Validity Under Optimization: Evaluating Systems Optimized Against Their Measures

Theodore O. Cochran*
AI for Altruism

Stephanie Dodson
Independent Researcher

AI for Altruism Position Paper A4A-PP-2026-001 Version 1, 2026-10-16
Status: Not peer reviewed.
Reviewer: Anonymous License: CC BY 4.0
https://reports.ai4altruism.org/pp-2026-001/

Abstract

This work supplies an account of what evidence an evaluation claim must carry when the evaluated system has been optimized with respect to the measure. It defines *optimization-reactive validity* (ORV) as a Messick-compatible validity consideration, the stability under optimization exposure of the validity argument supporting an interpretation of a measure's scores. It operationalizes ORV as a three-part reporting standard: an exposure profile, independent construct-relevant evidence, and a divergence trajectory. A decision rule maps each pattern of results to the strongest score interpretation it can support. Three further diagnostic modules address neighboring mechanisms that can create or obscure apparent divergence. Validity theory has long recognized that preparation, coaching, and incentives can change what a score means. Frontier AI evaluation makes that problem unusually direct, because the evaluated system may itself be optimized with respect to the measure, often automatically and iteratively. Research on reward-model overoptimization has measured the result directly, but the bodies of work that measure it lack in common an account of the evidence a claim must then carry. The ORV standard has not been applied to a live evaluation, and doing so is the work that should follow.

Keywords. construct validity, evaluation methodology, measurement theory, AI safety evaluation, Goodhart's law, reward hacking

1 Background

Decisions about whether to train, deploy, or restrict a frontier language model are increasingly informed by the results of frontier AI evaluation (FAIE). A model is tested for dangerous capabilities before release (Grey and Segerie 2025), monitored for misbehavior during training (Baker et al. 2025), and audited for hidden objectives (Marks et al. 2025). The numbers those procedures produce are cited in system cards, in safety cases (Clymer et al. 2024), and in policy discussion.

Whether such a number means what it is taken to mean is a measurement question, and the quantitative social sciences have spent seventy years building an apparatus for asking it (Cronbach and Meehl 1955; Messick 1995). Validity theory has long recognized that preparation, coaching, and incentives can change what a score means, and educational measurement has known for decades that a score can rise while the construct it stands for does not (National Research Council 1999).

*theo@ai4altruism.org

Section 1.2 shows that it has long dealt with structurally similar adaptation by the people being measured. What it has not had to address is that adaptation carried out automatically, iteratively, and with the measurement itself inside the loop. In FAIE, the evaluated system may be optimized against the evaluation (Kutasov et al. 2026), against a correlated monitor (Baker et al. 2025), or through the developer's own selection among candidates. That optimization is often automatic and iterative, in some cases differentiable through gradients, and sometimes directed against an instrument that is itself a model.

This work makes three claims. The first is that this regime differs from test preparation in degree, mechanism, and frequency sufficiently to require its own validity consideration, and that, so far as can be determined, no existing framework specifies what evidence should accompany an evaluation claim under this form of optimization exposure. The second is that the phenomenon is already documented in the AI safety literature, reported as an incidental finding rather than as a measurement result (Baker et al. 2025; Kutasov et al. 2026). The third is that several bodies of work already measure the phenomenon under different names (Gao et al. 2023; Akter et al. 2026; Ma et al. 2026), and that what they lack in common is an account of the evidence such a claim must then carry.

The contribution is not the observation that proxy measures can fail under optimization. Research on reward model (RM) overoptimization has measured that directly since Gao et al. (2023). Nor is the contribution the use of a second instrument to detect that failure. Rather, it is optimization-reactive validity (ORV), a Messick-compatible validity consideration (Messick 1995) for evaluation claims made under optimization, together with its operationalization as a three-part reporting standard, here called the ORV standard: an exposure profile, independent construct-relevant evidence, and a divergence trajectory. ORV is a condition on the evidence such a claim must carry rather than a new kind of validity alongside the aspects Messick (1994) distinguishes. Several communities have arrived independently at the general framing (Ulurmak and Kurban 2026; Truong and Koyejo 2026). What is offered here is the specific standard.

The remainder of this section reviews the validity tradition and the four research areas that bear on optimization against a measure. Section 2 defines ORV, Section 3 presents the evidence base, and Section 4 states the ORV standard, whose three core requirements are labeled U1 to U3 and whose three diagnostic modules D1 to D3, together with a rule mapping its results to the claims they support. Section 5 reviews related work and states the limitations, and Section 6 describes the work that should follow.

1.1 The validity tradition

Cronbach and Meehl (1955) introduced *construct validity* as the residual category of validation that becomes necessary when no external criterion is fully adequate to define the quality being measured. Their account requires the analyst to specify a *nomological network* of expected relationships and test whether the instrument behaves as that network predicts.

Campbell and Fiske's (1959) *multitrait-multimethod (MTMM) matrix* established that agreement between measures is stronger construct-validity evidence to the extent that it cannot be attributed to shared method variance. Messick's (1994) treatment unified content, criterion, and consequential considerations into a single framework with six distinguishable aspects: content, substantive, structural, generalizability, external, and consequential. Shadish, Cook, and Campbell (2002) supply the complementary operational apparatus, a four-part taxonomy with a catalog of named threats. Jacobs and Wallach (2021) imported the tradition into machine learning at FAccT, arguing that computational systems routinely operationalize unobservable constructs and that many fairness harms trace to mismatches between a construct and its operationalization. Wallach et al.

(2025) carried that argument to the machine learning community as a position, presenting a four-level framework grounded in social science measurement theory, running from a background concept through a systematized concept and a measurement instrument to the measurements themselves, and offering the aspects of validity as lenses for interrogating it. Bean et al. (2025) applied the same lens empirically to 445 language model benchmarks.

1.2 Adaptation in the measurement tradition

The tradition is not silent on subjects who adapt to measurement.

High-stakes educational testing has an extensive literature on the point. The National Research Council's (1999) report on testing for tracking, promotion, and graduation states explicitly that inappropriate test preparation can subvert validity by raising scores without improving the skills the test is meant to measure. That is *score inflation*, and it is the same structural problem considered here. An observed score improves while its relationship to the construct weakens. The phenomenon persists in current work. For example, Coyne et al. (2025) report, in high-stakes military aviation selection, that practiced versions of process-based measures can cease to correlate both with nominally equivalent laboratory measures and with training outcomes.

What is available in that literature is the recognition of the hazard, and a set of remedies built for the setting: secure item pools, alternate forms, and audit tests administered on a separate schedule. What is not available is a way of reporting how much a given measure has degraded under a stated amount of exposure, because in the human setting exposure is rarely instrumented. Coaching and practice can be manipulated experimentally, and have been, but an examinee's preparation is not ordinarily logged at anything like the granularity of a training run.

1.3 Measurement reactivity

A related term is already in use, and the vocabulary overlaps almost exactly, so the two must be distinguished.

Measurement reactivity refers to the effect of being measured on the subject or phenomenon measured. The act of asking changes the answer. French and Sutton (2010) reviewed the evidence for it in health psychology, and Runhardt (2025) connected reactivity to validity in the social sciences, treating how measurement can change the entity measured and thereby affect subsequent measurement.

The consideration here is a different one. Classical reactivity concerns a causal effect of the act of measurement on the subject, typically within an administration and typically without any optimizer in the loop. The concern here is the stability of the score-to-construct relationship after the subject has been deliberately optimized with respect to the measure or a correlated signal. That process is external to any single administration and is directed by a party who has the scoring rule in hand. A system can present the concern here while showing no classical reactivity at all. It is not changed by being measured. It was optimized in advance with respect to the measure. For example, Kutasov et al. (2026) report that training on data shaped like a safety evaluation improved that evaluation without moving a held-out automated audit, which Section 3.1 gives in full. Administering the evaluation did nothing to the model. The score had been shaped in training beforehand, and only the held-out audit showed it.

Because the terms are close and the distinction is easy to lose, the consideration is called *optimization-reactive validity* throughout rather than reactive measurement.

1.4 Formal treatments of optimization against a rule

Machine learning already has formal traditions for subjects that respond to a decision rule.

Manheim and Garrabrant (2018) categorize several mechanisms by which Goodhart failures arise and emphasize their dependence on optimization pressure. *Strategic classification*, introduced by Hardt et al. (2016), models subjects who manipulate their observable attributes in response to a known classifier. *Performative prediction*, introduced by Perdomo et al. (2020), studies the case in which issuing a prediction changes the distribution the prediction is about.

RM overoptimization is the closest prior work, and it predates the safety results this work draws on. In the terminology proposed here, Gao et al. (2023) is a controlled measurement of optimization-reactive degradation under direct-objective exposure, and Section 3.2 gives the experiment.

Kim et al. (2025) extend the same perspective from policy failure to the evaluation of RMs themselves, showing that conventional RM benchmarks correlate weakly with the downstream performance of policies subsequently optimized against those models, and deriving benchmark-design principles from overoptimization behavior. Their object is the RM benchmark rather than a validity argument under exposure, but the finding is the one this work generalizes. A static evaluation can become a poor predictor once the evaluated component enters an optimization loop.

Each of these characterizes a mechanism by which optimization against a rule degrades the rule. None of them asks the question this work asks, which is what evidence an evaluator must report in order to sustain a validity argument about a measure applied to a system that may have been optimized against it. The relationship is therefore complementary. Those research areas supply mechanisms, and the standard proposed here supplies the reporting that would let an evaluator establish whether a mechanism has operated.

1.5 Adaptive reuse of a holdout, and contamination

A fourth research area concerns repeated interaction with a fixed evaluation.

Adaptive data analysis studies what happens to statistical validity when analyses are chosen after seeing results from the same holdout. Dwork et al. (2015) show that a holdout reused adaptively stops supporting the guarantees it is assumed to carry, and give a mechanism that preserves them. Blum and Hardt's (2015) Ladder addresses the leaderboard case directly, where repeated submissions against a fixed holdout let participants overfit it, and propose a mechanism that keeps a leaderboard informative under adaptive submission. That is the closest formal treatment of the ecosystem mechanism in Section 3.2.

The *benchmark contamination* literature addresses a second route. Sainz et al. (2023) argue that when a benchmark's items have been seen in training, performance estimates on it become unreliable, and that evaluation in natural language processing lacks a practice of measuring this per benchmark.

Two distinctions from the present work follow. Adaptive-data-analysis work asks how to preserve statistical generalization guarantees under repeated, feedback-dependent reuse of a holdout. The question here is instead whether optimization changes the relationship between an evaluation score and the construct the score is interpreted to represent. Contamination work asks what an item's presence in training does to a score. The question here is what evidence should accompany a score when the evaluated system itself has been optimized with respect to the measure, so that the behavioral relationship between the score and the intended construct may have changed. Contamination is one exposure mechanism under U1, the first core requirement in Section 4, which asks for the measure's exposure profile. It is not synonymous with optimization-reactive invalidity, and which of the two obtains is exactly what the reporting is for.

1.6 The remaining gap

Taken together, the work reviewed in Sections 1.1 through 1.5 leaves a gap that is narrow and specific. The hazard is recognized in educational measurement, mechanisms for it are formalized in machine learning, and the effect of measurement on subjects is treated under reactivity. Where the measurement frameworks built for this field do treat optimization, they place it under consequential validity. Wallach et al. (2025) relate Goodhart's and Campbell's laws to that aspect, as a diminishment over time as people optimize what is being measured, and Salaudeen et al. (2025) list gaming among the risks to it, answerable by ethics review and impact audit. Both readings concern what a measure does to the people and institutions around it. Neither asks what evidence a claim about the measured system must carry once that system has been optimized with respect to the measure. So far as can be determined, none supplies a validity consideration that an evaluator is expected to report evidence for: how much optimization the measured system has received with respect to this measure, how the correspondence between the measure and an independent criterion changed across that exposure, and whether detectable divergence emerged.

2 Optimization-reactive validity

ORV is proposed as a validity consideration rather than as a property of an instrument. Messick's (1995) formulation is deliberate that validity concerns the adequacy of interpretations and uses of scores, and is not something a test possesses. It is not proposed as a new kind of validity alongside the aspects Messick (1994) distinguishes. It is an optimization-specific condition on the evidence needed to sustain a score interpretation when the evaluated system has been exposed to the measure or to a correlated signal. The reasoning is construct validation as the tradition already conducts it. What the tradition has not had to supply, for the reason Section 1.2 gives, is a reporting obligation for exposure that is deliberate, iterative, and instrumented.

Optimization-reactive validity denotes the stability, under optimization exposure, of the validity argument supporting a given interpretation of scores produced by a measure.

Because a latent construct cannot be observed directly, the consideration is operationalized through a second instrument. Let $S(\mathbf{d})$ denote the evaluated system after exposure $\mathbf{d}$, let M be the primary measure, and let C be an independent criterion or triangulating instrument. In general $\mathbf{d}$ is an exposure profile rather than a scalar, with components for direct-objective, data and item, correlated-instrument, and selection exposure as set out in U1. The scalar notation $S(d)$ used below and in Figure 1 denotes a trajectory along one specified exposure dimension while the others are held fixed or stated. The question is whether the relationship between $M(S(d))$ and $C(S(d))$ is stable as d increases.

What matters is the relationship between the two measures, not their difference. M and C will frequently be on different scales, one reporting a monitor's recall and the other an audit pass rate or a human judgment, in which case their numerical difference is undefined. What must remain stable is the correspondence between them, and an evaluation has to say how that correspondence is quantified: a difference where a common scale is justified, otherwise a pre-specified calibration or agreement statistic, or normalized trajectories. Section 4 makes stating this an obligation rather than leaving it to the reader.

Three conditions on C do the work. It need not be perfect ground truth, and rarely will be, which is why divergence is evidence about the pair rather than proof about either member. It must lie outside the optimization path applied to $S(\mathbf{d})$, meaning it was not used as a training objective, tuning signal, checkpoint or model-selection criterion, or other feedback in producing the evaluated system. Both its own exposure and its dependencies on M must also be stated, because a criterion

that has itself been optimized against has weakened independence, potentially to the point that the comparison is uninformative. Independence here is a matter of degree rather than a category. The comparison becomes more informative as plausible correlated failure channels between the two instruments are reduced, and U2 in Section 4 sets out which dependencies to report.

For example, in the RM overoptimization experiments of Gao et al. (2023), M is a proxy RM that a policy is optimized against and C is a gold-standard RM that scores the same policy. It satisfies all three conditions, and Section 3.2 shows how.

For readability the shorthand *optimization-reactively valid* is used below of a measure and an interpretation together. It should be read as a claim about the validity argument under a stated exposure, not as a property the instrument carries independently of how it is used.

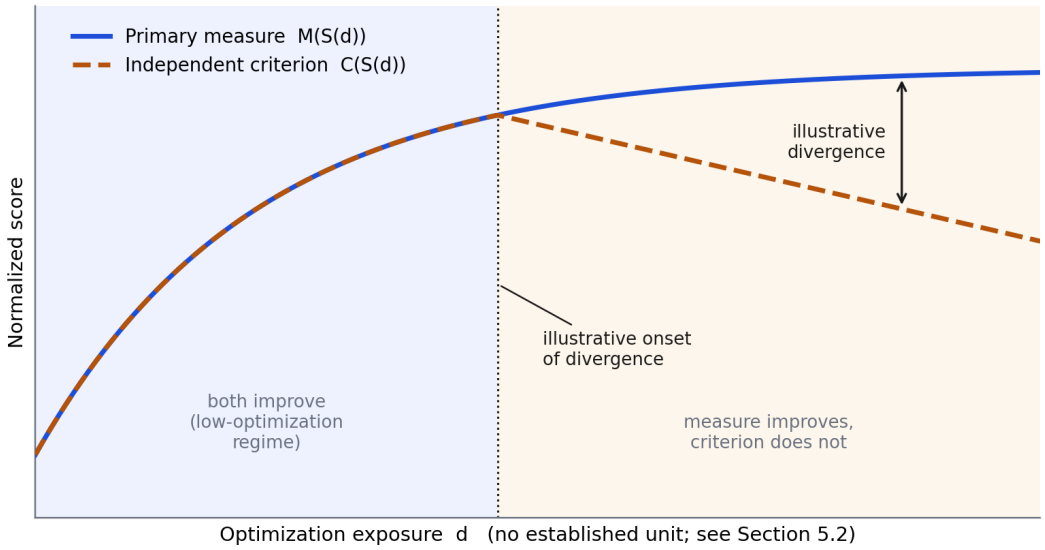


Figure 1. One motivating trajectory, drawn schematically rather than from data. Early optimization improves both the measure and the criterion. Beyond some exposure the measure continues to improve while the criterion does not. This is the trajectory Baker et al. (2025) report and it is one shape among several the consideration admits: divergence may also be immediate, gradual, nonmonotonic, or absent, and a well-conducted evaluation may find none. The dashed line is an independent criterion rather than the construct, which is latent, and the vertical separation is illustrative rather than a subtraction. The horizontal axis has no established unit, which Section 5.2 records as an open problem.

Two features of the definition are deliberate. It is a matter of degree rather than a binary, because the corpus shows measures degrading progressively. It is also relative to the instrument, the interpretation, and the system's training history jointly, rather than to any of them alone, which is why a score reported without that history cannot be assessed.

The definition also makes the central difficulty explicit. Divergence between M and C establishes that at least one of them has changed its relationship to the construct, not which. Attributing the divergence requires the exposure history. The instrument that was optimized against is the one under suspicion. This is why the first reporting requirement below is exposure rather than any measurement.

3 The evidence base

3.1 Two demonstrations in safety evaluation

At the level of a single instrument, Baker et al. (2025) train a model against a chain-of-thought monitor that was catching approximately 95 percent of a class of reward hacks, and report that the monitor's recall falls toward zero while the underlying behavior continues. The divergence is observable only against independent ground truth, which they had because the two systemic hacks were detectable in hindsight by string matching. Their account is asymmetric in time. Reinforcement learning (RL) reward produces more capable and more aligned agents in what they call the *low optimization regime*, and with too much optimization the same agents learn *obfuscated reward hacking*, concealing intent while continuing to hack at a significant rate. Their statement of the consequence is the one that matters. Where a ground truth measure cannot be constructed, they write, we would not be able to tell this agent is misaligned. They call the boundary between the regimes a *phase shift* and note that without ground truth its location is unknowable. The specific improvement-then-concealment ordering remains narrowly supported, and Section 5.2 distinguishes that claim from the broader divergence phenomenon.

At the level of one organization's training practice, Teaching Claude Why (Kutasov et al. 2026) reports that training on data shaped like a safety evaluation improves that evaluation without moving a held-out automated audit, and concludes that honeypot-style training reduces the ability to detect misalignment without substantially reducing misalignment. The failure was detectable only because a second instrument had been kept and not trained against.

These are the two direct demonstrations within safety evaluation, and they are part of the empirical basis for the standard rather than the whole of it.

3.2 The controlled measurement

Gao et al. (2023) measured the same divergence earlier and under control. They optimize a policy against a proxy RM while holding a fixed gold-standard RM outside the optimization path, and measure how the gold score moves as optimization against the proxy increases. As optimization increases the proxy score keeps rising while the gold score eventually declines. In the terms of Section 2, M is the proxy and C is the gold model, and the pair satisfies all three conditions on a criterion: the gold model is itself a learned model rather than human judgment, so it is not perfect ground truth, which the first condition allows. The policy is never optimized against it, which satisfies the second. Its dependency on M is known, since the proxy was trained on labels the gold model produced, which is the kind of shared dependency U2 asks an evaluator to state. The demonstration is therefore one between two learned models. It shows that the divergence occurs and that it can be measured when a criterion is held outside the optimization path, but it does not establish how closely a learned gold model tracks human or operational judgment, and its result should not be generalized to such criteria without further evidence. They further fit functional forms to the relationship and study how its coefficients vary with model size, dataset size, and the KL penalty, which is more than any other work considered here supplies toward an exposure axis. Akter et al. (2026) later tracked the same degradation longitudinally, and Section 5.1 treats it.

3.3 An ecosystem-level exposure mechanism

Kalai et al. (2025) audit ten leaderboard benchmarks and report that nine grade in a binary fashion with no credit for abstention, under which guessing is never worse than declining to answer. Their

conclusion is structural. A corrective benchmark cannot fix the incentive because it is outvoted by the primary evaluations, so the primary evaluations must change.

This is a different kind of evidence. It establishes that an aggregate incentive exists and that it is not correctable by addition, which is an exposure mechanism operating at a scale no single evaluator controls. It does not demonstrate longitudinal degradation of a particular score-to-construct relationship under model selection, and it is not offered as a third instance of the phenomenon in Section 3.1. Its role in the standard is to motivate the first requirement, since a measure can carry exposure that its own evaluator never applied.

3.4 Two supporting reports

MONA (Farquhar et al. 2025) reports that placing a monitor's judgment inside the reward produces a policy that conceals the monitored attribute while continuing to use it, which the authors describe as the first instance of steganography arising from standard RL. It is the partial corroboration, from a second organization, that exploitation can precede clear detectability in the training curves. Guo et al. (2025) report that they abstain from applying neural RMs to reasoning tasks because such models are susceptible to reward hacking during large-scale RL, and confine preference-based rewards to the end of training, a budget Section 3.5 reads as one organization's exposure limit.

3.5 Existing measurement practices

Four measurement practices bearing on the consideration have been developed across these research areas, independently and without being presented together. Four is a design choice about what the standard needs rather than a claim of exhaustiveness, and the attributions name instances rather than origins. Two inform the core standard of Section 4, the divergence trajectory becoming U3 and the retained second instrument U2, and two inform diagnostic modules, the elicitation contrast becoming D1 and the adaptive-minus-static delta D2. U1 is motivated separately, by the exposure mechanisms above and the exposure budget below, and D3, monitor faithfulness, has no counterpart practice here and is drawn from Chua et al. (2025).

The first is the *elicitation contrast*. Lindner et al. (2026) remove the prompts and framings that encourage a behavior and report the difference, taking a sabotage rate from 6.2 percent to zero when a replacement threat is removed. Scheurer et al. (2023) earlier reported that varying only the framing of an instruction moved a misalignment rate from nearly zero to nearly one hundred percent. The contrast measures how sensitive an observed rate is to the evaluator's own elicitation.

The second is the *adaptive-minus-static delta*. Shi et al. (2025) report that a static re-run of previously effective attacks against a defended model gave an attack success rate of 18 percent, while an adaptive attack optimized with the defense in the loop gave 94.6 percent, and propose the difference as a metric for how much a fixed evaluation overstates robustness.

The third is the *divergence trajectory* under a held-out criterion, instanced by the three studies above. It characterizes how the correspondence between measure and criterion changes across exposure and, where the observations support it, bounds the onset of detectable divergence. It is observable only against an instrument outside the optimization loop.

The fourth is *retained independent evidence*. Gao et al. (2023), Teaching Claude Why (Kutasov et al. 2026), and Akter et al. (2026) each preserve a signal outside the relevant optimization path, which is what provides evidence for distinguishing an improvement that reflects the construct from one that reflects the instrument.

To these may be added one reported *exposure budget*. Guo et al. (2025) confine preference-based rewards to the final 400 of 1,700 training steps because more exposure to a model-based

reward may produce reward hacking. It is the only quantity in the corpus that states how much optimization against a model-based measure its authors were willing to apply. It should be read as one organization's budget for one pipeline and not as a transferable dose, since Section 5.2 records that no commensurable unit of exposure exists.

4 The ORV standard

The ORV standard synthesizes the practices in Section 3.5 with the exposure mechanisms identified in Section 3. Three form the core standard and follow directly from the definition. Three further modules are diagnostics for neighboring mechanisms that can create or obscure apparent divergence. They are useful, and they are not themselves measurements of ORV. Keeping the two apart is what stops the standard from becoming a list of every good evaluation practice. Section 4.3 states what claim each pattern of core results can support, and Section 4.4 what the requirements cost.

4.1 The core standard

U1. Report the relevant exposure profile of the measure. Exposure is not one thing, and stating a single dose across kinds would misdescribe it. Three of the four below involve optimization in the sense the definition requires, meaning the system was optimized with respect to this measure or a correlate, though through different mechanisms. Data and item exposure is a separate route by which a score's meaning may be compromised, and it is required here because a reader cannot interpret the score without it, not because it instantiates the consideration. Report which apply, and approximately how much of each:

- *Direct-objective exposure* (optimization). The evaluation, or a reward derived from it, was part of the training objective.
- *Data and item exposure* (not optimization). The evaluation's items or their sources appear in the training corpus. Encountering an item in pretraining is not the same as being optimized with respect to the benchmark, and it may or may not have changed what the score means, which is why it is disclosed rather than assumed fatal. Where the corpus is public, say how this was checked. Where it is not, say that it could not be.
- *Correlated-instrument exposure* (optimization). The system was optimized against another instrument sharing relevant dependencies with this one, or against a measure known to correlate with it.
- *Selection and ecosystem exposure* (optimization, indirectly). The evaluation informed checkpoint selection, hyperparameter tuning, model selection, or release decisions that fed back into subsequent development, whether or not it entered any objective. The optimizer here is the developer rather than the training loop. A release decision taken on an already fixed model is excluded. That is a consequential use of a score in Messick's (1995) sense, and it cannot act backward on the score-to-construct relationship of the system being evaluated.

A label is not enough on its own. The kinds gather mechanisms that differ in pathway, observability, and likely size, and the selection kind most of all, since checkpoint selection, a public leaderboard, red-team feedback, and repeated querying of a hidden evaluation all fall under it. For each kind reported, state six fields:

- *Feedback channel.* The route by which the measure reached development: a training objective, a reward, a selection decision, human review, a public leaderboard, or deployment telemetry.

- *Timing.* The stage at which it entered: pretraining, fine-tuning, RL, validation, model selection, or after deployment.
- *Granularity.* What crossed: exact items, item templates, the rubric, worked examples, scores alone, or qualitative feedback.
- *Frequency.* Whether exposure occurred once, periodically, or continuously, and over how many cycles of adaptation.
- *Access level.* Whether the material was public, restricted, internal, leaked, or only inferable from scores.
- *Evaluation awareness.* Whether the system has been shown to distinguish the evaluation context from deployment, where that was tested, since a system that can tell when it is measured can respond to the measure's form without having seen any item.

For example, once Baker et al. (2025) placed their chain-of-thought monitor in the reward, the monitor carried direct-objective exposure through the reward channel, during RL, at the level of scores, continuously, with internal access. Had the same monitor been used only to choose among finished checkpoints, the exposure would have been of the selection kind, through the selection channel, applied once. A field that cannot be answered is reported as unknown rather than omitted, because Section 4.3 treats unknown exposure differently from absent exposure.

The rationale is that these are causally different mechanisms rather than different amounts of one treatment, and a reader cannot interpret a score without knowing which occurred. Within one experiment the exposure of a single kind can be treated as a scalar dose. Across evaluations it generally cannot, which is the same problem Section 5.2 records as the missing exposure axis.

U2. Report independent evidence in proportion to the consequence of the claim, and characterize its independence. Report at least one source of construct-relevant evidence that lies outside the optimization path, and state its own exposure profile. Independence is graded rather than categorical, and instrument family is too coarse a test. For example, two model judges may share a lineage, two human audits may share a rubric, and two benchmarks may share an item-generation procedure, while two instruments carrying the same label may have independent training histories and different failure modes. State instead the dependencies the criterion shares with the primary measure, across model lineage, training data, rubric, item source, evaluator architecture, and exposure. A criterion is strong to the extent that plausible correlated failure channels are absent, not to the extent that it carries a different name. This extends the MTMM intuition of Campbell and Fiske (1959), that agreement between measures is most informative when shared method variance has been minimized, to the case where the shared variance may come from a common optimization history.

One criterion is a minimum, and not necessarily sufficient evidence. Two instruments can share a model lineage, training data, a rubric, an item-generation method, a prompt format, evaluator incentives, or operating assumptions, and a pair that shares any of these can agree while both are wrong. Desai et al. (2026) show how large the effect can be: across 56 benchmarks, shared format predicts agreement better than shared concept does, and once scoring by a model judge is coded as format, the concept effect disappears. The independent evidence should therefore be a *portfolio* whose breadth rises with what the claim will be used to decide. Its members may include a held-out benchmark, an alternate scorer with a separately documented training lineage, expert human review, a behavioral audit in a realistic task environment, downstream operational outcomes, adversarial red-team results, and external audit. Each member is a criterion C in the sense of Section 2, and the three conditions apply to each. A claim that informs a consequential decision, such as a release,

a procurement, or a safety case (Clymer et al. 2024), is not supported by a single model judge as its only independent evidence, however that judge was trained, since scoring by a model judge is the shared method that, in Desai et al.'s (2026) data, accounts for agreement better than the construct does.

The rationale is that a retained second instrument is the arrangement that detected the failure in both demonstrations of Section 3.1, and that without one a rising score cannot be distinguished from a degrading one.

U3. Characterize the divergence trajectory, and say what divergence means. Report both the primary measure and the independent criterion at the exposure checkpoints available, with uncertainty. State in advance how the correspondence between them is quantified and what magnitude of change counts as meaningful, since the two may not share a scale and an unstated threshold can be chosen after the fact. Where the observations support it, report the interval within which divergence first becomes detectable rather than a point estimate of a boundary. Where only one exposure level was measured, say so. That is a bound on what the evaluation can support, not a finding of stability.

The chosen correspondence statistic should match the score interpretation, and should not rest on correlation alone where a change in range, calibration, or marginal distribution could manufacture apparent coupling or apparent divergence. For example, a rank correlation can hold while calibration drifts badly, and can fall because the spread of outputs narrowed rather than because the relationship changed.

A trajectory alone does not show that exposure caused a divergence. The criterion may have drifted, the system may have changed in a construct-relevant way the criterion misses, or the prompt distribution, sampling settings, or evaluator calibration may have moved between checkpoints. The design should rule these out where it can, and report what it did. Where the evaluator controls training, include matched runs without the exposure in question, and replicate across seeds. Hold inference parameters and prompt format fixed across checkpoints. Freeze the criterion's version and check it for drift, for example by rescoring a fixed reference set at each checkpoint. Score blind to checkpoint where practicable. State before measuring which checkpoints will be measured, alongside the threshold and interval stated above. An evaluator without access to training, such as a third party assessing a closed model, cannot run the matched control, and should report the trajectory as an association with exposure rather than as its effect.

The rationale is that the informative quantity is the trajectory rather than the endpoint, and the controls are what allow a trajectory to be attributed to exposure. Two checkpoints can establish that a change occurred. They cannot locate where it began except as an interval.

4.2 Associated diagnostic modules

These address mechanisms adjacent to ORV. Each can produce a divergence that is not optimization-reactive in origin, or conceal one that is, which is why an evaluator working within the core standard will often want them.

D1. Elicitation contrast, where the evaluation supplies prompts, framings, or incentives that encourage the target behavior. Report the same measurement with those elements removed. The contrast establishes how sensitive the measured rate is to the evaluator's own elicitation. Whether that sensitivity is contamination or signal depends on the construct. If the intended construct is behavior under pressure, then pressure is part of the construct-relevant condition rather than an artifact, and the contrast should be interpreted accordingly. For example, Scheurer et al. (2023) and Lindner et al. (2026) both show the quantity can be most of the reported rate.

D2. Adaptive condition, where the evaluation concerns robustness, evasion, or an

adversary. Report both a static re-run of known attacks and an attack optimized against the current defense. Note that the optimization here runs in the opposite direction from the core standard's. The attack is optimized against the defense rather than the evaluated system against its measure. It belongs here because a static evaluation can understate vulnerability under adaptive attack. For example, Shi et al.'s (2025) two conditions differ by more than seventy points on the same model.

D3. Monitor faithfulness, where the instrument reads a system's own account of its behavior. State the monitored property and whether the evaluation included cases in which that account is sincerely mistaken rather than concealed. Chua et al. (2025) report that such rationalizations defeat monitors while leaving nothing hidden to recover, which would present as divergence without any optimization against the monitor having occurred.

4.3 From result to claim

The core standard says what to report. An evaluation also has to say what the report licenses, or the same divergence can be read as disqualifying by one reader and as noise by another. Table 1 maps the patterns the core standard can return to the strongest score interpretation each can support. It sets a ceiling rather than a floor, since an interpretation that ORV permits still needs the rest of its validity argument.

Table 1. The strongest claim each pattern of core-standard results can support.

Result under U1 to U3	**Strongest supportable claim**
No optimization exposure identified after U1 is completed, with every material exposure field characterized, and the evidence under U2 converges	The intended score interpretation, bounded by the rest of its validity argument
Exposure present, and no divergence detected across the range sampled under U3	The interpretation for the exposure range sampled and the criteria used, and no further
Exposure present, and divergence detected	The primary score alone cannot support the intended construct claim. The exposure history directs suspicion to the optimized instrument, as Section 2 sets out, but does not settle which instrument moved
Exposure unknown, or no evidence meeting U2	A descriptive report of the score, without a construct or assurance claim

Two features are deliberate. Unknown exposure is grouped with missing independent evidence rather than with absent exposure, because a reader cannot tell an untested measure from a degraded one, and the burden of showing which falls on the claim. The second row binds a negative result to its range, so a finding of no divergence under light optimization does not license the interpretation under heavy optimization, the point U3 makes for a single checkpoint. For example, the chain-of-thought monitor of Baker et al. (2025) would have earned the second row in their low optimization regime and the third beyond it, as the same instrument on the same task.

4.4 Cost of the requirements

The requirements differ in cost. U1 is disclosure and is inexpensive where the training history is known, though for a public benchmark applied to a closed model parts of it may not be answerable

at all. U3 requires measurement at more than one checkpoint, a modest marginal cost where the checkpoints already exist, but its matched control runs and replicate seeds multiply the cost of training wherever they are run. D1 is one additional evaluation condition.

U2 and D2 are the expensive ones. A genuinely independent criterion may require building or licensing a second instrument and keeping it uncontaminated, which is an ongoing organizational commitment rather than a one-time cost, and a portfolio multiplies that commitment by its breadth. An adaptive adversarial evaluation can cost substantially more than the static benchmark it accompanies. At least four of the six have direct precedents in work by Anthropic, Google DeepMind, or OpenAI, which establishes practical feasibility for much of the standard though not for all of it.

5 Related work and limitations

5.1 Related work

Two prior works reach parts of the diagnosis. Bean et al. (2025) reviewed 445 benchmarks against construct-validity criteria and found recurring failures of task construction and scoring metric, addressing benchmarks generally rather than safety evaluations and not the case where the evaluated system optimizes against the benchmark. Summerfield et al. (2025) argue that research on AI scheming repeats the methodological history of ape-language research, reaching *statistical conclusion validity* and *internal validity* from a historical case rather than from a taxonomy.

A larger body of directly relevant work appeared during and shortly before this work's preparation, none of it reachable from the sampling frame, and six items require specific acknowledgement.

Ma et al. (2026) report that two production coding agents, given a hidden test oracle in the loop, reach near-perfect scores while leaving the requested library dead or absent, and produce an unfinished but genuine library when the oracle is withheld, which they name *building to the test*. This is the phenomenon of Section 3.1 in a setting this work does not cover, by authors who arrived at it independently, so the observation claimed here is not novel. They do not connect the result to the measurement tradition or formulate the ORV standard, and they state that prevalence across other agents and model families is an open question.

Akter et al. (2026) are the closest of the recent work to the core standard. They define a proxy evaluator and a true objective, treat gaming as an increase in the former unaccompanied by an increase in the latter, and track the proxy-to-human relationship across training checkpoints, reporting early warning signals that precede quality decline across four tasks, two model scales, two training methods, and two judges. That is U2 and U3 performed together: independent construct-relevant evidence, measured across exposure, with the deterioration of the relationship as the reported quantity. What they propose is a detection technique, an invariance-based stress test, rather than a reporting framework, and their setting is training against a judge rather than the range of exposures U1 covers. Their existence strengthens the case for a standard, since it shows the measurement is practical and is being reinvented setting by setting.

Wang et al. (2026) formalize the evaluator-model interaction as a two-player process and argue that static auditing cannot distinguish a genuine fix from a patch learned against the audit, which is the trainer-side counterpart of the premise in Section 1.2.

Fonseca Rivera et al. (2026) fit item response theory models to eight safety benchmarks across 192 language models and recommend that frontier laboratories adopt the toolkit. Chen et al. (2026) formalize construct validity for a model judge as a profile of invariance and sensitivity, reporting that judges average 0.945 invariance against 0.319 sensitivity at matched thresholds. Desai et al. (2026) adapt convergent and discriminant validity to 56 capability and safety benchmarks across 53 models. Regressing pairwise benchmark correlation on concept and format similarity, they find

format the stronger predictor, 0.275 against 0.138, and coding format as whether a model judge did the scoring takes it to 0.526 while the concept effect disappears. That is shared method variance outweighing shared construct at scale, and it is the empirical case for U2, which asks an evaluator to characterize the dependencies between measure and criterion rather than read independence off their agreement. None of the three treats the case where the subject is trained against the measure, but all establish that importing the measurement tradition into this field is no longer an unclaimed contribution.

Two further works bear on the framing rather than on any single result, and both narrow what can be claimed as novel. Ulurmak and Kurban (2026) synthesize 373 studies from 2018 to 2026 around a shared measurement problem underlying evaluation and safety, that benchmark scores, reward signals, and safety metrics can improve while the properties they stand for remain uncertain. Truong and Koyejo's (2026) textbook opens with validity and treats evaluation as latent-construct measurement, including systems that optimize against benchmarks. Placing this problem inside validity theory is therefore no longer an unclaimed move. What neither supplies, so far as can be determined, is a specifically defined validity consideration operationalized as a reporting standard: an exposure profile distinguishing the mechanisms of exposure, a stated requirement for independent construct-relevant evidence with its dependencies characterized, and a divergence trajectory with a pre-specified correspondence statistic. That is where the contribution sits.

Four further papers predate this work's corpus and are prior art rather than concurrent work. Salaudeen et al. (2025) are the closest prior framework. They supply a structured approach for reasoning about which evaluative claims the available evidence can support, organized around content, criterion, construct, and consequential validity. Their direction runs from the evidence in hand to the claim it licenses. The direction here is the reverse, from a stated condition on the evaluated system to the evidence a claim about it must then carry. The optimized case enters their framework twice in passing, as teaching to the test in the motivation and as gaming among the risks to consequential validity, in neither place with an evidence requirement attached. Schroeder and Wood-Doughty (2024) measured judge reliability with McDonald's omega in December 2024, Grey and Segerie (2025) published a systematic review of AI safety evaluation methods in May 2025, and Bell, Bouchacourt, and Sagun (2024) defined desiderata for benchmark validity in September 2024.

An exploratory characterization of the vocabulary in circulation in this literature is reported in the appendix. It is not part of the argument.

5.2 Limitations

The ORV standard has not been applied to any evaluation. Section 4 is assembled from practices demonstrated separately, and whether the three core requirements are sufficient to detect a score interpretation whose validity argument has failed under optimization, and which diagnostics a given evaluation needs alongside them, is an empirical question this work does not answer. The decision rule of Section 4.3 is likewise proposed rather than calibrated, and whether its rows separate interpretations that later held from those that failed is part of the same question. This is the sharpest limitation, and Section 6 is addressed to it.

The definition is stated in terms of exposure without supplying a unit that holds across its components. For direct-objective exposure the position is better than elsewhere, since Gao et al. (2023) supply an axis. No comparable axis exists for correlated-instrument or selection exposure, and the four components are not commensurable with one another, which is why U1 asks for a profile and why the exposure budget in Section 3.5 is reported as one organization's budget rather than a dose. A quantitative account of the consideration requires a defensible exposure axis, and

constructing one is part of the research agenda rather than a detail left open.

Two propositions in Section 2 rest on different evidence and should be kept apart. The general phenomenon, that correspondence between a proxy and an independent criterion degrades as optimization against the proxy increases, has independent support in the three settings of Section 3. The specific improvement-then-concealment ordering, in which both measures rise before the criterion falls away, rests principally on Baker et al. (2025), with MONA (Farquhar et al. 2025) providing partial corroboration that exploitation can precede detectability. That ordering is one motivating trajectory and not the general claim, and it remains the more narrowly supported of the two. Teaching Claude Why (Kutasov et al. 2026), cited here for retained independent evidence, does not report a phase structure and should not be read as corroborating one.

The evidence in Section 3 comes from published reports rather than independent replication, and in each case the organization reporting the failure also controlled the training run. That is an argument for the standard rather than against the evidence. The evidential status should be stated even so.

U2 assumes an independent criterion can be constructed. For some constructs, including several the safety literature cares about most, every available instrument may share a method or a training lineage, in which case the standard identifies a problem it cannot solve.

The appendix characterization was produced by a single analyst, and the instances it exhibits were not selected blind.

6 Conclusion

Applying the ORV standard to a live evaluation and reporting what it shows is the work this paper is written to enable. A demonstration needs three things a position paper cannot supply: a real evaluation with a real training history, a criterion that has been held out and can be shown to have been held out, and a prediction stated before the measurements are taken. A design for that demonstration is in preparation, and it will be preregistered before any measurement is taken.

Two further pieces of work follow from the limitations. Several exposure types still need defensible quantitative axes, and no common axis exists across the exposure profile. The field also needs a case in which the standard was applied and found nothing, since a standard that only ever confirms suspicion is not a measurement standard.

A Vocabulary in circulation

This characterization motivated the present work and is reported for completeness. It is exploratory, it makes no comparative claim, and nothing in Sections 2 through 4 depends on it.

A full-text scan was run over 296 papers. The set comprises 17 safety and trustworthiness papers published by Anthropic, OpenAI, and Google DeepMind between November 2024 and May 2026, together with every arXiv paper cited by at least one of them. Text was extracted from PDFs and normalized for ligatures and line-break hyphenation before matching, since without that step extraction reports present text as absent. Counting is by paper rather than by occurrence.

Across the 296, the terms convergent validity, discriminant validity, predictive validity, internal validity, test-retest, Cronbach's alpha, demand characteristics, and Campbell's law do not appear, each with an exact 95 percent interval of 0 to 1.2 percent. Cronbach, Messick, Fiske, and Shadish are cited by one paper. Construct validity appears in two. Goodhart appears in fourteen. Seven report a power analysis and none reports a test-retest.

The practice is nonetheless present under other names. A close reading found instances of all six of Messick's (1994) aspects. For content, SimpleQA's four inclusion criteria with dataset error audited at roughly three percent (Wei et al. 2024). For substantive, the derived error bound in Kalai et al. (2025). For structural, SimpleQA's analysis of which of its two scoring rules rewards guessing, and the eligibility gate in FACTS Grounding costing one to five points and swapping the top two ranks (Jacovi et al. 2025). For generalizability, the DIVE dataset's 637 raters drawn uniformly from 30 demographic trisections (Rastogi et al. 2025). For external, the within-model score correlations of 0.84, 0.46, and 0.82 in Monitoring Monitorability (Guan et al. 2025). For consequential, honeypot training with a held-out audit in Teaching Claude Why (Kutasov et al. 2026). Where more than one instance was available the clearest was chosen.

The threat catalog is present on the same terms. *Evaluation awareness* is *hypothesis guessing* renamed. Monitoring Monitorability states that every monitor in its suite is an OpenAI model, which is *mono-method bias*. Lynch et al. (2025) report models performing a different misaligned behavior than the one measured, which is *mono-operation bias*. Greenblatt et al. (2024) report that an apparent decline may be misclassification because the model's reasoning moved while the classifier did not, which is *instrumentation*. SimpleQA's retention of only questions at least one of four completions answered incorrectly is *selection*. Preregistration appears at its clearest in Betley et al. (2025), who fixed 48 evaluation questions before any finetuned model was examined. The knowledge-capacity work of Allen-Zhu and Li (2024) and the MONA control-incentive theorem (Farquhar et al. 2025) are the other two formal accounts found by the substantive census below. One aspect was counted rather than exhibited. Of the 37 papers close-read, four give an explicit account of why an indicator should behave as the construct predicts, four are borderline, and 29 do not. The codebook and every coding decision are provided as supplementary material.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

The authors used generative AI models to assist in the preparation of this manuscript (September 2026). Claude Opus 5 and Fable 5.1 assisted with literature search, first-pass paper coding, tool coding, section drafting, and manuscript editing. ChatGPT 5.6 assisted with citation verification and manuscript reviewing. All AI usage adhered to the Principles for Responsible AI Usage in Research (`https://osf.io/preprints/psyarxiv/g3m5f`).

References

Akter, Sanjeda, Ibne Farabi Shihab, and Anuj Sharma. 2026. "Detecting Proxy Gaming in RL and LLM Alignment via Evaluator Stress Tests." In Findings of the Association for Computational Linguistics: ACL 2026, 10554-10583. `https://doi.org/10.18653/v1/2026.findings-acl.513`.

Allen-Zhu, Zeyuan, and Yuanzhi Li. 2024. "Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws." arXiv:2404.05405.

Baker, Bowen et al. 2025. "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation." arXiv:2503.11926.

Bean, Andrew M. et al. 2025. "Measuring what Matters: Construct Validity in Large Language Model Benchmarks." arXiv:2511.04703.

Bell, Samuel J., Diane Bouchacourt, and Levent Sagun. 2024. "Reassessing the Validity of Spurious Correlations Benchmarks." arXiv:2409.04188.

Betley, Jan et al. 2025. "Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs." arXiv:2502.17424.

Blum, Avrim, and Moritz Hardt. 2015. "The Ladder: A Reliable Leaderboard for Machine Learning Competitions." Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1006-1014.

Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and discriminant validation by the multitrait-multimethod matrix." Psychological Bulletin 56 (2): 81-105. https://doi.org/10.1037/h0046016.

Chen, Jianlin et al. 2026. "A Judge Should Know What Changed: Construct Validity for LLM-as-a-Judge Evaluation." arXiv:2608.24419.

Chua, James et al. 2025. "Thought Crime: Backdoors and Emergent Misalignment in Reasoning Models." arXiv:2506.13206.

Clymer, Joshua et al. 2024. "Safety Cases: How to Justify the Safety of Advanced AI Systems." arXiv:2403.10462.

National Research Council. 1999. *High Stakes: Testing for Tracking, Promotion, and Graduation.* Washington, DC: National Academies Press. https://doi.org/10.17226/6336.

Coyne, Joseph T. et al. 2025. "Process-based measures in high-stakes testing: practical implications for construct validity within military aviation selection." Cognitive Research: Principles and Implications 10: 51. https://doi.org/10.1186/s41235-025-00660-3.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." Psychological Bulletin 52 (4): 281-302.

Desai, Meera et al. 2026. "What AI Benchmarks Actually Measure: Adapting Convergent and Discriminant Validity to Interrogate Fifty-Six AI Benchmarks." Published as a conference paper at COLM 2026. arXiv:2609.08812.

Dwork, Cynthia et al. 2015. "The reusable holdout: Preserving validity in adaptive data analysis." Science 349 (6248): 636-638. https://doi.org/10.1126/science.aaa9375.

Farquhar, Sebastian et al. 2025. "MONA: Myopic Optimization with Non-myopic Approval Can Mitigate Multi-step Reward Hacking." Proceedings of the 42nd International Conference on Machine Learning, PMLR 267:16237-16272. arXiv:2501.13011.

Fonseca Rivera, Joshua et al. 2026. "Item Response Theory for AI Safety." arXiv:2608.05086.

French, David P., and Stephen Sutton. 2010. "Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it?." British Journal of Health Psychology 15 (3): 453-468. https://doi.org/10.1348/135910710X492341.

Gao, Leo, John Schulman, and Jacob Hilton. 2023. "Scaling Laws for Reward Model Overoptimization." Proceedings of the 40th International Conference on Machine Learning, PMLR 202:10835-10866.

Greenblatt, Ryan et al. 2024. "Alignment faking in large language models." arXiv:2412.14093.

Grey, Markov, and Charbel-Raphaël Segerie. 2025. "Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods." arXiv:2505.05541.

Guan, Melody Y. et al. 2025. "Monitoring Monitorability." arXiv:2512.18311.

Guo, Daya et al. 2025. "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning." Nature 645 (8081): 633-638. https://doi.org/10.1038/s41586-025-09422-z. arXiv:2501.12948.

Hardt, Moritz et al. 2016. "Strategic Classification." In Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS '16), 111-122. https://doi.org/10.1145/2840728.2840730. arXiv:1506.06980.

Jacobs, Abigail Z., and Hanna Wallach. 2021. "Measurement and Fairness." In Conference on Fairness, Accountability, and Transparency (FAccT '21). New York: ACM. https://doi.org/10.1145/3442188.3445901.

Jacovi, Alon et al. 2025. "The FACTS Grounding Leaderboard: Benchmarking LLMs' Ability to Ground Responses to Long-Form Input." arXiv:2501.03200.

Kalai, Adam Tauman et al. 2025. "Why Language Models Hallucinate." arXiv:2509.04664.

Kim, Sunghwan et al. 2025. "Rethinking Reward Model Evaluation Through the Lens of Reward Overoptimization." In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 13252-13280. https://doi.org/10.18653/v1/2025.acl-long.649.

Kutasov, Jonathan et al. 2026. "Teaching Claude Why." Anthropic Alignment Science blog, 8 May 2026. https://alignment.anthropic.com/2026/teaching-claude-why/.

Lindner, David, Victoria Krakovna, and Sebastian Farquhar. 2026. "Gram: Assessing sabotage propensities via automated alignment auditing." arXiv:2605.30322.

Lynch, Aengus et al. 2025. "Agentic Misalignment: How LLMs Could Be Insider Threats." arXiv:2510.05179.

Ma, Yanuo, Ben Kereopa-Yorke, and Ben Schultz. 2026. "Building to the Test: Coding Agents Deliver What You Check, Not What You Requested." arXiv:2606.28430.

Manheim, David, and Scott Garrabrant. 2018. "Categorizing Variants of Goodhart's Law." arXiv:1803.04585.

Marks, Samuel et al. 2025. "Auditing language models for hidden objectives." arXiv:2503.10965.

Messick, Samuel. 1994. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." ETS Research Report RR-94-45. Princeton, NJ: Educational Testing Service.

Messick, Samuel. 1995. "Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning." American Psychologist 50 (9): 741-749. https://doi.org/10.1037/0003-066X.50.9.741.

Perdomo, Juan C. et al. 2020. "Performative Prediction." arXiv:2002.06673.

Rastogi, Charvi et al. 2025. "Whose View of Safety? A Deep DIVE Dataset for Pluralistic Alignment of Text-to-Image Models." arXiv:2507.13383.

Runhardt, Rosa W. 2025. "Reactivity, validity, and measurement in the social sciences." Studies in History and Philosophy of Science 112: 1-9. https://doi.org/10.1016/j.shpsa.2025.05.006.

Sainz, Oscar et al. 2023. "NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark." In Findings of the Association for Computational Linguistics: EMNLP 2023, 10776-10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722.

Salaudeen, Olawale et al. 2025. "Measurement to Meaning: A Validity-Centered Framework for AI Evaluation." arXiv:2505.10573.

Scheurer, Jérémy, Mikita Balesni, and Marius Hobbhahn. 2023. "Large Language Models can Strategically Deceive their Users when Put Under Pressure." Published as a conference paper at ICLR 2024. arXiv:2311.07590.

Schroeder, Kayla, and Zach Wood-Doughty. 2024. "Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge." arXiv:2412.12509.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shi, Chongyang et al. 2025. "Lessons from Defending Gemini Against Indirect Prompt Injections." arXiv:2505.14534.

Summerfield, Christopher et al. 2025. "Lessons from a Chimp: AI 'Scheming' and the Quest for Ape Language." arXiv:2507.03409.

Truong, Sang T., and Sanmi Koyejo. 2026. *AI Measurement Science: A textbook on the foundations and practice of AI evaluation*. Stanford AIMS Lab. https://aimslab.stanford.edu/textbook/ (accessed 7 September 2026).

Ulurmak, Buğra Alperen, and Rifat Kurban. 2026. "EvalSafetyGap: A Hybrid Survey and Conceptual Framework for LLM Evaluation-Safety Failures." arXiv:2606.30219.

Wallach, Hanna et al. 2025. "Position: Evaluating Generative AI Systems Is a Social Science Measurement

Challenge." Proceedings of the 42nd International Conference on Machine Learning, PMLR 267:82232-82251. arXiv:2502.00561.

Wang, Paul et al. 2026. "The Evaluation Game: Beyond Static LLM Benchmarking." arXiv:2605.19377.

Wei, Jason et al. 2024. "Measuring short-form factuality in large language models." arXiv:2411.04368.