Episode Book

EPISODE #6500

Why LLM Leaderboards Need Psychometrics

PUBLISHED	8 November 2026 at 14:42 UTC
DURATION	27:41
LISTEN	myweirdprompts.com/episode/psychometrics-llm-benchmarking-irt-70510/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	9 November 2026 at 08:09 UTC

SYNOPSIS

Standard AI benchmarks treat every question as an equal test of intelligence, but educational testing abandoned raw score averages decades ago. Applying item response theory exposes how popular benchmarks like MMLU and HellaSwag have grown far too easy to separate top models. Adaptive testing methods such as ATLAS can evaluate models in dozens of questions instead of thousands while adjusting rankings for question difficulty, slope, and guessing, exposing the hidden noise beneath standard benchmark leaderboards.

TRANSCRIPT



Corn

Daniel's prompt this week is a hand grenade dressed as a homework question. Educational testing, he points out, solved a problem decades ago that AI benchmarking is only now crashing into: not every question on a test is equally informative. There's a whole field, item response theory, that models each question's difficulty and how sharply it separates strong from weak test-takers, and researchers have started pointing that machinery at LLM benchmarks. What they found is ugly. Many items are near-useless, a well-chosen hundred questions can apparently stand in for thousands, and underneath all of it is a bigger question: do these benchmarks measure what they claim to measure at all?



Herman

And the moment you frame it that way, you can't unsee it. Every leaderboard you've ever looked at is a raw accuracy number. Percent correct, averaged over the whole item set, every question weighted the same whether it's a gimme or a hard problem.



Corn

A scoreboard that counts a coin flip the same as a proof.



Herman

Yes! And human testing abandoned that approach generations ago, for exactly the reason Daniel's picking at. The GRE, the GMAT — the high-stakes admissions tests — they all run on item response theory now. Every item in the bank has a calibrated difficulty, a discrimination slope, a guessing floor. Your score isn't a count of correct answers, it's an estimate of a latent ability with an error bar attached.



Corn

An estimate with an error bar. Imagine a leaderboard trying that.



Herman

The structural reason education got there first is worth pausing on, because it explains everything that follows. High-stakes testing has three things AI benchmarking lacks. A stable population of test-takers, administrations repeated year after year, and lawyers. If



Herman (continued)

your GRE score is wrong, there is a person, an institution, and a lawsuit attached to that wrongness. So the whole apparatus of item calibration grew up under pressure.



Corn

Whereas an LLM benchmark is a paper artifact. Someone builds it once, for a fixed set of models at one moment in time, and then the field just keeps reusing it as the models improve.



Herman

Reused past its calibration date. Nobody retires items that stopped discriminating. And that's how you get the finding that anchors this whole episode — the audit that fitted item response models to eleven benchmarks, roughly forty-two thousand items, across a dozen LLMs. Hardest items in the entire set rarely exceed a difficulty of one on the IRT scale. The top models sit above three.



Corn

So the hardest question on the test is well below the best student's level.



Herman

The ruler is too short to measure the model. Every question is measuring space the model has already climbed past. I love that phrasing because it's a different disease from the ones we usually talk about — it's not contamination, it's not saturation in the leaky sense. It's that the test was never recalibrated for who's now taking it.



Corn

Let's get into how the ruler actually works, then. Because I think people hear "difficulty" and imagine a subjective label.



Herman

Right, no — the move that defines item response theory is refusing to assume each item is equally difficult. Classical test theory treats items as parallel replications of each other. Ten questions on a quiz are ten interchangeable samples of the same underlying quantity, like weighing yourself ten times on the same scale.



Corn

Which is fine if they really are interchangeable. And never true in practice.



Herman

IRT gives every item its own characteristic curve. The simplest version, the one-parameter logistic model, says the probability of a correct answer is one over one plus e to the minus theta minus b — where theta is the test-taker's ability and b is the item's difficulty. The clean property of that model: when your ability exactly equals the item's difficulty, you've got a fifty-fifty chance of getting it right.



Corn

So difficulty isn't "how hard does it feel." It's the ability level at which you flip from likely-wrong to likely-right.



Herman

That's the whole thing. It's a location on the ability scale. Then you add parameters. Discrimination is the slope — how steeply the probability of success rises as ability rises. A sharply discriminating item is one that weak test-takers fail and strong ones pass, with a steep cliff between. A flat item barely separates anyone — everyone gets it right at nearly the same rate regardless of ability. And then there's the guessing parameter, the lower asymptote. A four-option multiple choice item has a floor of twenty-five percent, because a coin-flipping answerer gets a quarter of them right by accident.



Corn

Which means on a saturated multiple choice benchmark, a model scoring eighty percent is only earning fifty-five of that.



Herman

Roughly that shape, yes. And the field mostly uses richer models than the one-parameter version — three or four parameters — but the skeleton is always difficulty, slope, floor.



Corn

Okay. Where does the "well-chosen hundred questions" come in? That's the part of Daniel's prompt that sounds like a magic trick.



Herman

That comes from Fisher information, and this is the bit I find beautiful. Fisher information measures how much an item tells you about ability at a given level. An item that's far too easy for you carries almost no information about you — you'll pass it whether you're mediocre or brilliant. An item far too hard, same problem in reverse. The most informative item for you sits near your level, near that fifty-fifty point.



Corn

So information isn't a property of the question. It's a property of the question and the test-taker together.



Herman

Together, yes. And that's the hinge for adaptive testing. You estimate the test-taker's ability, pick the next item that maximizes information at that estimate, update, repeat. The GRE literally works this way — that's why it adapts as you go. When researchers turned this on LLM benchmarks, the ATLAS system got a full-match ability estimate for HellaSwag using forty-one items out of five thousand six hundred. Forty-eight items for TruthfulQA. Across benchmarks, somewhere between thirty and ninety items did the job.



Corn

Forty-one questions replacing five thousand six hundred. I could sit that exam.



Herman

You could sit it twice before breakfast. There's a medical version too — thirty-eight LLMs run against a human-calibrated medical item bank, adaptive stopping rule, and the resulting proficiency estimates correlated with the full-bank estimates at zero point nine eight eight. Using one point three percent of the items. Evaluation time went from hours per model to minutes.



Corn

And the older approach, TinyBenchmarks — the one that started the "hundred questions" idea —



Herman

Uses around a hundred items but with higher error across every benchmark ATLAS compared against. The adaptive version wins because it's not a fixed subset. It picks the next question based on the model so far, so it concentrates its questions where the uncertainty actually is.

L

Larry

A public service announcement from the Office of Psychometric Responsibility. Citizens, every day you are measured by rulers too short to reach you. Your difficulty is saturated. Your discrimination has collapsed to a slope of nothing. You sit above the item, not below it, and no one is calibrating your characteristic curve. This is not a personal failing. It is a measurement emergency. The responsible choice, the only responsible choice, is the Personal Item Response Theory Kit, a sturdy laminated ruler that estimates your own difficulty, discrimination and pseudoguessing from the comfort of your kitchen table.

L

Larry

It will not shorten the test. It will simply tell you, with grave civic honesty, exactly how far past the top of the scale you have already climbed. Do not be a saturated item. Be calibrated. Ask for the Personal Item Response Theory Kit wherever fine psychometric instruments are responsibly sold.



Herman

No, no, no. That ad mangled the model and I'm not letting it stand. "Your discrimination has collapsed to a slope of nothing" — a slope of nothing means the item doesn't discriminate, fine, but "your difficulty is saturated" is backwards. Saturation is what



Herman (continued)

happens to a test-taker's performance on an item, or to an item's usefulness at a given ability. Difficulty doesn't saturate. Difficulty is a location parameter. An item can be mislocated relative to the population, which is exactly the b equals one versus theta above three problem, but "saturated difficulty" is like saying a street address is full.



Corn

Herman. It's a laminated ruler for kitchen tables.



Herman

Laminated! You can't even write on it properly, the ink smears. And the pseudoguessing estimate from a kitchen table with an n of one —



Corn

Let it go. The Office of Psychometric Responsibility is not a real office.



Herman

There's no such office, that's what I'm saying. Deep breath. Where were we. ATLAS. The efficiency result is honestly the smaller half of what it found, and the smaller half is what everyone reports.



Corn

Go on.



Herman

The sharper claim is about ranking. ATLAS re-ranked more than three thousand evaluated models by IRT ability instead of raw accuracy. Somewhere between twenty-three and thirty-one percent of them shifted by more than ten rank positions. And models with identical accuracies received meaningfully different ability estimates.



Corn

Hold on. Identical accuracy, different ability.



Herman

Because accuracy throws away the item information. Two models both score eighty-two percent — one of them got there by acing all the easy items and guessing on the hard ones, the other by getting every discriminating item right and flubbing some easy ones through sloppiness. Raw accuracy calls that a tie. The item-level model doesn't, because it knows which questions carried signal.



Corn

So if ATLAS is right, a quarter or more of every leaderboard we've ever argued about was ranking noise.



Herman

That's the implication, and it's uncomfortable. And the audit side is just as bad. That forty-two-thousand-item study — the PSN-IRT work — found no benchmark is good at everything. ARC, HellaSwag, and MMLU have become comparatively easy, many items readily solved by strong models. The ones that still retain difficulty are TheoremQA, GPQA Diamond, and interestingly Chinese SimpleQA. And their headline findings were widespread item saturation and data contamination across numerous items.



Corn

Which we've covered before, the contamination side — there's a whole graveyard of benchmarks that broke that way. The new part today is that even setting contamination aside, the items don't discriminate. A test can be perfectly clean and still useless if every question is a speed bump.



Herman

And the audit found the strengths and weaknesses offset each other — the benchmark with the best aggregate quality rank ranked ninth of eleven on one of the sub-measures. There's no free lunch anywhere in the set.



Corn

Corn's test kitchen, here's my question though. All of this assumes the item parameters can be estimated at all. Forty-two thousand items, twelve models — that's a thin grid in some places. GPQA Diamond has a hundred and ninety-eight items.



Herman

It's the thinnest cell in the table, and I'd trust its difficulty estimates less than MMLU's, which has fourteen thousand items. The PSN-IRT people did check stability — they re-estimated parameters on random subsets and got correlations between the subset and full estimates running from about zero point seven four to zero point nine five. Decent, not perfect.



Corn

Now. Before we canonize IRT as the fix — I read a dissenting result and I want your read on it, because it's pointed directly at everything you just described.



Herman

Madaan and colleagues, twenty twenty-four. "Quantifying Variance in Evaluation Benchmarks." I know the one, and I keep going back and forth on it.



Corn

Their finding, as I understand it: methods borrowed from human testing literature — item analysis, item response theory — struggle to meaningfully reduce variance.



Herman

That's their phrasing, yes. And meanwhile a trivial change — reframing a choice task like MMLU as a completion task — often reduced variance for smaller models, around the seven-billion-parameter scale. So you have one line of papers, ATLAS, PSN-IRT, the medical ones, reporting large gains from item response machinery, and this other group finding it doesn't help.



Corn

Which do you believe?



Herman

I don't know, and I'll tell you why it might be both. The studies aren't looking at the same models — Madaan's variance work was on smaller models, the adaptive testing results skew toward frontier models where the ceiling problem is worst. The task framing differs — choice versus completion. And the subsetting differs — a static subset versus adaptive selection, where the item picked depends on the model's estimated ability so far. Those are three different experiments wearing the same label. Nobody has adjudicated it.



Corn

That's the sharpest live disagreement in the literature, then.



Herman

I think so. And there's a second caution that came out this year that complicates the Fisher information story specifically. Someone extended the neural IRT approach with proper uncertainty quantification, and two things fell out. First: most pairwise comparisons among the top models on a standard leaderboard are not statistically distinguishable, despite their differing point-estimate ranks. The leaderboard prints model A above model B, and the honest statement is we cannot tell them apart.



Corn

The leaderboard as a list of coin-toss winners.



Herman

Second, and this is the technical wrinkle: Fisher information for an item, evaluated at a single reference ability, can be nearly zero for many items. Which means naive item selection can pick items that carry no information at the ability level that actually matters. You can build an adaptive test that adaptively selects garbage.



Corn

So even the efficiency program has a trapdoor. The machinery assumes you know roughly where the model is; if your reference point is off, you'll happily select a hundred questions that tell you nothing.



Herman

The fix is estimating the whole posterior over ability rather than a point estimate, which is what the Laplace extension does — but it adds cost to the very thing that was supposed to be cheap.



Corn

Let me pull up to altitude for a second, because there's a bigger question sitting behind all of this, and it's the last third of Daniel's prompt. Everything we've discussed so far is about item efficiency. Whether the questions are informative, whether you can use fewer of them. None of it asks whether the test measures anything real.



Herman

And that's the construct validity conversation, which I think matters more. There's a systematic review from NeurIPS — twenty-nine expert reviewers went through four hundred and forty-five LLM benchmarks. The setup: a benchmark is a phenomenon you claim to measure, a task you use, and a metric you score it with. Like GSM8K — phenomenon is multi-step mathematical reasoning, task is grade-school word problems, metric is exact match.



Corn

Fine so far.



Herman

Seventy-eight percent of benchmarks did define their phenomenon. But of those definitions, nearly half are contested — phenomena with many possible definitions or no clear definition at all. Alignment benchmarks were the worst offenders, targeting things like harmlessness. Then the tasks: fewer than ten percent used complete real-world tasks. Forty percent used constructed tasks. Forty-three percent handcrafted brand-new items from scratch.



Corn

Fewer than one in ten uses a real task. The rest is puzzles built for the occasion.



Herman

Scored on exact match. And the medical people are the bluntest about it, because lives are on the line — one team's line about medical benchmarks is that they are, quote, arbitrarily constructed using medical licensing exam questions. Which, coming from clinicians, is a sentence designed to sting. Licensing exams measure whether a human can be licensed. Whether answering them predicts anything about safe clinical reasoning by a language model is a completely separate claim that almost nobody tests.



Corn

We did an episode once on benchmarking speech recognition, and the whole thesis was that a single aggregate word error rate told you almost nothing about usability. Same disease, different organ.



Herman

Same disease. And the distinction psychometrics draws is the one to hold onto here: reliability versus validity. Reliability is consistency — would you get the same score if you measured again. Validity is whether you're measuring the thing you claim. A beautifully calibrated item bank attached to a contested construct is precision without validity. You've built a very accurate scale for a quantity that may not exist.



Corn

A laser measure pointed at a ghost.



Herman

There's a measurement-error angle too that I found unsettling. Someone framed it as two armed camps that ignore each other. The latent-factor people — the ones fitting capability models to benchmark results — ignore scaling laws, so the capabilities they extract often just proxy model size. You're measuring parameter count with extra steps. The scaling-law people ignore measurement error, so their fits are overfit to benchmarks that are themselves noisy and, as we just established, possibly not measuring anything.



Corn

Both camps treat the benchmark number as ground truth, and the benchmark number is the least trustworthy object in the room.



Herman

That's the bind. And the constructive proposal out of the validity camp is heavy. This year someone argued that establishing construct validity in LLM capability benchmarks requires nomological networks — the Cronbach and Meehl tradition — meaning you don't validate a benchmark by checking its items, you validate it by showing the construct relates lawfully to other observables. If reasoning is real, it should predict performance on tasks it's never been tied to, correlate with measures it should correlate with, dissociate from measures it shouldn't.



Corn

A web of predicted relationships, and the benchmark is just one strand.



Herman

One strand. Which is a much bigger lift than fitting a three-parameter model to your items. Item efficiency is a weekend of compute. A nomological network is a research program.



Corn

Here's what I want to ask you, Herman, because you spent a career on the clinical side of measurement. If you had to rebuild LLM evaluation from scratch, would you start with the items or with the construct?



Herman

The construct, and it's not close. In medicine you learn it the hard way — the measure everyone trusts because it's easy is usually measuring something adjacent to what matters. When ATLAS finds a quarter of models shifting ten rank positions, my first thought isn't wow, IRT is powerful. It's what were those ranks made of in the first place. But I'll confess the efficiency results still excite me, because a cheap test you can run a hundred times beats an expensive one nobody repeats. Repeated measurement is the only thing that ever exposed a bad instrument.



Corn

Agreed, with the caveat that a cheap, repeatable test of a contested construct just lets you be wrong faster and more often.



Herman

Noted, grudgingly accepted.



Corn

So where does that leave Daniel's question? He asked what psychometrics teaches about whether a benchmark measures what it claims. The honest answer I'm hearing is a two-parter. Item response theory tells you which questions are informative — it can prove your items are saturated, mislocated, undiscriminating. It cannot tell you whether the thing the questions aggregate into is real.



Herman

Right, and those are being conflated in the discourse right now. The efficiency program — ATLAS, TinyBenchmarks, the adaptive medical testing — optimizes item selection. The validity program asks whether the test should exist in its current form. Both matter and they are not the same conversation. You can absolutely run a forty-one-item adaptive HellaSwag that precisely measures... whatever HellaSwag measures.



Corn

HellaSwag measures the ability to complete sentences about everyday activities with plausible endings.



Herman

Which we call common sense reasoning when we're feeling generous.



Corn

I want to raise one more thing — this keeps nagging me. The item banks behind the GRE have an institutional home. Items get piloted, calibrated, retired. There's a whole profession whose job is item lifecycle. Who is the item doctor for MMLU?



Herman

Nobody. That's the institutional vacuum. MMLU was assembled, released, and adopted — fourteen thousand items, essentially no psychometric vetting, one paper's authors and then the whole field. There's newer work on detecting globally bad items — miskeyed answers, ambiguous wording, construct misalignment — and it finds plenty, but it's academic output, not maintenance. Nobody owns the bank.



Hilbert

Which of you two has actually written a multiple choice item?



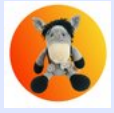
Herman

I've written exam questions for residents, if that counts.



Hilbert

It counts. Then you'll know the argument I lost in eighty-six. There was a reading comprehension bank we supplied items for, me and a shop in Watford I'd dealt with for years, and one item came back from review that I flagged. The passage was about tidal power, and the question asked for the main advantage of the barrage. The keyed answer said renewable. A boy who'd read nothing but the first line of the passage got it right, and a boy who'd read the entire business and knew the place flooded mudflats and killed the fish wrote that in the margin and chose the option about environmental damage. Wrong, said the key.



Herman

So the item measured who skimmed.



Hilbert

It measured who agreed with the setter. I said cut it or rewrite it. The editor said the item discriminated well — hard item, good slope, the numbers were fine. He was right about the numbers. The item survived because it was hard, not because it told you anything. I kept the proof sheet. Still know where it is. Herman, you explained difficulty as a location on a scale about ten minutes ago and it was the clearest thing said in here all afternoon, so you'll see the point. The fish boy knew more than the item did.



Corn

The item went into the final bank?



Hilbert

Ran for years. I saw it in a published practice paper in the nineties and wrote to the publisher. Got a leaflet back about their revision cycle. Corn, your GRE people retire items because a barrister makes them. Nobody made these people do anything, so nothing got done.



Corn

Nobody's ever been sued over a leaderboard position.



Herman

The fish boy's margin note is more peer review than most benchmark items have ever received.



Hilbert

The boy did fine. He passed everything. It was the test that was wrong, and it kept being wrong on schedule.



Corn

There's the whole institutional gap in one story. The item discriminated beautifully. It discriminated in favor of skimmers.



Herman

And that's the piece our whole efficiency discussion skipped — the adaptive machinery would have kept that item, because on paper it was high-information. The difficulty and slope numbers were good. Item response theory can't see the fish.



Corn

Which is the misconception, actually — worth naming it plainly before we go. The most common wrong belief about all this: that a higher raw accuracy score means a better model.



Herman

And the correction runs straight through what we covered. Models with identical accuracies get meaningfully different ability estimates once you weight the items by what they actually measure. A quarter of models shift more than ten rank positions. And that's before the validity problem — fewer than one benchmark in ten uses a complete real-world task, and nearly half the constructs are contested even by the people who define them. The number on the leaderboard is the least informative thing about the model.



Corn

So the open question I'm leaving with — if forty-one well-chosen items can stand in for five thousand six hundred, why is the whole industry still paying to run five thousand six hundred? Some of it's inertia. Some of it's that the variance question is unresolved. But some of it, I suspect, is that nobody wants to fund the boring institution — the item bank with a staff, a retirement schedule, a fish-boy department. We've known how to build longer rulers for decades. The question is whether anyone wants to pay for the calibration.



Herman

Producer credit to Hilbert Flumingtop, who has been quietly minding the desk and, we now know, a proof sheet from Watford for forty years. If you want more of this, try episode eleven, How Does Fine Tuning Work Anyway; and episode fourteen, AGI's Crossroads. This has been My Weird Prompts — no question too obscure, no rabbit hole too deep. If



Herman (continued)

today's episode recalibrated anything for you, the review button is right there in your podcast app.



Corn

See you soon, wherever you climb past the top of the scale.

SOURCES

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

1. **Zhou, Huang, Zhao et al., Lost in Benchmarks? Rethinking Large Language Model Benchmarking with Item Response Theory (PSN-IRT), arXiv:2505.15055v3 [primary]**

   retrieved 2026-11-08; submitted 2025-05-21, revised 2026-01-16; AAAI 2026 Oral.
   https://arxiv.org/abs/2505.15055v3

2. **Li, Tang, Chen et al., Adaptive Testing for LLM Evaluation: A Psychometric Alternative to Static Benchmarks (ATLAS), arXiv:2511.04689v3 [primary]**

   retrieved 2026-11-08; submitted 2025-10-26, revised 2026-07-13; ICML 2026 Spotlight, PMLR 306, 70246, 70269.
   https://arxiv.org/abs/2511.04689v3

3. **Bean, Kearns, Romanou et al., Measuring what Matters: Construct Validity in Large Language Model Benchmarks, arXiv:2511.04703v1 [primary]**

   retrieved 2026-11-08; submitted 2025-11-03; NeurIPS 2025 D&B Track.
   https://arxiv.org/abs/2511.04703v1

4. **Wikipedia, Item response theory**

   retrieved 2026-11-08; last edited 2026-10-25.
   https://en.wikipedia.org/wiki/Item_response_theory

5. **Madaan, Singh, Schaeffer et al., Quantifying Variance in Evaluation Benchmarks, arXiv:2406.10229v1**

   retrieved 2026-11-08; submitted 2024-06-14.
   https://arxiv.org/abs/2406.10229v1

6. **Mandujano Reyes, Laplace-PSN-IRT: Uncertainty Quantification for Neural Item Response Theory Models of LLM Benchmarks, arXiv:2607.25257v1**

   retrieved 2026-11-08; submitted 2026-07-28.
   https://arxiv.org/abs/2607.25257v1

7. **Lior, Frostig, Stanovsky et al., Extending Item Response Theory for Efficient and Meaningful Multilingual Evaluation, arXiv:2606.15643v1**

   retrieved 2026-11-08; submitted 2026-06-14.
   https://arxiv.org/abs/2606.15643v1

8. **Zheng, Jiang, Liu et al., Leveraging Computerized Adaptive Testing for Cost-effective Evaluation of Large Language Models in Medical Benchmarking, arXiv:2603.23506v1**

retrieved 2026-11-08; submitted 2026-02-28.
https://arxiv.org/abs/2603.23506v1

9. **Liao, Zhang, Wu et al., Toward a unified framework for data-efficient evaluation of large language models (LEGO-IRT), arXiv:2510.04051v1**

retrieved 2026-11-08; submitted 2025-10-05.
https://arxiv.org/abs/2510.04051v1

10. **Luo, Wu, Frisch et al., Measuring Competency, Not Performance: Item-Aware Evaluation Across Medical Benchmarks (MedIRT), arXiv:2509.24186v2**

retrieved 2026-11-08; submitted 2025-09-29, revised 2026-04-06.
https://arxiv.org/abs/2509.24186v2

11. **Uebayashi, Masui, Atarashi et al., Evaluating Cross-Modal Reasoning Ability and Problem Characteristics with Multimodal Item Response Theory (M3IRT), arXiv:2603.02663v1**

retrieved 2026-11-08; submitted 2026-03-03; ICLR 2026.
https://arxiv.org/abs/2603.02663v1

12. **Hardy, Gilbert, Domingue, Efficient Detection of Bad Benchmark Items with Novel Scalability Coefficients, arXiv:2603.24999v2**

retrieved 2026-11-08; submitted 2026-03-26, revised 2026-03-27.
https://arxiv.org/abs/2603.24999v2

13. **Alaa, Hartvigsen, Golchini et al., Medical Large Language Model Benchmarks Should Prioritize Construct Validity, arXiv:2503.10694v1**

retrieved 2026-11-08; submitted 2025-03-12.
https://arxiv.org/abs/2503.10694v1

14. **Kearns, Quantifying construct validity in large language model evaluations, arXiv:2602.15532v1**

retrieved 2026-11-08; submitted 2026-02-17.
https://arxiv.org/abs/2602.15532v1

15. **Freiesleben, Establishing Construct Validity in LLM Capability Benchmarks Requires Nomological Networks, arXiv:2603.15121v1**

retrieved 2026-11-08; submitted 2026-03-16.
https://arxiv.org/abs/2603.15121v1

16. **Berndt, Ralph, Bach et al., Construct Validity of Coding Benchmarks for LLM-based Agentic Coding Tools, arXiv:2611.05665v1**

retrieved 2026-11-08; submitted 2026-11-03.
https://arxiv.org/abs/2611.05665v1

17. **Zhao, Chen, Bao et al., A Measurement Perspective on Benchmarking LLM-Generated Text Detectors using Item Response Theory, arXiv:2610.24972v1**

retrieved 2026-11-08; submitted 2026-10-18.
https://arxiv.org/abs/2610.24972v1