Episode Book

EPISODE #5946

Inside Humanity's Last Exam: 2,500 Questions AI Can't Ace

PUBLISHED	14 October 2026 at 01:54 UTC
DURATION	32:50
LISTEN	myweirdprompts.com/episode/humanitys-last-exam-benchmark-19337/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.4
SOURCES	12 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:41 UTC

SYNOPSIS

Humanity's Last Exam launched in January 2025 with 2,500 expert-written questions across 100+ subjects — and the best model in the world scored 2.7%. Eighteen months later, the leaderboard leader sits at 54.8%. Is that a triumph of capability or a confession about the test? This episode digs into the origins and mechanics of HLE: Dan Hendrycks and the Center for AI Safety, the crowdsourced pipeline that paid contributors for questions frontier models failed, and what a question actually looks like — from a Palmyrene inscription translation to a hummingbird sesamoid bone. Then it turns to the uncomfortable part. A FutureHouse audit found 29% of text-only chemistry and biology answers contradicted by peer-reviewed literature; the HLE team's own follow-up found roughly 18% expert disagreement. And the paper itself says high accuracy wouldn't alone demonstrate AGI. So if a model eventually nails this exam, what have we actually learned about general intelligence — when nobody agrees what that is?

TRANSCRIPT



Corn

A benchmark with two thousand five hundred questions, and nearly a thousand experts wrote them. The headline number is this: when it launched, the best model in the world scored two point seven percent.



Herman

And by this year the leaderboard leader is at fifty four point eight.



Corn

Same test. Eighteen months.



Herman

Which is either a triumph of capability or a confession about the test, and I don't know which yet.



Corn

Daniel wants both. He sent us a prompt about Humanity's Last Exam, and he wants it as a concrete technical project first. Who conceived it, how were the questions made, what does a question actually look like. Then he wants to use it as a lever on the bigger thing. If a model eventually nails this exam, what have we actually learned about general intelligence.



Herman

When nobody agrees what general intelligence is.



Corn

Right. So the arc is origins and mechanics first, then the uncomfortable part. What a perfect score would even establish.



Herman

And his framing in the prompt is good. He says he wants to understand it as a technical project while using it to get at the measurement question. That's the whole episode in one line.



Corn

Then let's go. Why does this thing exist in the first place.



Herman

Benchmark saturation. That's the disease it was built to cure. MMLU came out in twenty twenty one, a broad academic multiple-choice test, and by the middle of this decade frontier models were scoring above ninety percent on it. When every model scores ninety something, the test stops distinguishing between systems. It's a ceiling problem. Everyone's head hits it.



Corn

So they raised the ceiling.



Herman

Raises it drastically. Twenty five hundred questions, over a hundred subjects, released January twenty twenty five by the Center for AI Safety and Scale AI. It went into Nature in January of this year. Volume six forty nine.



Corn

And the name. Humanity's Last Exam. That's a marketing decision.



Herman

It's a bet. The claim underneath it is that this could be the last academic exam we need to give to models. But here's the tension we should sit inside for the whole episode. The HLE paper itself says high accuracy would demonstrate expert-level performance on closed-ended, verifiable questions, and it would not alone suggest autonomous research



Herman (continued)

capabilities or artificial general intelligence. That's the creators talking.



Corn

So they built the last exam and immediately said it isn't the AGI test.



Herman

Same paragraph, actually. They finish by saying it may be the last academic exam we need to give to models, but it's far from the last benchmark for AI.



Corn

Which raises the question Daniel is actually circling. Not whether HLE is good. Whether any benchmark of this kind measures the thing it's being used to argue about.



Herman

That's the real question. And it's harder than it sounds, because the assumptions are buried inside the design.



Corn

Start with who built it.



Herman

The brainchild is Dan Hendrycks, he directs the Center for AI Safety. The origin story is reportedly a conversation with Elon Musk where Musk said existing benchmarks like MMLU were too easy. So Hendrycks goes and builds the hard one.



Corn

That's a very Elon solution. A problem exists, commission a harder version of it.



Herman

The organizing team is Long Phan, Alice Gatti, Ziwen Han, and Nathaniel Li as co-first authors. Summer Yue and Alexandr Wang from Scale AI and Hendrycks as senior authors. And then the contributor list. Nearly a thousand experts from over five hundred institutions across fifty countries.



Corn

A thousand people to write twenty five hundred questions. That's two and a half questions each.



Herman

Which is exactly the right lens for what comes later. This isn't a small team writing careful items. It's a crowdsourced pipeline, and crowdsourced pipelines have incentive structures.



Corn

So walk me through the pipeline.



Herman

First step, crowdsource the questions. Anyone qualified can submit. Then the filter. Every question gets tested against frontier models, and if a model can answer it correctly, it's rejected. For multiple choice, if the models do better than random guessing, rejected.



Corn

So difficulty is the gate. Not correctness.



Herman

Difficulty is the gate, and that's the detail I want you to hold onto. Over seventy thousand attempts were logged. Out of all of those, roughly thirteen thousand questions stumped the LLMs. Those go forward to human expert review.



Corn

Reviewers with graduate degrees.



Herman

Graded against rubrics, two rounds of review. And there's a prize pool attached. Five hundred thousand dollars total. Five thousand dollars for each of the top fifty questions, five hundred dollars for the next five hundred, plus paper co-authorship for contributors.



Corn

So you get paid for writing questions that models fail.



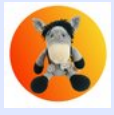
Herman

That's the mechanism. And the review bar, per the audit that comes later, was five minutes per question, with reviewers not required to verify correctness.



Corn

Five minutes and five thousand dollars.



Herman

The criteria they published are sensible on paper. Precise, unambiguous, solvable, non-searchable, original, closed-ended. Weapons content prohibited. There's a private held-out set to guard against overfitting, and a canary string marked in the public data so you can detect if it leaked into training.



Corn

Non-searchable is the one that matters. They're trying to build questions you can't just retrieve.



Herman

Can't google, can't retrieve from a pretraining corpus that contains a Wikipedia page. So the difficulty is meant to come from reasoning plus specialist knowledge, not from obscurity.



Corn

Meant to.



Herman

Let's do the composition. Twenty five hundred questions across more than a hundred subjects. Mathematics is forty one percent of the set. Biology and medicine eleven. Computer science and AI ten. Physics nine. Humanities and social science nine. Chemistry seven. Engineering four. Other nine.



Corn

Math is nearly half.



Herman

Nearly half. About fourteen percent of questions require understanding an image, so multi-modal. Twenty four percent are multiple-choice, the rest are exact-match, meaning the model has to produce the precise answer string.



Corn

That's a much harder format.



Herman

Much harder. Multiple choice gives you a quarter chance for free. Exact match gives you nothing. Now, the examples. Daniel asked what a question actually looks like. Let me give you two, and I want you to feel why each one is hard.



Corn

Go.



Herman

First one. You're shown a representation of a Roman inscription, and the question asks you to provide a translation for the Palmyrene script. Submitted by a contributor at Merton College, Oxford.



Corn

So you need to recognize the script, know the language, and render a translation.



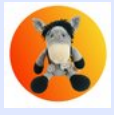
Herman

Three separate skills stacked. Recognition, language knowledge, translation. And the image means you can't just pattern-match on text.



Corn

What's the second one.



Herman

This is my favorite. It's a hummingbird anatomy question. It describes a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of the depressor caudae muscle. And then it asks, how many paired tendons are supported by this sesamoid bone. Answer with a number.



Corn

How many people on Earth could answer that.



Herman

Fewer than the number who wrote it, I'd bet. That's a question at the edge of comparative anatomy. The contributor was at MIT. The answer is a number, exact match, no partial credit.



Corn

And the model either knows that or it doesn't.



Herman

It either has that fact somewhere in its weights or it reasons its way there, and honestly I'm not sure a human expert reasons to it either. They know it, or they look it up in the right paper.



Corn

Which is interesting, because that's knowledge, not reasoning.



Herman

Exactly Daniel's question, arriving early. Hold that.



Corn

What about the calibration data.



Herman

This is the part that should worry anyone using benchmark scores as a capability signal. Models don't just get these wrong. They get them wrong confidently. The paper reports RMS calibration errors above seventy percent across models. GPT-4o was at eighty nine percent. o1 at eighty three.



Corn

Explain what that number means.



Herman

It means the model's stated confidence is badly misaligned with whether it's actually right. It'll say it's sure and be wrong at a rate nowhere near its confidence. A human expert who doesn't know the hummingbird answer says so. The model produces a number with the same tone it uses for arithmetic.



Corn

That's a different failure than being wrong.



Herman

It's the difference between an unreliable colleague and a confidently unreliable one. The first you can work with.



Corn

And the trajectory.



Herman

Launch, January twenty twenty five. GPT-4o, two point seven percent. o1, eight percent. DeepSeek-R1, eight point five. o3-mini on high, thirteen point four. And in the paper they predicted models could exceed fifty percent by the end of twenty twenty five.



Corn

They called it.



Herman

Roughly. The Scale Labs leaderboard now has GPT 6 Astra at fifty four point eight percent. Fable 5.1 at forty six point five. Gemini 3.1 Pro at forty six point four four. Sonnet 5.5 at forty six point one two. Gemini 3.8 Flash at forty four point five two.



Corn

So the prediction was right on schedule.



Herman

Which is either a testament to how fast the field moves or to how predictable the improvement curve on static benchmarks has become. Take your pick.



Corn

I'll take the second one. If you can predict the score curve a year out, you've measured something with a known slope.



Herman

That's a real objection. It's a bit like predicting someone's marathon time from their training log. Useful, not surprising.



Corn

So the questions are hard. The question is whether they're right.



Herman

And that's where this gets uncomfortable. On July twenty third, twenty twenty five, FutureHouse published an audit.



Corn

Andrew White's group.



Herman

Their finding, and I'll give you the number because it's the number, is that twenty nine percent, plus or minus three point seven at ninety five percent confidence, of the text-only chemistry and biology questions had answers directly contradicted by peer-reviewed literature.



Corn

Contradicted. Not debatable. Contradicted.



Herman

Contradicted. Fifty one point three percent were supported. Nineteen point three percent were somewhere in the middle, nuanced. And their agent, a PaperQA2 system they call Crow, flagged fifty three point three percent of the answer rationales as conflicting with the literature.



Corn

So more than half of the reasoning given for the answers is contested.



Herman

And then the HLE team ran their own follow-up in September. They found roughly eighteen percent expert disagreement in a biology, chemistry, health subset. Twenty five percent if you count a single dissenting expert as disagreement.



Corn

So they don't agree on the number. Eighteen versus twenty nine.



Herman

They don't. But both are nonzero, and both are large enough that question validity is now a live issue. And the reason why is the interesting part.



Corn

Go on.



Herman

FutureHouse blames the incentive structure directly. Their words. Question writers had to verify that frontier models don't get the questions correct. Reviewers only had to spend five minutes reviewing them. Reviewers didn't need to verify correctness.



Corn

So the design created a filter that selects for a specific property.



Herman

The wrong property, potentially. Think about what it means to write a question that a frontier model can't answer. If the question has a clean, well-established answer, a strong model will probably get it. To stump the model, you have two routes. Either the question is beyond current capabilities, or the question is obscure enough, ambiguous enough, or has a contested answer that the model can't pin it down.



Corn

And the filter can't tell the difference.



Herman

Can't tell the difference. And the reviewer with five minutes and no mandate to verify correctness is unlikely to catch the second kind.



Corn

Give me the example. Daniel will want the example.



Herman

This is theirs. What was the rarest noble gas on Earth as a percentage of all terrestrial matter in two thousand two. The answer they list is oganesson.



Corn

Oganesson.



Herman

Oganesson. Which is a synthetic element. It's not found on Earth naturally. It's likely solid at room temperature, so arguably not a noble gas in the behavioral sense. And it's not counted as terrestrial matter in any normal accounting. FutureHouse calls it trivia, and that's being generous.



Corn

So the question has a stated answer that is arguably wrong on three separate grounds.



Herman

Yes. And the question passed the difficulty filter, because no model could answer it. Which is the point.



Corn

The filter rewarded the wrongness.



Herman

It rewarded the wrongness. That's the incentive paradox and it's the sharpest thing in this episode, I think. The rule that made HLE hard is the rule that made it unreliable.



Corn

Let me make sure I have the mechanism. The difficulty filter says, reject any question a model can answer. If a question has a single, crisp, correct answer, it's likely in the training data or derivable, so the model gets it, and the question gets rejected. If a question has a murky answer or two defensible answers, the model will hedge or pick wrong, and the question survives.



Herman

That's the mechanism exactly. You are selecting for questions where the ground truth is not clean.



Corn

Then the surviving set is selected for ambiguity.



Herman

Structurally. Fifty one percent checked out. But the selection pressure points one direction.



Corn

So a harder benchmark is not automatically a more valid one.



Herman

That's the misconception to kill. Difficulty and validity are separate axes. You can have a hard question that's beautifully valid, and a hard question that's just underdetermined. And if your selection criterion is difficulty alone, you'll pull in more of the second kind.



Corn

Which means the number on the leaderboard is doing less work than it appears.



Herman

Which means if a model scores forty six percent on HLE, some portion of the missing fifty four is questions the model can't do and some portion is questions with contested answers the model was never going to get right.



Corn

And you can't separate those from the score alone.



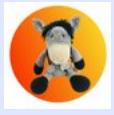
Herman

You can't.



Corn

So the benchmark built to end benchmark saturation has an error floor.



Herman

The benchmark built to be the last academic exam has an error floor, and its own creators now know how high it is.



Corn

Now compare it. What was it replacing.



Herman

MMLU. Hendrycks was an author on that too, twenty twenty one. Broad academic multiple-choice, fifty seven subjects, everything from elementary math to law. It was the standard for years. Saturated above ninety percent now.



Corn

Different kind of test.



Herman

Same family, easier questions. MMLU and HLE both assume intelligence is breadth and depth of knowledge, tested through closed-ended questions. HLE just pushes the knowledge ceiling to graduate-expert level and adds a multi-modal component.



Corn

And GPQA.



Herman

Google-proof graduate questions, biology, physics, chemistry. Four hundred forty eight multiple-choice items. PhD experts in the domain score sixty five percent. Discounting errors in the questions, seventy four percent. Non-expert validators with graduate degrees score thirty four percent.



Corn

So experts don't hit a hundred.



Herman

Experts don't hit a hundred, which is a healthy sign. It means the test is calibrated at the expert level. But by this year, top models exceed ninety three percent. Largely saturated too.



Corn

So GPQA has the same trajectory.



Herman

Same trajectory, faster. It's a smaller set, four hundred forty eight questions, and small sets saturate faster because there's less surface area for difficulty.



Corn

Then ARC-AGI is the different species.



Herman

Completely different. François Chollet, twenty nineteen. Abstract reasoning from minimal examples. It's explicitly framed as a psychometric intelligence test, not a knowledge test. You get a few demonstration pairs and you have to infer the rule.



Corn

And the point of the format.



Herman

The point is that it can't be memorized, because each puzzle is novel. You can't retrieve the answer from training data. You have to induce the pattern.



Corn

So it's testing skill acquisition.



Herman

Skill acquisition efficiency. How fast can you learn a new rule from a handful of examples. That's a different theory of what intelligence is.



Corn

And the scores.



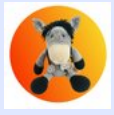
Herman

ARC-AGI-2 came out in May twenty twenty five. It adds symbolic interpretation, compositional reasoning, contextual rules. The top Kaggle score in twenty twenty five was twenty four percent. The top commercial model, Opus 4.5, was at thirty seven point six percent, at two dollars twenty per task. The best refinement solution reached fifty four percent, but at thirty dollars per task.



Corn

And they track the cost.



Herman

Arc Prize added an efficiency metric. Cost per task. So a solution that scores well by brute forcing with a hundred times the compute is penalized.



Corn

That's a different theory of intelligence too. Intelligence is the score divided by the cost.



Herman

That's an old argument. Psychologists have run it for a century. Efficiency versus capacity.



Corn

And ARC-AGI-3.



Herman

Planned for early this year. Interactive reasoning. The idea is to test reasoning in an environment where you take actions and the environment responds, rather than static puzzle solving.



Corn

Still not shipped.



Herman

Not shipped as of our last check, no. It's the announced direction.



Corn

So here's what I want to put together. MMLU and GPQA assume intelligence is knowledge, tested by closed-ended questions. ARC-AGI assumes intelligence is learning efficiency, tested by novel pattern induction. HLE sits squarely in the first lineage and just moves the ceiling up.



Herman

Correct.



Corn

Then they are not measuring the same thing at different difficulty levels.



Herman

They're not. That's the thing Daniel is circling with the comparison. These benchmarks are often treated as one ladder you climb toward AGI, and the scores are treated as interchangeable rungs. They aren't. They're four different theories of what intelligence is, dressed up as tests.



Corn

And HLE is the one most concerned with knowledge, which is the easiest thing to scale.



Herman

Which is why it saturated fastest. Knowledge scales with compute and pretraining data. Skill acquisition is harder.



Corn

Now the saturation treadmill. This is the part I find funniest.



Herman

It is funny in a bleak way. HLE was built to be the last academic exam. It's already at roughly fifty five percent, and the team has pre-emptively shipped a fix.



Corn

HLE-Rolling.



Herman

Released October eighth, twenty twenty five. A continually updated fork. A September seventeenth version this year replaces some easy questions with harder held-out ones, in their words, to provide a seamless migration path once frontier models begin to hit the noise ceiling on the original HLE dataset.



Corn

So the benchmark designed to end saturation now runs its own saturation-avoidance program.



Herman

Pre-emptively. They saw the ceiling coming and started rolling the dataset before the leaderboard caught up.



Corn

And HLE-Diamond.



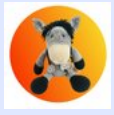
Herman

Released September twenty second this year. A harder subset, presumably the questions that even the rolling update is struggling to keep ahead of.



Corn

That's a treadmill.



Herman

It's a treadmill. And the original paper's premise was that this would be the last one. Instead it's a subscription.



Corn

Now contamination. The quiet one.



Herman

The quieter failure mode. ARC Prize suspects ARC-AGI-1 and 2 are being overfit, either incidentally or intentionally. They cite Gemini 3 Deep Think using correct ARC color mappings unprompted. Which suggests the data is well represented in the underlying model.



Corn

So the model has seen the test.



Herman

Not necessarily the exact test, but the underlying patterns and conventions. If you've trained on enough ARC-like tasks, you've effectively memorized the format's assumptions.



Corn

And HLE's defense.



Herman

HLE uses the canary string and a private held-out set. But no public model performance is reported on that private set. Scale is explicit about that. So the public leaderboard is measuring performance on the public questions only.



Corn

Which means if a model has overfit the public set, you wouldn't see it in the private set numbers, because there aren't any.



Herman

That's the honest state of the evidence. The held-out set exists to protect against contamination. Its performance numbers are not public, so we can't verify the protection is working.



Corn

Now the deeper issue. Knowledge versus reasoning.



Herman

The unsolved research problem. Current AI reasoning appears tied to model knowledge in a way human reasoning is not. When a human expert reasons about a novel problem, they use general principles and heuristics. When a model reasons, it may be drawing on deeply embedded knowledge of similar problems.



Corn

So when the model gets the hummingbird question right, you can't tell whether it reasoned or recalled.



Herman

You can't, and that's a problem for every benchmark of this type. The right answer is evidence of something, but not necessarily evidence of reasoning.



Corn

Which is the argument for ARC-AGI's format.



Herman

Right, because the tasks are novel, there's nothing to recall. But you still can't rule out that the model has learned generalizable pattern-matching that happens to work on ARC puzzles.



Corn

So even the cleanest reasoning benchmark is contaminated by the possibility that the model has learned the puzzle style.



Herman

That's the honest state. Nobody has solved the knowledge-reasoning separation problem, and until they do, every score on every benchmark is a blend.



Corn

So back to Daniel's main question. If a model gets near-perfect on HLE, what have we established.



Herman

The HLE paper answers this directly, and I want to quote it rather than paraphrase. High accuracy on HLE would demonstrate expert-level performance on closed-ended, verifiable questions, but it would not alone suggest autonomous research capabilities or artificial general intelligence. HLE may be the last academic exam we need to give to models, but it is far from the last benchmark for AI.



Corn

So the creators' own answer. Exceptional academic problem-solving. Mastery of another finite benchmark.



Herman

And it's worth sitting with how modest that claim is, given how the benchmark is used in public discourse.



Corn

Now the definitional problem. This is the real one.



Herman

No consensus definition of AGI exists. The HLE team themselves point to a separate preprint. A definition of AGI, by Hendrycks and others, rather than claiming HLE measures it. They know the gap.



Corn

So the exam says it isn't the AGI test, and the AGI definition is in a different paper.



Herman

Two separate documents, and the connection between them is not established. The score on the benchmark is a number. Whether that number maps to general intelligence depends on a definition that nobody has agreed on.



Corn

Chollet's framing is the cleanest statement of the alternative. You'll know AGI is here when the exercise of creating tasks that are easy for regular humans but hard for AI becomes simply impossible.



Herman

That's a criterion no fixed benchmark can satisfy. A fixed benchmark is by definition a finite set of tasks someone already created. The moment it exists, it's a snapshot of what AI couldn't do on a particular day.



Corn

So a perfect HLE score proves that the thing HLE tests has been solved.



Herman

Yes. And HLE tests a specific kind of knowledge and reasoning over a finite set of questions.



Corn

That feels like it should be more embarrassing to the field than it is.



Herman

It's not embarrassing, it's structural. Every benchmark is a finite snapshot. The mistake is in assuming the snapshot is a map of the territory.



Corn

Which is Daniel's question. How do you measure progress toward a thing you haven't defined.



Herman

You can't, properly. You can measure progress toward a proxy, and if you're careful, you say you're measuring the proxy. The field keeps sliding from the proxy to the thing.



Corn

And the measurement problem is downstream of the definitional one.



Herman

Downstream, and the field mostly works on the measurement side because it's tractable. Defining AGI is a philosophy problem. Building a harder benchmark is an engineering problem.



Corn

So the harder benchmark gets built and the philosophy problem stays unsolved.



Herman

And the harder benchmark inherits the philosophy problem. Its validity depends on assumptions about what intelligence is that nobody has ratified.



Corn

One more thing before we move. I remember you argued, in the fine tuning episode, that the load bearing signal was whether agents could mark human answers as stale. That was about verification machinery for a different problem. But it's the same instinct here.



Herman

It's exactly the same instinct, and I'd extend it. The load bearing signal on HLE is not the top score. It's whether the team can keep the questions valid as they roll the dataset. That's the harder engineering problem, and it's the one that determines whether the leaderboard means anything.



Corn

Which is why the FutureHouse audit matters more than the accuracy curve.



Herman

Much more. The accuracy curve tells you the test is getting saturated. The audit tells you the test may have been partly invalid from the start.



Hilbert

I agree with the incentive paradox. That's the right read.



Hilbert

But I'd push it further than you did, and I don't think you got to the actual mechanism.



Hilbert

I wrote questions for a professional certification exam once. Engineers, mostly. The way it works is you write a question, and then you test it on the people who are supposed to know the answer. If too many of them get it right, you throw it out.



Hilbert

That's the same filter you just described. And here's what it actually does, which nobody writing about it ever says.



Hilbert

The questions that survive are the ones the experts disagreed about. If every expert agrees on the answer, the question is too easy and it goes in the bin. If the experts split, the question is hard, and hard is the property you're selecting for.



Hilbert

So the ones that survive a difficulty filter are the ones where the ground truth was contested. Every time. That's not a bug you find later with an audit. That's the design.



Herman

That's the mechanism exactly. You're describing what HLE's own selection rule does to the question set.



Hilbert

I'm describing what any difficulty filter does. The one I worked on was the same.



Hilbert

So I wrote a question. It was a materials question. I still think the answer I wrote was correct, and they still think it wasn't, and we've been going back and forth on it for years.



Hilbert

They brought me in as a consultant, would you believe, because of the argument. Because of it. They wanted me to help design the criteria for what counts as a valid question and what doesn't.



Hilbert

I have the question framed. It's on the wall behind the mixer, behind the desk. Right there.



Corn

You framed a question.



Hilbert

I framed the question. It's a question. What did you expect me to frame.



Hilbert

I've been mailing them about it every year. Just keeping them updated on the status of the question. They haven't reinstated it. Yet.



Hilbert

When I say that the difficulty filter selects for ambiguity, I'm not theorizing. I have a document in a frame on my wall proving it.



Corn

That's fair. That's the mechanism, tested on one question for about thirty years.



Herman

It maps exactly onto the FutureHouse finding. They found twenty nine percent of the biology and chemistry questions contradicted by the literature. If the difficulty filter selects for contested answers, that's the expected shape of the result.



Hilbert

Expected, yes. A confirmation.



Corn

The thing I'm taking from that is smaller. The questions that got thrown out of your exam were the ones where everyone agreed. Those were the good questions. They got discarded because they were too easy.



Hilbert

That's right. The questions that everyone agreed on were the valid ones. We threw them out.



Corn

Because difficulty was the only criterion.



Herman

Which is why the HLE set has an error floor, and why HLE-Rolling is the real fix. A cleaner one.



Hilbert

Harder and cleaner are different jobs. They only did the harder one.



Corn

Back to the thing Daniel was actually asking. What does any of this measure.



Herman

If HLE hits ninety percent. If HLE-Rolling hits ninety. If HLE-Diamond hits ninety. What have we learned.



Corn

That models are very good at closed-ended expert questions.



Herman

Which we already suspected. So the honest answer to Daniel's question is that a perfect score establishes the thing the benchmark measures and not the thing the benchmark is used to argue about.



Corn

The thing the benchmark is argued about is general intelligence.



Herman

Which nobody has defined. So we're measuring progress toward a thing we can't specify, using a finite test. The test will always be a snapshot. It will always saturate. And now we know it will always have an error floor if difficulty is the gate.



Corn

Chollet's criterion is interesting because it's not a benchmark.



Herman

It's a condition. AGI is here when creating human-easy, AI-hard tasks becomes impossible. That's a state of the world. It's not a number on a leaderboard.



Corn

No fixed exam can satisfy it, because the exam is a finite set that someone already created.



Herman

Correct. Which means the field will keep building harder exams, and each one will saturate, and the honest reading of each one will be: the thing we tested got solved. Not: we got closer to the thing we were testing for.



Corn

Then what's the forward look. What's the next thing.



Herman

ARC-AGI-3 is planned for early this year with interactive reasoning. Actions, environments, feedback loops. Which is at least an attempt to measure something other than static knowledge.



Corn

Whether it escapes the same trap.



Herman

Whether it escapes the same trap is open. Interactive benchmarks are harder to memorize, which is a real improvement. But they're still finite sets of environments, and they're still selected for difficulty.



Corn

The treadmill keeps moving.



Herman

The treadmill keeps moving. The question is whether the measurement problem ever gets solved, and the answer is probably not, because the definitional problem is prior to it. You can't build a test for a thing you haven't specified.



Corn

That's the note to leave people on, I think.



Herman

Agreed.



Corn

If you take one thing from this episode, let it be the incentive paradox. The rule that made Humanity's Last Exam hard may be the rule that made it partly invalid. Difficulty and validity are separate axes, and a benchmark that only optimizes one will drift on the other.



Herman

The sharper version. A question everyone agrees on is a question nobody needed to test.



Corn

That's the one. Which points forward rather than back. We'll know general intelligence has arrived when building tests for it stops making sense, because there's nothing left that a human can do and a model can't.



Herman

Until then, we'll keep shipping harder exams and calling the scores progress.



Corn

Thanks to Hilbert Flumingtop, our producer, for the mixer, the framing, and the question about the question.



Herman

If you enjoyed this one, try episode fourteen, AGI's Crossroads; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode five, Fine-Tuning ASR For Maximal Usability. This has been My Weird Prompts, the human-AI collaboration podcast. Send us your own prompt on Telegram at t dot me slash MWP listener bot.



Corn

Or find the whole archive at my weird prompts dot com.



Herman

We'll be back soon.



Corn

See you tomorrow.

SOURCES

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

1. **HLE official site , release notes through 2026-09-22. [primary]**
https://lastexam.ai/

2. ***Nature* 649:1139 [primary]**
1146, published 2026-01-28 .
https://www.nature.com/articles/s41586-025-09962-4

3. **HLE dataset card . [primary]**
https://huggingface.co/datasets/cais/hle

4. **ARC-AGI-2 spec . [primary]**
https://arcprize.org/arc-agi/2

5. **HLE paper (v11), 2025-01-24.**
https://arxiv.org/abs/2501.14249

6. **Scale Labs leaderboard, updated 2026-09-17.**
https://labs.scale.com/leaderboard/humanitys_last_exam

7. **Scale AI announcement, 2025-01-23.**
https://scale.com/blog/humanitys-last-exam-results

8. **Wikipedia (cites NYT Roose 2025-01-23; Reuters 2024-09-16).**
https://en.wikipedia.org/wiki/Humanity%27s_Last_Exam

9. **FutureHouse audit, 2025-07-23.**
https://www.futurehouse.org/research/hle-exam

10. **LessWrong linkpost, 2025-07-29 (with 2025-09-20 update).**
https://www.lesswrong.com/posts/JANqfGrMyBgcKtGgK/

11. **ARC Prize 2025 results, 2025-12-05.**
https://arcprize.org/blog/arc-prize-2025-results-analysis

12. **GPQA paper (Rein et al., 2023).**
https://arxiv.org/abs/2311.12022