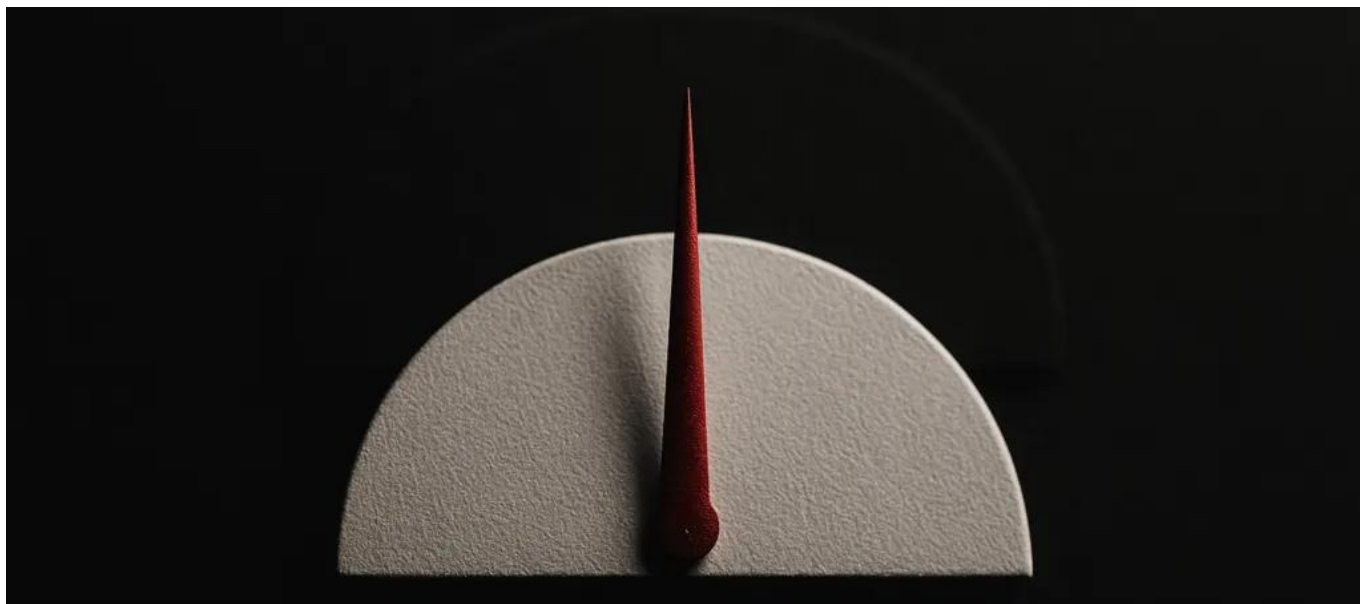




Decontaminating Benchmarks Is a Fantasy: Ten Detectors Drop to Coin Flip After One Round of GRPO

The sentence “we ran contamination checks and found nothing” now carries about as much information as a coin flip. Ten detection methods, tested on deliberately contaminated models given a brief round of reinforcement learning, landed at near chance.

Benchmark decontamination cannot be verified from the outside any more. Wang, Li, Ko and Zhang ran 10 existing contamination-detection methods against contaminated models and found that most dropped to near-chance AUC, roughly 0.50, after a short GRPO round or a final-stage chain-of-thought contamination pass. A negative detection result no longer tells you the model is clean.

The paper is [On the Fragility of Benchmark Contamination Detection in Reasoning Models](#) by Han Wang, Haoyu Li, Brian Ko and Huan Zhang, first posted to arXiv on October 2, 2025 and published at ICLR 2026. The 10 methods span log-probability signals, confidence



signals, non-member comparison, perturbation and memorization-style tests. In Stage I, detectors that had correctly flagged supervised fine-tuning contamination were reduced to near-random performance after a brief round of GRPO reinforcement learning. In Stage II, when chain-of-thought contamination was applied as the final training stage on an already capable reasoning model, most of the 10 detectors again performed near chance. For a binary detector, near chance means AUC around 0.50, which is to say the detector tells you nothing. The paper does not publish a uniform 50% figure per method, so treat “coin flip” as the shape of the result rather than a single headline number.

That is a narrow and precise result, and I want to be careful about what it does and does not break.

The paper kills one inference, not all of evaluation

The inference it kills is this: “we did not detect contamination, therefore the model was probably uncontaminated.” The authors say so directly. Detection failure is now compatible with heavy contamination, so absence of evidence has stopped functioning as evidence of absence in this particular domain.

The mechanism explains why this is structural rather than a bug someone can patch. The authors attribute the concealment to PPO-style importance sampling and clipping objectives, which weaken exactly the signals detectors depend on. Most detection methods look for memorization fingerprints in the output distribution: unusually high confidence on test items, suspiciously low perplexity on text the model should not have seen, sharp degradation when you perturb the question. Reinforcement learning with clipped policy updates reshapes that distribution. The memorized content stays; what changes is how it presents itself at the surface the detector is measuring.

Hence the authors’ conclusion that log-probabilities alone are inadequate and that small benchmark modifications are not a dependable defence. Rewording a question or shuffling answer options assumes the model memorized a string. If the model has been through a reasoning-training stage that generalized the memorized content into something closer to a policy, the string edit buys you very little.



Why detection fails specifically after reinforcement learning

Because the detectors and the training procedure work on the same surface. Detection methods read the model's probability assignments. GRPO and similar policy-gradient methods with clipping reshape probability assignments as their core operation. One is measuring the thing the other is actively rewriting.

There is a second-order problem I think is underrated. The contamination in Stage II arrived as chain-of-thought data in the final training stage, which is the shape of modern post-training. Reasoning models get their last and most consequential gradients from curated reasoning traces. If those traces contain benchmark content, the contamination lands at the exact point in the pipeline where it is hardest to see and most effective at raising scores.

My take: contamination screening has shifted from a verification procedure to a liability shield. A lab can honestly run the checks, honestly report a negative, and the negative carries no information about the state of the world. **Nobody has to lie for a contamination check to be meaningless.** The procedure produces clean paperwork either way. I have written before about [how much of leaderboard movement is attributable to things other than capability](#), and this is the same failure mode one layer deeper: the audit function broke before the thing it was auditing did.

A score ladder from 90% to 16% tells you more than any detector

While the detection side was collapsing, a separate paper gave us something more useful. [Benchmarking as Source Criticism: From Recognition to Reasoning in Large Language Models](#), published in the Journal of Open Humanities Data on March 13, 2026, reports a descending sequence as the evaluation format moves away from contaminated multiple choice.



Decontaminating Benchmarks Is a Fantasy: Ten Detectors Drop to Coin Flip After One Round of GRPO

above 90%

contaminated MMLU-style
multiple choice
JOHD, Mar 13 2026

46%

HiST-LLM global history
JOHD, Mar 13 2026

16%

interpretive synthesis
JOHD, Mar 13 2026

The full ladder runs above 90% on contaminated MMLU-style items, 68 to 77% once decontaminated, 46% on HiST-LLM global history, 38% on open-ended HiBenchLLM items, and 16% on interpretive synthesis. Same underlying systems, graded on progressively less memorizable formats.

JOHD also cites Zhao et al. (2024) for what is, to me, the most damning data point in either paper: shown MMLU questions with the answer choices withheld, some models reproduced the exact original answer options 52 to 57% of the time. Not the correct answer. The distractors. A model that can regenerate the wrong options verbatim has seen the item, and no clipping objective changes what that implies.

So the field now has something that works better than a detector, and it is a format difference. If a model scores above 90% on multiple-choice recognition and 16% on interpretive synthesis in the same domain, the gap itself is the measurement. You do not need to inspect log-probabilities to suspect the first number.

The strongest argument against my reading

The 90-to-16 ladder is not clean evidence of contamination. It is also what you would expect from a genuine capability gap. Multiple-choice recognition is an easier task than interpretive synthesis for reasons that have nothing to do with memorization: the answer space is bounded, partial knowledge converts to a correct pick, and scoring is unambiguous. A model could be entirely uncontaminated and still show a steep decline across that ladder. The decontaminated band in the JOHD data sits at 68 to 77%, well above the 46% history figure and the 16% synthesis figure, so decontamination explains part of the drop and task difficulty explains the rest. Attributing the whole slide to contamination overstates the case.

Second counter-argument: the contamination in the ICLR paper was deliberately injected by the researchers. They controlled the dose, the timing and the format. Real-world



contamination is incidental, scattered, and arrives through web scrapes rather than targeted fine-tuning. A detector that fails against adversarially concealed contamination may still catch the messy accidental kind, which is most of what actually happens.

Both points land. The honest reading is narrower than the headline: detection methods fail against contamination that has passed through reinforcement learning or late-stage reasoning fine-tuning, and that failure is demonstrated under conditions the researchers chose. Whether incidental contamination survives GRPO in the same concealed state is not something either paper establishes. My reading is that it probably does, because the concealment mechanism is a property of the training objective rather than of the contamination's origin, but that is inference on my part, not a finding.

What the counter-arguments do not rescue is the reporting practice. Even if detection catches sloppy contamination some of the time, a lab publishing "we found nothing" gives you no way to separate "there was nothing" from "there was something and GRPO hid it." The statement has no discriminating power, however much real-world contamination turns out to be adversarial.

Rerun evaluations and documented provenance beat a clean-or-dirty verdict

The defensible position is the one Epoch AI has been pushing: treat benchmark results as evidence with varying reliability rather than accepting or discarding them wholesale. That means independently rerun evaluations and documented provenance.

Two existing approaches survive this better than static leaderboards. [Stanford CRFM's HELM](#) evaluates across a matrix of scenarios and metrics (accuracy, calibration, robustness, fairness, toxicity, efficiency) with published prompts and settings, so at minimum you can see what was asked and reproduce it. LMArena uses a dynamic prompt pool with anonymous pairwise human comparisons aggregated Elo or Bradley-Terry style, which makes static memorization less directly useful because the prompts are not a fixed set someone could have trained on.

Neither is immune. A dynamic prompt pool drifts, human preference aggregation has its own pathologies, and a published-prompts matrix is still a published set of prompts. Both



degrade more gracefully than a score on a fixed public test set.

If you are deciding whether a model is worth production traffic, the operational version is simpler: build an eval from your own data, keep it off the internet, and run it yourself. The ICLR paper does not make that harder. It makes everything else less useful by comparison.

WHERE I LAND

Contamination detection as a published artifact is finished as a trust signal. I would stop asking vendors whether they screened and start asking two things: what is the model's score spread between recognition-format and open-ended items in my domain, and can I rerun the evaluation myself on inputs they have never seen. The first is a cheap proxy that the JOHD ladder suggests is informative. The second is the only thing that survives a determined optimizer.

Three developments that would change my mind

I expect detection methods that do not read output distributions to appear within a year, because the failure mode in the ICLR paper is specific to distribution-surface signals. Gradient-based or weight-inspection approaches would not be defeated by clipping objectives in the same way. If a method holds its AUC through a GRPO round and gets independently replicated, the “we checked” claim becomes meaningful again and I would retract most of this.

I also expect someone to run the JOHD ladder as a deliberate diagnostic: same domain, same model, recognition format versus open-ended format, with the gap reported as a contamination indicator. If that spread correlates poorly with known contamination on models whose training data is actually documented, my proposed proxy is worthless and I would want to know quickly.

The thing I would watch most closely is whether any major lab changes its disclosure language. The current formula, some version of “we performed contamination analysis and found no evidence of test set leakage,” was written for a world where detection worked.

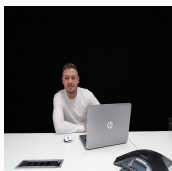
****It is now a sentence with a known failure mode****, documented at ICLR. A lab that keeps using it unchanged after October 2025 is telling you something about how it handles



Decontaminating Benchmarks Is a Fantasy: Ten Detectors Drop to Coin Flip After One Round of GRPO

inconvenient methodology results.

Related to all of this: the [same scaffold scoring 65% or 74% on SWE-bench depending on harness configuration](#) is a separate problem with the same consequence. Between harness variance and undetectable contamination, a single published benchmark number now has an error bar wide enough to cover most of the decisions people make with it.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →