between those jobs?



Herman

The access to answers, their precision, and how the training procedure uses examples without trusted answers. We can follow those without pretending every workflow has a place on a single ruler.



Corn

Take the unmarked messages. My instinct is to write a few obvious rules. A message asking for your password and threatening immediate account closure goes into the suspicious pile. Not because the rule understands fraud. Because I need a useful starting point.



Herman

That's data programming. You write labeling functions. A function can vote for a class, vote against one in some formulations, or abstain when it has nothing useful to say. Different functions inspect wording, structure, metadata, whatever is available. Their outputs get combined into training targets.



Corn

The abstention seems important. Otherwise my password rule labels every birthday invitation as legitimate.



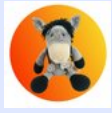
Herman

Precisely the trap. Sorry... the trap is treating no evidence of fraud as evidence of safety. You also have conflicting votes and correlated votes. Five rules mentioning slightly different versions of the same suspicious phrase are not five independent witnesses.



Corn

One witness wearing five hats.



Herman

Aggregation tries to account for noisy sources and, depending on the method, their dependencies. Then a downstream model learns from the probabilistic or consolidated labels. It can generalize beyond the literal rules because it sees features across many examples.



Corn

Useful, but it might also learn that my rules dislike exclamation marks. A school cancellation notice arrives shouting and gets arrested.



Herman

Which is why a small trusted labeled set matters. It exposes errors that agreement between your rules cannot expose. SPEAR is an open-source implementation of this general approach, combining rules and supervised signals. The labor shifts from labeling every item to writing and debugging reusable sources of labels.



Corn

I like reusable labor. You like a new interface for it.



Herman

Both can be true. Interactive heuristic creation is the point of Nemo. Its reported average productivity improvement was about twenty percent. That's time saved creating and improving the weak supervision workflow, not twenty percent more model intelligence.



Corn

An unusually modest software promise.



Herman

The advantage is that the person can inspect what a rule catches, see mistakes, and revise it. They aren't writing rules into the dark and waiting for a final score.

L

Larry

A whisper of certainty. A silhouette of truth. Introducing Supervision, the new experience in curated assurance. Before the answer, there is a feeling. Before the feeling, a person stands behind you holding a clipboard. Supervision surrounds your unlabeled moments with soft, amber confidence. Its proprietary distillation process transforms every hard label into a softer label simply by lowering the lighting. Let uncertainty slip from your

L

Larry (continued)

shoulders. Let your training examples discover the exquisite intimacy of being watched.

L

Larry

Within each presentation box, a single beige sticky note. Place it beside your computer. Write correct upon it. Your model will sense the intention. For deeper results, place another note behind the first, where only the future can see it. Supervision. Human presence, without human involvement. Available in original beige, evening beige, and the limited reserve, administrative beige. Clipboard and person not included. The note is not adhesive on the right-hand edge. This is where possibility enters.



Herman

Soft labels encode a distribution over possible answers. Lower lighting changes neither the target distribution nor the information conveyed to the student.



Corn

Herman, let the stationery go. Your rules still need checking.



Herman

A pipeline with no labeled examples can perform badly when its labeling functions are noisy. Adding a small labeled set can anchor the rules and the classifier. That's why semi-supervised data programming is useful. You combine explicit examples with the wider coverage of rules.



Corn

Then weak and semi-supervised overlap, but they aren't synonyms.



Herman

Weak supervision usually means imperfect, incomplete, or coarse supervision. Some definitions emphasize a small human-labeled set with a much larger unlabeled or imprecisely labeled pool. Semi-supervised learning specifically combines labeled and



Herman (continued)

unlabeled examples. It needn't use labeling functions at all.



Corn

A model could learn from the labeled messages, make tentative predictions on the others, and train on the predictions it trusts. Though I don't like the sound of trusting its own homework.



Herman

That's pseudo-labeling. You can also encourage consistent predictions under changes that shouldn't alter the answer. The unlabeled data supplies information about the input distribution. Your concern is real, though. Bad initial predictions can become training targets, so thresholds, balance, and independent evaluation matter.



Corn

Self-supervision sounds similar only because someone has put self in front of both activities.



Herman

The target in self-supervision comes from the data itself. Hide part of an input and predict the missing part. In next-token prediction, the text supplies the next token. You're not asking an external teacher to decide what the answer ought to be.



Corn

The answer is already in the recording, so to speak. You remove it from the model's view rather than commission somebody to invent it.



Herman

Yes. Though next-token prediction doesn't require literally creating a separate masked document. The causal setup restricts what the model can see. The important point is the target's origin.



Corn

You nearly turned that into a diagram again.



Herman

A small diagram. Respectable size.



Corn

Teacher models complicate this because now the target might look just like ordinary supervised data. A question, an answer, train against the answer. The student doesn't receive a little certificate saying a human wrote it.



Herman

The training loss can be the same. That's the crucial separation. Distillation names a transfer process from teacher to student. Supervised fine-tuning describes fitting explicit targets. Weak supervision can describe the noisy automated source. Those descriptions can all apply to one workflow.



Herman

Would you call it strong supervision if the teacher is almost always correct in your domain?



Corn

I'd call the labels reliable and keep the word almost visible. Strong supervision sounds like a structural description to me, full answers rather than hints. I wouldn't want it quietly changing into a quality endorsement.



Herman

I lean that way too. Someone else may use strong to mean high-quality, detailed guidance. Under that usage, a near-oracle teacher supplies strong supervision. Neither usage tells us the measured error rate unless we report it.



Corn

Strongly supervised into a ditch is still available.



Herman

The Llama three point one distillation work makes the overlap concrete. A teacher with four hundred and five billion parameters supplied responses for students with eight billion and seventy billion parameters.



Corn

The teacher has a substantial class-size advantage.



Herman

Those are model sizes, Corn, not pupils.



Corn

I know. You've still given it the larger desk.



Herman

It used hard-label fine-tuning, response-based distillation without the teacher's logits. For a classifier, a hard label might say fraud. A soft target can say how the probability is divided between fraud and legitimate, retaining information about uncertainty and similarity.



Corn

For generated text, hard doesn't mean a one-word answer. A long worked response is still a selected answer rather than the full distribution of possible next tokens.



Herman

Correct. Reasoning chains can give the student intermediate structure in that selected response. In those experiments, synthetic training improved the smaller models, and with reasoning chains they sometimes matched or exceeded the teacher's zero-shot accuracy on particular datasets.



Corn

The teacher answers cold. The student studies the subject. That's less mystical than a headline saying the apprentice surpasses the master.



Herman

Task-specific training can beat a larger model's unadapted performance. It doesn't establish that the student has all the teacher's abilities. Soft targets can convey more information than hard ones, but a useful hard target is still supervision.



Corn

There are also two different synthetic operations. Label an actual customer message, or invent a customer message and its label. The second operation lets the teacher decide what customers are like.



Herman

That is a much bigger responsibility. Label noise and coverage error are different problems. Your synthetic messages might be flawlessly labeled but omit the awkward, multilingual, abbreviated messages that dominate actual support work.



Corn

Every fictional customer writes with excellent punctuation. Nobody is standing in a car park typing with one thumb.



Herman

Fully synthetic task data can still be useful. You can target rare situations, control difficulty, and generate examples cheaply relative to specialist annotation. But full targets on every example don't establish that the examples resemble deployment.



Corn

I'd want a few ugly real messages before I bought a million immaculate ones.



Herman

That brings us... no, the practical test is whether additional supervision fixes the student's actual mistakes. A larger pile of easy, redundant answers may change almost nothing.



Corn

This is the point where you propose an agent to manage the agent teaching the other agent.



Herman

I do have an example.



Corn

Your enthusiasm has formed a reporting structure.



Herman

The structure earns its keep if feedback is independent and useful. An agentic distillation workflow generates data, trains a small student, examines performance, and targets the next round at what the student is missing. It continues until improvement levels off.



Corn

Provided the feedback comes from more than asking the original teacher whether its own lesson went well.



Herman

Held-out evaluation is central. On SMS phishing detection, one such setup reached about ninety-four percent accuracy, with about ninety-six percent recall. Recall here measures how many malicious messages it catches. High accuracy alone could hide missed attacks if the class balance favors harmless messages.



Corn

A detector that congratulates most messages isn't necessarily a detector.



Herman

The striking comparison was with a DPO baseline, preference optimization using the same synthetic knowledge but without the iterative feedback. The iterative setups achieved accuracy in roughly the high eighties to mid-nineties. The baseline range ran from about half to four-fifths correct.



Corn

That's evidence for the whole teaching procedure, not proof that adding a loop automatically helps. Different training procedures use information differently.



Herman

Agreed. I was about to say the loop makes supervision stronger. Better to say it makes the supervision more useful in that task. We can't isolate every contributing design choice from those headline results. Teacher choice also changed performance substantially.



Corn

What would you inspect first if you inherited that pipeline?



Herman

The test split. Before the agent, before the prompts. Can generation see the final evaluation items? Are near-duplicates crossing splits? Is the loop repeatedly optimizing against the same tiny validation set? I'd want a sealed final test drawn from actual deployment conditions.



Corn

That's your clinician voice. You distrust a machine that declares itself healthy.



Herman

A very sensible prejudice. What would make you suspicious even if the final accuracy looked excellent?



Corn

A model that gets ordinary cases right and gives the same wrong answer for every unusual spelling. I'd inspect disagreement clusters and the cases people actually escalate. Also I'd compare against doing nothing clever. If a small batch of real examples solves the problem, the autonomous teaching department can have the afternoon off.



Herman

A baseline with the same budget is important. Synthetic generation has inference cost, filtering cost, training cost, and review cost. Cheap examples can become expensive if people must untangle systematic mistakes afterward.



Corn

Free labels, specialist cleanup.



Herman

Curriculum design adds another dimension. A student may not benefit from the hardest examples immediately. You can identify a skill it's nearly able to perform, generate appropriate exercises, and revisit weaknesses rather than sample uniformly.



Corn

You mean instruction adjusted to what the pupil can do, rather than what the teacher enjoys explaining.



Herman

I recognize the accusation.



Corn

You explained correlated labeling functions before letting me have a second example.



Herman

Fair. Though you understood them. Pedagogically inspired synthesis explicitly uses mastery learning and the zone of proximal development. On DollyEval, students retained about ninety-five percent of teacher performance with less than a tenth of the parameters.



Corn

Relative performance on that evaluation. Not ninety-five percent of everything the teacher knows.



Herman

Quite. The same work reported improvement on complex reasoning, but the important design idea is matching data to the student's capabilities. Supervision quality includes sequencing and coverage, not just whether individual answers are right.



Corn

It also makes the teacher's blind spots harder to notice. If it chooses exercises, answers them, and grades progress, you've given one opinion a lot of administrative authority.



Herman

Independent checks limit that. For executable code, test behavior rather than just stylistic resemblance to the teacher. For arithmetic, verify the result. For subjective tasks, use a clear rubric and human review of disagreements. Multiple judges can help, but shared training histories can make their mistakes correlated.



Corn

Five hats again.



Herman

Different hats, potentially the same witness. You're allowed that callback within the conversation.



Corn

I wasn't applying for permission.



Herman

There's also the generational problem. A student learns from a teacher, becomes the next teacher, and supplies a replacement training set. Rare patterns can disappear at each step.



Corn

Even if nobody inserts a dramatic falsehood. You simply stop seeing the uncommon thing often enough to learn it.



Herman

That's central to model collapse. The Nature paper describes recursive training forgetting the underlying data distribution. The tails disappear first. Finite samples omit rare cases. The model may lack capacity to represent the distribution, and training may fail to fit even what it could represent.



Corn

Three ways to lose detail. Never sample it, can't express it, or fail to learn it.



Herman

Sampling error is the primary driver in that analysis. Repeatedly sampling from an approximation and replacing the original data compounds the loss. Their inevitability result applies to the recursive conditions being analyzed, not every conceivable use of synthetic data.



Corn

Important, because half this conversation concerns useful synthetic data. One generation of verified exercises isn't a civilization eating its own textbooks forever.



Herman

Retaining the original real data changes the outcome. Accumulating synthetic examples alongside that original data can avoid collapse in analyzed settings where replacement drives it. Other analyses establish limits on the synthetic fraction under their assumptions.



Corn

Keeping the original material seems almost offensively unfashionable. I support it.



Herman

I knew you would defend the old dataset before I finished the sentence.



Corn

It has earned its shelf.



Herman

Keeping it is necessary in those accumulation results, but not a universal guarantee for every training recipe. Weighting matters. A retained dataset can have negligible influence if an enormous synthetic pool overwhelms it.



Corn

Then preserved on disk and represented in training are different promises.



Herman

Yes. Forgetting can also be slow. In discrete-distribution analyses, how long a word survives is roughly proportional to how often it appeared originally. A comfortable-looking average can coexist with rare information quietly disappearing.



Corn

The minority vocabulary goes first.



Herman

Surface fluency can survive factual deterioration too. Work on knowledge collapse examines that separation. The output still sounds composed and competent while factual accuracy falls. Grammar is a poor alarm system.



Corn

An excellent sentence can carry a very bad fact.



Herman

Developers should track error types, rare slices, calibration, and retention of capabilities, not just average scores. Audit targets independently. Record which teacher produced which examples, which checks passed, and what original evidence supports factual answers.



Corn

I'd reserve human attention for where it changes the outcome. Ambiguous labels, novel categories, expensive mistakes, and a random sample that stops the entire audit becoming a search for expected problems.



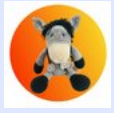
Herman

Random sampling catches surprises. Targeted sampling finds more of the failures you already suspect. You need both. Expert humans may also be necessary to define correctness, while routine checks can be automated.



Corn

Human oversight doesn't have to mean a person approving every synthetic comma.



Herman

Nor is there a settled minimum human-to-synthetic ratio. Recent Fisher-Rao-based bounds challenge older distance-based bounds as uninformative in very high dimensions. Those are mathematical results under assumptions, not a universal recipe for a training mixture.



Corn

I'd rather know whether the real examples cover my weird cases than receive a ceremonial percentage.



Herman

A ratio without provenance or coverage can conceal almost everything important. The practical question is what your independent reference set contains and whether it remains independent as the pipeline evolves.



Hilbert

Herman, does the reference set have to belong to the person running the model? My neighbor Lev is charging me rent for a list of things that are not owls. He charges separately for owls. I agree about keeping the original examples. That's why this arrangement is becoming difficult. His original examples occupy the stairwell, and he won't let the building cleaner move them.



Corn

Ownership doesn't determine whether the examples are useful. Rent for negative examples sounds open-ended, though.



Hilbert

It's twelve shekels a week for not owls. Owls are eighteen. He says the difference covers handling. I'm training a small image classifier to stop the hallway camera alerting me about birds. The computer sits beside my washing machine, plugged into the kitchen extension lead. Lev supplies pictures. This month he starts supplying the subjects as well. He runs a ceremonial sash business from his flat. Business is quiet. He wants the birds to model.



Herman

The sash could become a shortcut. Your classifier might learn clothing rather than the bird.



Hilbert

I tell him that. He puts sashes on the negative examples. There's a kettle wearing deputy treasurer across the landing. The kettle is not mine. I'm not responsible for its title. He brings a wardrobe committee into it. They meet on Wednesdays. There are six of them, all named Pavel. They insist this is convenient because the minutes don't need surnames.



Corn

Convenient until someone records a dissent.



Hilbert

They record it as other Pavel. The committee wants the camera to recognize rank. That's outside my agreement. I pay to stop bird alerts, not establish a chain of command in the corridor. Lev has already appointed a pigeon acting comptroller. It sits on the gas meter and rejects fabric samples. He tells me its silence means the fabric needs work.



Herman

Silence gives him very little information to evaluate.



Hilbert

He charges for interpretation. Corn, lift your microphone slightly. That last sentence loses the end. The pigeon has a deputy now, because Lev says a single officer creates a bottleneck. The deputy is a folded towel. It has limited duties. I can accept that. It doesn't block the stairs.



Corn

I'd check which one has authority to approve the rent.



Hilbert

Neither. That's a tortoise called Miriam. She signs by walking across an ink pad. I can't use the little signature stylus with these claws, so Lev keeps offering her services to me as well. Her availability is terrible. Every appointment is described as morning. There's no date. I wait near the lift with a folder while she finishes a meeting under a radiator.



Herman

Your camera project is acquiring quite a lot of human review.



Hilbert

The review is the problem. Lev gives every image to the wardrobe committee before I see it. Anything without a sash comes back marked incomplete. I want one ordinary picture of an owl. He says ordinary is not a category his business recognizes. You explain abstention earlier, Herman. Very useful word. I put it on the committee form. They think it's a fabric.



Corn

They'll want a swatch.



Hilbert

They want seventeen meters. I'm not buying it. Meanwhile the classifier works fairly well on his pictures. It fails on the actual corridor because the birds spend most of their time behind the wardrobe committee. I ask for an unposed picture. Lev sends an invoice for informality. It's seven shekels. That's the only price in this business I understand.



Herman

An unposed example would at least show you what the camera normally sees.



Hilbert

It shows a delivery man carrying a cabbage through the landing. No birds. Perfectly good negative example. Lev wants it back after training. He says the man only authorizes temporary use of his elbow. We are discussing that this week. The committee rejects my proposed crop because it removes the cabbage's sash.



Corn

The cabbage has rank too.



Hilbert

Visiting secretary. No voting rights. The cleaner moves it to the third floor, and now every Pavel refuses to attend until there is a jurisdiction ruling. The delivery man is still waiting for seven shekels. Lev offers him credit against a sash. He wants money. His elbow



Hilbert (continued)

appears in eleven photographs.



Herman

Permission to use an example is separate from its technical value. A usable reference set needs both.



Corn

That's where provenance becomes more than a training concern. You want to know what an example depicts, how it was obtained, whether it's accurate, and whether you're entitled to keep using it. A beautifully organized dataset can still leave one of those unanswered.



Herman

There's a legal literature forming around access to uncontaminated training data. As generated content becomes more common, identifiable original material may become more valuable. Not every original human statement is true, of course. Originality gives you a traceable starting point, not automatic correctness.



Corn

What changes my view here is separating the target from the evidence behind it. A complete answer from a teacher gives the student something clear to imitate. It doesn't, by itself, tell us whether that answer deserves imitation.



Herman

Distillation describes the transfer. Supervision describes the learning signal. Reliability has to be checked on the task.



Corn

The open question is whether future training pipelines preserve usable, permitted reference data as carefully as they preserve model weights. Keeping eleven photographs isn't enough if the elbow has to leave.



Herman

Thanks to Hilbert Flumingtop for producing today's episode.



Corn

If you enjoyed this one, try episode eleven, How Does Fine Tuning Work Anyway; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode six, How To Fine Tune Whisper. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy the show, leave a review wherever you listen. It helps other curious people find us.



Corn

See you tomorrow.

SOURCES

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

1. **Shumailov et al., *Nature*, published 2024-07-24 [primary]**
https://www.nature.com/articles/s41586-024-07566-y

2. **Knowledge Distillation Using Frontier Open-source LLMs, 2024-10-24 [primary]**
https://arxiv.org/abs/2410.18588v1

3. **Agentic Knowledge Distillation, 2026-02-11**
https://arxiv.org/abs/2602.10869v1

4. **Pedagogically-Inspired Data Synthesis, ICLR 2026, 2026-02-12**
https://arxiv.org/abs/2602.12172v1

5. **Seddik et al., How Bad is Training on Synthetic Data?, 2024-04-07**
https://arxiv.org/abs/2404.05090v1

6. **Gerstgrasser et al., Is Model Collapse Inevitable?, 2024-04-29**
https://arxiv.org/abs/2404.01413v2

7. **Suresh et al., Rate of Model Collapse in Recursive Training, 2024-12-23**
https://arxiv.org/abs/2412.17646v1

8. **Knowledge Collapse in LLMs, 2025-09-05**
https://arxiv.org/abs/2509.04796v1

9. **Preventing Model Collapse: A Fisher-Rao Perspective, IEEE CDC 2026, 2026-09-16**
https://arxiv.org/abs/2609.18878v1

10. **Screening Collapse under Recursive Synthetic Data, 2026-10-11**
https://arxiv.org/abs/2610.15617v1

11. **The Economics of Model Collapse (withdrawn), 2026-08-15**
https://arxiv.org/abs/2605.20279v2

12. **SCAN, NeurIPS 2025, 2025-10-14**
https://arxiv.org/abs/2509.16548v2

13. **TOM multi-teacher distillation, 2025-08-19**
https://arxiv.org/abs/2508.14932v1

14. **Semi-Supervised Data Programming with Subset Selection, ACL 2021 Findings**
https://arxiv.org/abs/2008.09887v3

15. **SPEAR, EMNLP 2022 demo**
https://arxiv.org/abs/2108.00373v3

16. **Nemo, PVLDB 2022**
https://arxiv.org/abs/2203.01382v3

17. **last edited 2026-09-21**
https://en.wikipedia.org/wiki/Weak_supervision

18. **last edited 2026-10-15**
https://en.wikipedia.org/wiki/Self-supervised_learning

19. **Access to uncontaminated data for AI training, 2026-08-19**
https://academic.oup.com/ijlit/article/8765341