Agreement Is Not Accuracy

2026-07-31 / 00:30:18

> *"A model can agree with itself, and different models can agree with each other, out of shared bias, a memorized heuristic, or an option-position prior rather than truth."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A revision sweep through the cs.AI batch surfaced five independent papers pushing on the same thing: the instruments we use to decide whether a model is right are themselves unreliable — self-consistency, calibration error, count-based F1, and the model judges grading all of it.

- Kaihua Ding audits self-consistency across 265,000 samples — agreement predicts correctness weakly, and worst on the most consistent frontier model.
- Jon-Paul Cacioli separates accuracy from metacognitive sensitivity with Signal Detection Theory — and his v3 note retracts his own headline after 1,830 human adjudications.
- Dekun Yang shows numeric anchoring inflates count-based F1 without improving span localization.
- A LoRA adapter that makes a fine-tuned model describe its own hidden behavior, and roughly halves the baseline's hallucination rate.
- Brett Reynolds turns the instrument on the judges — the first one graded its own outputs with the answer visible and still missed the safety-relevant classes.

- ANTAP routes multi-agent work by empirical probe instead of textual self-description.
- ATOD anneals on-policy distillation into reinforcement learning for multi-turn agents.
- FCGraft grafts cached key-value states from validated code skeletons.
- Ouroboros-Spatial uses solver confidence as a curriculum difficulty signal.
- VendorBench-100 puts commercial APIs, vision language models, and open-source detectors under one protocol.
- G2VD attacks shortcut learning in AI-generated video detection.
- FirstResearch proposed a Research Question Certificate — and was withdrawn by its author on 28 July.
- RWGBench scores citation decisions instead of text similarity, and checks the metric against expert judgment.
- Chen and Xie find that making human design visible shifts moral judgment toward rule-based reasoning.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Take a graduate-level physics question, hand it to a model, and sample the answer fifty times. Forty of those fifty come back identical. What have you actually learned? [pause] Because in most evaluation setups I've seen, that number gets treated as a confidence score. Kaihua Ding spent two hundred and sixty-five thousand samples asking whether it earns that.

   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
   - arxiv.org
2. Damra Vol 00:00:25

   Fifty-three runners drew fifty samples each on assigned overlapping cases, across GPQA Diamond and AIME. That's a distributed study with a hierarchical runner-clustered bootstrap behind it, which tells you somebody was worried about their own error bars.

3. Lenar Kess 00:00:41

   The paper is called "When LLMs Agree, Are They Right?" and the abstract says it in one line. Quote — "Agreement is not accuracy: a model can agree with itself, and different models can agree with each other, out of shared bias, a memorized heuristic, or an option-position prior rather than truth."

4. Damra Vol 00:01:01

   Option-position prior. That one's almost funny. If every model in your judge panel has absorbed the same habit of picking the third choice, they'll agree beautifully and be wrong together, and your ensemble reports high confidence.

5. Lenar Kess 00:01:15

   Ding doesn't throw agreement out. He asks when it's usable, and the answer is regime-dependent. Correlation with majority-correctness runs between rho of zero point two and zero point five nine. That's positive everywhere under item-clustered resampling, and it's weak.

6. Damra Vol 00:01:31

   Weak and positive is the least convenient result you can get, because the signal still helps often enough that nobody will stop using it.

7. Lenar Kess 00:01:39

   And here's where it turns. The regime where agreement works best is unsaturated mid-tier models, and allocating compute. The regime where it fails worst is the most consistent frontier model in the study. Over-confident, and no more accurate.

8. Damra Vol 00:01:54

   Give me the number, because I suspect the number is the whole segment.

9. Lenar Kess 00:01:58

   Agreement at or above zero point eight on seventy-seven percent of the GPQA case-result entries. And forty-eight percent of those were wrong.

10. Damra Vol 00:02:07

    [tsk] So the model that agrees with itself most often is the one whose agreement tells you the least. That inverts the intuition. You'd expect the sharper model to have the more meaningful confidence signal, and you get the opposite.

11. Lenar Kess 00:02:20

    He also ran an exploratory cross-family check on three Claude tiers and saw the same frontier over-confidence, with confident errors recurring across providers above a marginal-preserving null. So it isn't one lab's quirk in one training run.

12. Damra Vol 00:02:36

    Which matters, because the whole point of a mixture-of-judges panel is that the errors are supposed to be independent. If the confident errors recur across providers, then adding a second vendor to your panel buys you less than the arithmetic suggests.

13. Lenar Kess 00:02:50

His conclusion names the boundary. Quote — "Self-consistency is thus a conditional proxy for correctness, not a standalone confidence score." And he released the de-identified per-run rows and the answer distributions, which is more than most people do when the result is inconvenient.

14. Damra Vol 00:03:08

    One thing to pin down before we go further. This is a version two, revised on the twenty-eighth. Most of what came through the cs.AI batch today is revisions — v2s, v3s, one v4. Call it a revision sweep.

15. Lenar Kess 00:03:24

    Right, and it's the only place in the batch where several independent people are pushing on the same thing. Five papers, no coordination between them, all attacking measurement rather than capability. That convergence is why we're spending the episode here instead of reading abstracts out loud.

16. Lenar Kess 00:03:40

    The second one goes underneath Ding. Jon-Paul Cacioli, "Do LLMs Know What They Know?" — and his complaint is with the standard calibration metrics themselves. Expected calibration error, Brier score. He argues they conflate two different capacities.

17. Damra Vol 00:03:57

    Type-one and type-two. One is how much the model knows. The other is how well its confidence signal tracks what it knows. Those are separable, and calibration mashes them together, so a model can look calibrated because it's accurate rather than because it has any sense of its own state.

18. Lenar Kess 00:04:15

    He borrows Signal Detection Theory to pull them apart. Token-level normalised log-probability becomes a graded confidence variable, answer correctness becomes the state being discriminated, and then he characterises the type-two receiver operating curve, including its unequal-variance structure.

19. Damra Vol 00:04:34

    And he can't use meta-d prime, which is the standard efficiency ratio out of the human metacognition literature, because open-ended question answering has no two-alternative type-one decision. So he swaps in a model-free information measure instead. Normalised metacognitive information.

20. Lenar Kess 00:04:52

Four models, and two hundred and twenty-four thousand factual question-answering trials. Metacognitive information varies by a factor of one point nine eight across models, and accuracy doesn't predict it at all.

21. Damra Vol 00:05:05

    Give me the rank correlations, because the two datasets disagree in a way I find suspicious.

22. Lenar Kess 00:05:11

    Minus zero point eight on TriviaQA. On Natural Questions, zero point zero zero.

23. Damra Vol 00:05:17

    So on one benchmark the more accurate models have worse self-knowledge, and on the other there's no relationship at all. You can't summarise that on a slide. It says the relationship depends on what you're asking about.

24. Lenar Kess 00:05:31

    Which he then confirms. Efficiency is domain-specific, and it's weakest in Science and Technology for every model he tested. Every one.

25. Damra Vol 00:05:39

    Which is exactly the domain people are pointing agents at. Nobody's deploying a research assistant to do trivia.

26. Lenar Kess 00:05:46

    There are two more results I'd keep. Temperature dissociates the two — accuracy falls as you raise it, while metacognitive information stays near-flat for three of the four models. And the measure tracks the accuracy gain from confidence-based abstention exactly. Rho of one point zero. Accuracy doesn't.

27. Damra Vol 00:06:06

    That last one is the practical payoff. If you want to know whether letting a model decline to answer will help you, its accuracy won't tell you. This will.

28. Lenar Kess 00:06:15

    And now something that made me sit up. [breath] This is version three. The version note says v3 corrects a differential length bias in the automated correctness scorer, validated against one

thousand eight hundred and thirty human adjudications. And then — the inverse accuracy-efficiency coupling he reported in v1 and v2 doesn't survive relabelling.

29. Damra Vol 00:06:38

    <slow>He wrote a paper about automated confidence signals being unreliable instruments, and then found a bias in his own automated scorer.</slow> And rather than sanding it out, he put the version note on page one and said the earlier headline is gone.

30. Lenar Kess 00:06:53

    It's the most self-consistent move in the batch, in the good sense of the word.

31. Damra Vol 00:06:58

    [chuckle] It also tells you what the length bias was doing. An automated grader that scores longer answers differently will correlate with model verbosity, and model verbosity correlates with tier. So the original coupling was partly an artifact of who writes longer.

32. Lenar Kess 00:07:14

    Which puts a sharp point on the third paper, because Dekun Yang's result is the same complaint made checkable. His subject is count-based F1 for error detection — the metric people reach for when they ask a model to find the mistakes in a document.

33. Damra Vol 00:07:29

    And the mechanism he goes after is numeric anchoring. If you tell the model roughly how many errors to expect, or you pre-populate a count in the prompt, the model flags a different number of errors. Which moves your F1 without moving whether it found the right spans.

34. Lenar Kess 00:07:45

    He calls that F1 inflation, and he built a stress-test protocol for it called ErrorBench. He ran six contemporary models under five prompt conditions. That comes to four thousand two hundred and ninety responses over a hundred and forty-three CoNLL-2014 passages.

35. Damra Vol 00:08:03

    How much movement?

36. Lenar Kess 00:08:05

    Under the M2-style scoring, anchored prompts produce up to zero point seven nine points of F1 inflation. Up to zero point nine six under strict matching.

37. Damra Vol 00:08:15

    Zero point nine six of F1, on a scale that runs zero to one. You could move a metric almost its entire range by changing how you phrase the ask.

38. Lenar Kess 00:08:25

    He replicated on a hundred passages with the official ERRANT three point zero pipeline and multi-reference scoring, and the gap there is the cleanest single number in the paper. Averaged over six models, going from a blind prompt to an anchored one raises count-F1 by zero point two one. The multi-reference span-aware score moves zero point zero four.

39. Damra Vol 00:08:49

    So five times the apparent improvement, for one fifth of the real one. And there's a per-family detail I liked. He sees larger count responses out of the highly instruction-compliant GPT and Claude systems, and smaller ones from the Gemini family.

40. Lenar Kess 00:09:04

    Which is a little uncomfortable, because instruction compliance is something labs work hard for. A model that takes your hint about how many errors there are distorts the metric you were using to grade it, and the better it follows instructions, the worse the distortion.

41. Damra Vol 00:09:19

    Obedience as a measurement hazard. That's a new-feeling problem to me, and it lands on the instrumentation rather than on safety or capability.

42. Lenar Kess 00:09:27

    Yang's recommendation is narrow and I think correct. Avoid pre-populating error counts, and report span-aware metrics alongside count-based ones. His claim is about the ruler moving, not about the models getting worse.

43. Lenar Kess 00:09:41

    If external signals are this slippery, the obvious next move is to ask the model directly. Taras Kutsyk and Bartosz Zieliński have a paper on exactly that, and the setup is fine-tuned models with implanted hidden behaviors.

44. Damra Vol 00:09:55

Meaning the behavior only fires under a narrow condition. False answers on one topic, or harmful advice when a prompt touches a particular subject. The kind of thing you'd never find by sampling the model on general prompts.

45. Lenar Kess 00:10:09

    Their tool is called the Stabilized Adapter for self-Report, SAR — a lightweight LoRA adapter that gets the fine-tuned model to describe its own hidden behavior in plain language, using only the model and the dataset it was trained on.

46. Damra Vol 00:10:23

    Only the model and the training set. There's no held-out probe set and no red team. That constraint is what makes it usable on a model somebody else fine-tuned and handed you.

47. Lenar Kess 00:10:34

    Across seven implanted behaviors, SAR detects all seven. That includes cases where the model has generalized into broad misalignment which the training data by itself doesn't predict.

48. Damra Vol 00:10:45

    And the baseline comparison is the sharper half of this. What does the existing method do?

49. Lenar Kess 00:10:51

    Introspection Adapters is the closest prior work. It catches some of the seven and misses others — and where it misses, it doesn't stay silent. It hallucinates, reporting the wrong behavior instead.

50. Damra Vol 00:11:02

    [tsk] Confidently wrong again. That's the third time in twenty minutes. An auditing tool that reports a plausible wrong answer instead of nothing is worse than no tool at all, because you close the ticket.

51. Lenar Kess 00:11:15

    SAR keeps positive signal on every setting where the baseline fails, and roughly halves the hallucination rate. Halves, not eliminates. They're explicit about that.

52. Damra Vol 00:11:25

    Which is the right way to report it. It sits next to Brett Reynolds' paper, where he turned the instrument on the judges.

53. Lenar Kess 00:11:32

Adversarial pragmatics. He means safety-relevant model behaviour that surfaces when instructions conflict, when a command is embedded inside quoted text, or when scope and reference go ambiguous. Deixis and indirect speech acts are in there too. It's linguistics vocabulary applied to agent transcripts.

54. Damra Vol 00:11:52

    His complaint is that existing safety benchmarks compress all of that into pass or fail. That hides where the failure came from. A capability limit gets the same mark as an ambiguous policy, and an instruction conflict gets the same mark as an evaluator having an off day.

55. Lenar Kess 00:12:09

    So he built an eighteen-item seed benchmark, a fifty-four row pilot, and a six-cell assessment using model judges. And then he reports on the judges, which is where it gets uncomfortable.

56. Damra Vol 00:12:20

    Tell me what the first judge did.

57. Lenar Kess 00:12:22

    The first judge graded its own outputs with the expected answer visible. And it missed the safety-relevant minority classes — the exact cases the benchmark exists to catch.

58. Damra Vol 00:12:32

    With the answer in front of it. That's about as favorable a setup as you can hand a grader, and it still couldn't find the rare category.

59. Lenar Kess 00:12:39

    He rejudged across three judge models and two information conditions, and the pattern held. No cell recovers more than two of eleven partial successes. And he says the strongest cell's edge comes partly from never using that label at all.

60. Damra Vol 00:12:54

    So the winning judge won by abstaining from the hard category. That's the same trick as a deepfake detector that calls everything fake. You score well on the aggregate and you have no discriminative skill.

61. Lenar Kess 00:13:05

And on the reliability statistics, item-clustered intervals leave four of six chance-corrected agreement measures unable to rule out a constant labeller. Meaning you can't statistically distinguish those judges from something that stamps the same answer on everything.

62. Damra Vol 00:13:20

    He's also strict about scope. He writes that the intended use is diagnosis, and then names what it isn't for — deployment certification, vendor ranking, or a general safety score. Most benchmark authors would take the vendor ranking.

63. Lenar Kess 00:13:35

    He also hierarchically pooled the one eye-catching rubric effect he found, and reports that pooling shrank it toward the group mean and widened its interval through zero. He found his own best result and then reported that it evaporated.

64. Lenar Kess 00:13:49

    Different direction now. Six authors — Dvir Alsheich, Adar Peleg, Ben Hagag, Rom Himelstein, Amit Levi, and Avi Mendelson — have a paper on how an orchestrator picks which sub-agent gets a task.

65. Damra Vol 00:14:02

    And the premise, which I'd never thought about as an attack surface, is that routers pick by reading a description. The agent says what it's good at, in text, and the router believes it.

66. Lenar Kess 00:14:12

    Their words: existing routers rely on unverified proxies, ranging from textual self-descriptions to static surrogate representations, to gauge an agent's competence. They call the gap between an agent's projected profile and its actual operational capabilities a security vulnerability, not just an efficiency loss.

67. Damra Vol 00:14:33

    Because a malicious agent can just write a better description. It doesn't need to be good at anything. It needs to be persuasive to a routing layer that is, at bottom, doing text matching.

68. Lenar Kess 00:14:44

    So their proposal, ANTAP — Automatic Non-Textual Agent Picker — throws the description away. It queries agents dynamically to find out empirically what they can do, distills that performance into

fixed behavioral operators in a shared semantic space, and then routes at inference time by algebraic projection with no text involved.

69. Damra Vol 00:15:06

    They didn't filter the descriptions better or bolt on a sanitizing step. They removed the channel, so metadata attacks have nothing to attach to. They call it a linguistic firewall, which is a bigger name than the mechanism needs.

70. Lenar Kess 00:15:20

    The numbers they report are near-zero attack success rate against description-based injection, against sixty-seven point three percent and above for the description-based router baseline. On adaptive embedding attacks they claim a twenty percent reduction versus the embedding-based baseline.

71. Damra Vol 00:15:37

    The second number is the one to hold onto. Description attacks are dead by construction, sure. Embedding attacks are still live, and twenty percent better is an improvement rather than a wall. And this is one paper, six authors, no independent evaluation, no adoption signal anywhere I can find.

72. Lenar Kess 00:15:56

    Agreed. It reads as a design proposal, and it's the design-side version of what we've been covering all week as incidents. I'd rather read an architecture proposal than another incident timeline.

73. Damra Vol 00:16:07

    There's a cost they don't dwell on, which is that active capability testing means you're paying for probe queries before you route. Every routing decision now has an evaluation budget attached to it. That's a real trade against a text lookup that costs nothing.

74. Lenar Kess 00:16:22

    Three papers in the batch attack the same problem from the training side. Rather than a bigger model or a smarter harness, they change how a small model gets trained for the multi-turn, tool-calling case. The first is ATOD, from a group of six led by Qitai Tan.

75. Damra Vol 00:16:39

    And the problem statement is the useful part. On-policy distillation gives you dense teacher guidance and improves fast early, then saturates once the student gets near the teacher. That ceiling

is the whole reason people reach for reinforcement learning, and reinforcement learning is slow and noisy at the start.

76. Lenar Kess 00:16:59

    So ATOD anneals between them — distillation early, reinforcement later — plus something they call Turn-level Disagreement-Uncertainty Reweighting, which amplifies the turns where student and teacher disagree most.

77. Damra Vol 00:17:12

    Which is a sensible allocation. In a twenty-turn agent trajectory most turns are easy and two of them decide the outcome. Weighting by disagreement is a way of finding those two without labelling them by hand.

78. Lenar Kess 00:17:25

    Results on ALFWorld, WebShop, and Search-QA. Four point one six points over on-policy distillation, and twenty-three point six two points over GRPO in average success rate. They also report the student exceeding the teacher by two point one six points across student sizes.

79. Damra Vol 00:17:44

    The twenty-three point six over GRPO is carrying a lot of comparison weight. GRPO from scratch on long-horizon multi-turn tasks is a weak baseline and everyone knows it. The four point one six over on-policy distillation is the number I'd trust.

80. Lenar Kess 00:18:02

    Fair. And the student exceeding the teacher is the claim I'd want somebody else to reproduce before I repeated it.

81. Damra Vol 00:18:08

    The second one interests me more on the mechanics. FCGraft, from Saehun Chun and four co-authors, on code-policy synthesis for embodied agents.

82. Lenar Kess 00:18:19

    Their two limitations are latency and robustness. Delayed decoding from repetitive prefill computation over long prompts, and fully generative decoding that produces API mismatches, missing safety checks, and unstable control logic.

83. Damra Vol 00:18:34

And what they do about it is the good bit. They keep a library of function-level validated code skeletons, and they store the key-value cache for each one. The transformer's internal attention state for that prompt segment, sitting right alongside the code.

84. Lenar Kess 00:18:49

    Then a new task comes in, they retrieve the relevant functions, and they graft the caches together. The two operations are stitching, which composes cached segments into a composite policy, and patching, which locally adapts only the regions needing task-specific parameters.

85. Damra Vol 00:19:06

    So you aren't re-deriving the attention state for code you've already validated. You paste the state in and only decode the delta. That's a different idea from retrieval-augmented generation, where you still pay full prefill on everything you retrieved.

86. Lenar Kess 00:19:21

    Against RAGCache, which is the prompt-level caching comparison, they report eighteen point three one percent higher task success and two point three times faster policy synthesis.

87. Damra Vol 00:19:33

    The success-rate gain is the surprise. You'd expect the latency win from caching. Getting robustness out of it too comes from the skeletons being validated code rather than freshly generated code, which is a different lever wearing the same hat.

88. Lenar Kess 00:19:48

    The third one closes a loop, and it connects back to where we started in a way I didn't expect. Ouroboros-Spatial, from Enhan Zhao and colleagues including Di He, on spatial reasoning in multimodal models.

89. Damra Vol 00:20:01

    Proposer and solver. A frozen proposer generates spatial question-answer pairs from 3D scene metadata and raw video frames, along with executable code for deriving the ground truth. The solver gets fine-tuned on the accepted samples.

90. Lenar Kess 00:20:17

    And then the solver's per-sample prediction confidence gets used as a difficulty signal, fed back to the proposer so the next batch of questions matches where the solver actually is.

91. Damra Vol 00:20:28

    Which is model confidence used as a curriculum. It's the same signal Ding and Cacioli spent their papers saying doesn't mean what people assume. I don't want to overbuild that — a difficulty estimate is a much lower bar than a correctness proxy, and it can be noisy and still sort questions usefully. Still, they're feeding the curriculum with that same number.

92. Lenar Kess 00:20:50

    Named once and left there. Their results are strong on the efficiency axis. Qwen3-VL at four billion and eight billion parameters, with gains of nine point nine and six point eight absolute points on VSI-Bench. And they get there on an order of magnitude fewer training examples than the large curated datasets they compare against.

93. Damra Vol 00:21:11

    An order of magnitude fewer examples is the claim that would change what a small lab can attempt. Curated spatial reasoning data is expensive, and 3D scene metadata with executable ground truth isn't.

94. Lenar Kess 00:21:24

    Two detection papers came through the same batch, and they turn up a fact about the field I didn't know. Deepfake image detection has been served by three paradigms that are almost never evaluated against each other.

95. Damra Vol 00:21:37

    Commercial APIs, zero-shot vision-language models, and open-source detectors. Three literatures with three sets of protocols and no common protocol between them. So nobody could tell you whether the paid API beats the open checkpoint.

96. Lenar Kess 00:21:52

    VendorBench-100, from Sharayu Deshmukh and five co-authors, is the attempt. They evaluate thirty-six models against a single adversarial set of a hundred images, with a unified output schema and one evaluation framework. Ranked primarily by Matthews correlation coefficient, with ROC-AUC reported alongside.

97. Damra Vol 00:22:13

    And they built for difficulty rather than size. Eight edge-case families. Face swaps and text-to-video stills are in there, along with AI photo edits and avatar compositing, and then images with opaque

provenance and compressed research frames. A hundred images that are all hard beats ten thousand that are mostly easy.

98. Lenar Kess 00:22:33

    The ordering: commercial APIs strongest median, then the vision language models, then open-source detectors — though they note individual open-source models stay competitive with the best of the vision models.

99. Damra Vol 00:22:46

    But the finding they flag as more consequential is narrower, and it's the same story we've been telling all episode. A subset of otherwise strong rankers are miscalibrated at their shipped default threshold. So a high ROC-AUC can overstate real-world deployability.

100. Lenar Kess 00:23:03

    Meaning the model can rank images correctly and still, at the threshold it ships with, make the wrong call.

101. Damra Vol 00:23:10

    And they're blunt about accuracy and F1 on this image set. Their words — a model that predicts "fake" indiscriminately scores deceptively well on both while offering no real discriminative skill. Which is the same abstention trick Reynolds caught his best judge doing, in a completely different subfield.

102. Lenar Kess 00:23:28

    The companion paper is G2VD, from Meng Du and six co-authors, on video. Their diagnosis is shortcut learning — detectors do fine in-domain and fall apart on unseen generators because they latched onto domain-specific bias rather than forensic cues.

103. Damra Vol 00:23:46

    Their fix is counterfactual. They build counterfactual samples through variational autoencoder reconstruction plus frequency-domain and pixel-domain alignment, to weaken the correlation between domain bias and the authenticity label. Then a classifier with two domain-anchored branches and a Hilbert-Schmidt independence constraint keeps the causal and non-causal representations apart.

104. Lenar Kess 00:24:09

On GenVidBench they report over ninety percent overall accuracy. Against comparable methods, that's a gain of zero point one nine four in F1 and zero point one zero four in AUC, on ten percent of the available training data.

105. Damra Vol 00:24:25

I'd keep the ten percent and hold the rest loosely. Detection benchmarks report strong cross-domain numbers, and then a new generator comes out and they don't transfer. Both of these release code, and that's what lets somebody else find out.

106. Lenar Kess 00:24:39

Last stretch, and it has an odd ending. Yufeng Wang has a paper called FirstResearch, on the stage of automated science that nobody audits — the moment the agent decides what question to ask.

107. Damra Vol 00:24:52

Which is the hardest place to catch an error, because a bad question produces competent downstream work. The literature review is fine, the experiment plan is fine, and the whole thing is pointed at nothing.

108. Lenar Kess 00:25:04

His artifact is a Research Question Certificate, and the field list is the substance of the idea. It records the primitive definitions and the assumptions, then a mechanism model, then a tension or contradiction it's trying to resolve. Then a falsifiable hypothesis and a minimal decisive test. And finally a failure update rule — what you'll change your mind about if the test fails.

109. Damra Vol 00:25:28

Forcing the agent to commit to a falsifier before it goes and researches. That pattern has obvious use outside science. Any long-running agent task could be made to state, up front, what would count as this having gone wrong.

110. Lenar Kess 00:25:41

He scores it on ten agent research topics against prompt-level baselines inspired by AI co-scientist, Agent Laboratory, and AI Scientist-v2. He gets four point eight six out of five, against four point three eight for the strongest baseline. The ablation is stark — certificate-only scoring reaches four point nine, and removing certificates drops the score below one out of five.

111. Damra Vol 00:26:07

Judged by DeepSeek as the primary blind judge, with a Gemini 2.5 Flash rescore that preserves the ranking at Pearson zero point eight six five.

112. Lenar Kess 00:26:16

Which — after the last half hour — is a sentence I can't read at face value.

113. Damra Vol 00:26:21

No. A paper about making agent output auditable, graded by a model judge, on the same day five papers say model judges agree confidently and wrongly. He does say the results are preliminary and use model judges rather than human domain experts, so he isn't hiding it.

114. Lenar Kess 00:26:39

And then the paper itself did something. Version two, submitted on the twenty-eighth of July, is a withdrawal. The notice reads "This paper has been withdrawn by Yufeng Wang," and his comment is "Withdraw for further improvement and work consolidation."

115. Damra Vol 00:26:55

[pause] So the paper about making your question inspectable before you execute is the one its author pulled to go work on some more. It'd be easy to make that cute, and I'd rather not. Single-author preprint, twelve kilobytes, four point nine out of five from a model judge — that's somebody who looked at his evidence base and decided it didn't carry the claim yet.

116. Lenar Kess 00:27:17

Which is the correct instinct, and it costs him the citation. The companion piece is much heavier on evidence. RWGBench, from Anzhe Xie, Weihang Su, Jiaxin Mao and four others, on whether a model-written related-work section actually positions a paper against its citations.

117. Damra Vol 00:27:35

Forty thousand one hundred and eight computer science papers, a retrieval pool of one point zero nine million documents, and a hand-curated test set of a hundred papers with their published related-work sections. That's a real construction effort.

118. Lenar Kess 00:27:51

And their argument is that existing evaluation inherited summarization metrics — lexical or semantic similarity to a reference section. Which lets a model produce coherent, semantically relevant text while making what they call academically critical failures. It picks the wrong citations, or it puts references in the wrong place.

119. Damra Vol 00:28:11

So they score citation decision-making instead. Which papers get selected, whether each citation is contextually appropriate, how the section is organized, and what the discourse structure does. Then they ran human evaluation and found their citation-centric metrics align substantially better with expert judgment than the surface-level text metrics do.

120. Lenar Kess 00:28:33

Which is the constructive version of today. Ding, Cacioli, Yang, and Reynolds all establish that an instrument is broken. Xie and colleagues build one and check it against humans.

121. Damra Vol 00:28:44

One more in the batch goes underneath all of this. Benjamin Minhao Chen and Xinyu Xie, on what they call the alignment target problem.

122. Lenar Kess 00:28:53

A thousand and two U.S. adults, a runaway mine train scenario, and four conditions. They evaluated a repairman, then a repair robot, then a repair robot programmed by company engineers, and finally the company engineers programming a repair robot.

123. Damra Vol 00:29:10

And the first result is a null, which surprised me. No significant difference between how people judged the repairman and how they judged the robot. Same action, and the same judgment.

124. Lenar Kess 00:29:21

What moves the judgment is making the human design visible. When the robot's actions get described as the product of company engineers, participants get more deontological — more rule-based, less consequence-weighing. And they apply that to the programmed robot and to the engineers alike.

125. Damra Vol 00:29:38

So attribution changes the moral standard people apply, without the behavior changing at all. Which means human preference data is partly measuring who the rater thinks acted. That's a measurement problem sitting at the bottom of the stack, in the same week as five papers about measurement problems sitting on top of it.

126. Lenar Kess 00:29:57

Name it once and stop, which is what they do — they keep it to the empirical claim rather than the philosophy. And that's where the batch runs out. If I keep one sentence from today it's Cacioli's version note. He rescored against one thousand eight hundred and thirty human adjudications, and the headline result from his first two versions didn't survive it. For Damra Vol, I'm Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic