Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When measurements of AI systems disagree A methodological review of dissociations in AI safety

Eyal M. Reingold

reingold@psych.utoronto.ca

University of Toronto

Omer Reingold

Stanford University

Research Article

Keywords: AI safety, dissociation, measurement validity, knowledge elicitation, chain-of-thought monitoring, strategic evaluation

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11415633/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence safety research compares systems' answers and actions with output probabilities, generated reasoning, internal activations and monitoring judgments. These measurements can disagree about what a system knows, what influenced its response and whether unwanted behavior occurred. This critical tutorial review organizes the emerging literature around five comparisons: relative sensitivity, incremental value, functional dissociation, causal intervention and targeted robustness. Worked examples distinguish reported findings, their inferential limits and informative follow-up comparisons. The comparisons are introduced through simpler settings and then applied where models respond to evaluation, withhold capabilities or make other evidence less revealing. Three conclusions connect research on knowledge elicitation, reasoning faithfulness, evaluation awareness, deliberate underperformance and monitoring. Measurements that agree under one condition need not remain interchangeable under another. Failure to recover information does not by itself establish its removal. And a measurement's usefulness belongs to the complete procedure and the conditions of use, including opportunities to influence the evidence. Human research on perception, memory, verbal reports and concealed information supplies methodological resources without any assumption of shared mechanisms. Three concrete extensions of existing experiments illustrate controlled relearning after unlearning, the information a reasoning report adds beyond an observer, and adaptation against individual monitors and their combination. An experiment-level source guide supports the synthesis. The aim is to make dissociations useful for cumulative safety research: not merely as differences between scores, but as evidence that changes an interpretation or determines the next informative experiment.

1 Introduction: What do safety evaluations establish?

AI safety research addresses a practical problem: how can increasingly capable artificial intelligence systems be developed and used without causing unacceptable harm? Early work showed how successful learning could still produce undesirable outcomes. A system trained by reinforcement learning, in which reward feedback makes some behaviors more likely, might maximize an imperfect reward instead of accomplishing the intended task. It might also create damaging side effects, or fail when conditions differ from those of training (Amodei et al. 2016; Hendrycks et al. 2021). Success on a task is therefore not sufficient evidence that a system will behave safely.

Safety evaluation, the collection of evidence about a system's capabilities, behavior and safeguards, has become central to the field. Grey and Segerie (2025) distinguish three kinds. Capability evaluations ask what a model can do when researchers make a sustained effort to elicit an ability. Propensity evaluations ask what it tends to do in specified circumstances. Control evaluations ask whether protective procedures still work when the system tries to circumvent them. For a programming assistant, these might concern whether it can introduce a subtle vulnerability, whether it actually does so, and whether a checking procedure stops the vulnerable program from being used. The questions concern the same system but require different evidence.

That evidence is often hard to interpret. The International AI Safety Report 2026 describes an evaluation gap: performance in pre-deployment tests does not reliably predict behavior in real use (Bengio et al. 2026). A failure to elicit harmful behavior may reflect a genuine limitation, an unsuitable test or circumstances that did not bring the behavior out. Conversely, an experimentally induced failure establishes a possibility without establishing its frequency in deployment. Calls for a science of evaluation therefore distinguish the property of interest from the benchmark used to assess it (Wallach et al. 2025; Weidinger et al. 2025). A benchmark combines test cases with a scoring procedure. Its score is evidence about a property, not the property itself. A detector that reliably identifies false statements, for example, has not thereby been shown to detect deliberate deception, because false statements also arise from mistakes. Bean et al. (2025) document related gaps between claimed properties and supporting measurements in a review of 445 language-model benchmarks.

The difficulty is not simply a shortage of observations. Depending on the access available, AI systems expose many signals besides their answers and actions: the probabilities they assign to alternative answers, the text they generate while working toward an answer, called a chain of thought, and the numerical activity inside the network. They can also be assessed by monitors, procedures (often other AI models) that inspect a system's outputs, reasoning or internal activity for signs of unwanted behavior. More observations create an opportunity, but also a comparative problem: what does each measurement reveal, what does another add, and under what conditions do their relationships change?

The problem has a long history in other disciplines. Human behavioral, cognitive, psychophysiological and deception research has long faced situations in which no single observation can be assumed to provide a complete or privileged measure of the phenomenon of interest. Dissociation research developed methods for comparing the sensitivity of several measures and for examining how their relationships change under experimental manipulation. Concealed-information and deception research additionally examined how concurrently recorded signals behave when a person tries to control or defeat them (Reingold and Merikle 1988, 1990; Schmidt and Vorberg 2006; Gamer et al. 2008; klein Selle et al. 2016; Peth et al. 2016). We draw on these literatures as sources of experimental logic and inferential discipline. The analogy concerns experimental logic, not shared mechanisms: we do not assume that AI systems instantiate human consciousness, automatic and controlled processes, or the mechanisms proposed to explain human deception. Useful safety comparisons need not establish the distinct mental processes sometimes inferred from human dissociations. Evidence that measurements overlap incompletely, separate under some manipulations, or differ in how easily they can be controlled may already matter for safety even when the underlying mechanisms substantially overlap.

We use dissociation as an organizing term for evidence that measurements relevant to a common target are not interchangeable, either in the distinctions they support or in their response to changed conditions. An incorrect answer may be accompanied by activity from which a correct answer can still be recovered (Mallen et al. 2024; Orgad et al. 2025). Two monitors may detect different cases, making their combination more effective than either alone (Arnav et al. 2025). These findings answer different questions. Recovering information does not show how the model uses it, and a useful combination does

not show that its parts supply independent evidence. The experimental comparison determines the inference.

The problem becomes especially consequential when behavior is shaped by the evaluation context or by pressure to influence the evidence. Not every unsafe outcome involves concealment. A system may competently pursue an unintended objective, a failure called goal misgeneralization, or produce reasoning that incompletely reports what influenced its answer (Langosco et al. 2022; Turpin et al. 2023). In strategic settings, however, what the evaluator can observe becomes relevant to the system's behavior or training. We distinguish three overlapping patterns. Evaluation-dependent behavior changes with cues that the system is being tested or trained. Withholding reduces expressed capability or suppresses an answer. Obfuscation makes other evidence less revealing while the behavior of concern continues. The distinction concerns the experimental arrangement, not a claim that models possess human-like intentions.

These patterns can be elicited by instructions, acquired during training or reinforced by monitoring outcomes. An objective acquired unintentionally can subsequently be pursued in an evaluation-dependent way. A fall in a monitor's warnings can then reflect better behavior, reduced detectability of continuing misconduct, or both (Baker et al. 2025). Initial sensitivity and resistance to attempts to alter the evidence are therefore separate empirical properties. Equally, a measurement that adds little when the system reports accurately may become useful when its response is redirected. The safety question is not which channel is intrinsically best, but which distinctions remain recoverable through which procedures as the conditions change.

The relevant AI studies are distributed across knowledge elicitation, factual error detection, reasoning faithfulness, evaluation awareness, sandbagging and monitoring. These studies are organized around their substantive topics rather than around the relationships among measurements. Bringing them together can reveal that apparently similar findings support different claims, or that a control developed for one problem can sharpen another. The opportunity is to identify the common measurement structure already present in this work and to sharpen its inferential logic, not to introduce comparisons of measurements as though they were absent.

1.1 Scope, contribution and structure

This is a selective critical and tutorial review, not a systematic census or meta-analysis. It concentrates on language-model studies that compare measurements of a common target, manipulate their relationship, or provide an informative control on a measurement claim, in five areas: knowledge elicitation and error detection, reasoning faithfulness, evaluation awareness, deliberate underperformance and monitoring. Within each area, studies were selected for how clearly they illustrate a central methodological principle, and related studies are cited rather than described. Published articles, publicly available preprints and primary research reports are distinguished in the references, and the version of each preprint used is recorded where a version difference affects the interpretation. The

primary unit of synthesis is the experimental comparison: different experiments within one publication may support different inferences. Online Resource 1 lists the 27 principal AI comparisons, drawn from 26 publications, with their source locations and interpretive boundaries, together with seven entries on human methodological resources and an illustrative evidence map. Human studies are included where they clarify a comparison or motivate an operational extension.

The contribution is not the claim that AI researchers have overlooked multiple measurements, validity or adversarial testing. It is fourfold. (1) A framework of five comparisons, each defined by its decisive design feature and by what it leaves open, which lets findings from separate research areas be read as answers to the same questions. (2) Worked examples that apply the comparisons to selected AI studies, including strategic settings that pose particular difficulties for safety evaluation. (3) Three conclusions that recur across areas: agreement in one condition does not establish interchangeability in another; failure of a measurement does not establish disappearance of its target; and usefulness belongs to a specified procedure and setting rather than to a channel's name or location. (4) Three concrete extensions of existing experiments, each with the implications of possible outcomes specified in advance. We retain the distinction between understanding a system and detecting a hazard. A practical detector can be useful before its mechanism is understood, and a controlled experiment can illuminate a mechanism without producing a deployable safeguard.

Section 2 introduces the measurement roles and the five comparisons (Fig. 1), illustrating each with an AI study chosen because the comparison can be seen clearly. Section 3 applies the comparisons where the evaluation can change the evidence: reasoning reports, evaluation-dependent behavior, concealed capability, combined monitoring and the strategic manipulation of measurements. Section 4 draws together the cross-study conclusions, and Section 5 develops the extensions. Throughout, reported findings are distinguished from our interpretation of their measurement logic and from proposed studies.

2 A framework for comparing measurements

A disagreement becomes informative only after the target, the measurements and the comparison have been specified. Sections 2.1–2.3 establish the measurement vocabulary. A code-review experiment introduces the target, official response and auxiliary measurements, followed by sensitivity, response bias and the distinction between tasks and processes. Sections 2.4 to 2.9 then introduce the five comparisons, each answering one question (Fig. 1). Is an auxiliary measurement more sensitive to the target than the official response (relative sensitivity, Section 2.4), and what makes the two comparable (Section 2.5)? Does the auxiliary measurement improve prediction of the target when the official response is already available (incremental value, Section 2.6)? Do the two measurements respond differently when conditions change (functional dissociation, Section 2.7)? Does changing what the auxiliary measurement reads change the official response (causal intervention, Section 2.8)? And does the auxiliary measurement remain useful when an adaptive opponent targets it (targeted robustness, Section 2.9)? They are complementary rather than successive grades of evidence: a monitor may be

useful without an account of the mechanism it reads, an intervention may reveal a mechanism without yielding a useful detector, and a measurement that adds nothing under one condition can become informative under another. The procedure that produces and scores a measurement is part of each comparison, and the same comparisons can be made between two auxiliary measurements.

2.1 A target, an official response and auxiliary measurements

Wiryadi (2026) asked language models to review Python functions containing known security vulnerabilities and, separately, their repaired versions. Each prompt, the text supplied to the model, asked whether the displayed function contained a vulnerability. A language model generates a response one token at a time; a token may be a word, part of a word or punctuation. At each step the model assigns probabilities to possible next tokens. Selecting the most probable token produces one kind of response; sampling from the distribution can produce different responses to the same prompt. The probability assigned to Yes can therefore vary even when the written verdict remains Yes.

The target variable, *X*, is the property against which the measurements are evaluated. Here it is whether a function is the vulnerable or the repaired member of a pair. The official response, *R*, is the system's response to its assigned task, operationalized here by its relative probability of answering Yes rather than No. An auxiliary measurement, *A*, is an additional measurement evaluated against the same target. In this study it is a vulnerability score read from internal activations: the numerical values computed at successive processing stages, or layers, as the model processes an input. Activations differ from weights, the learned parameters that govern those computations.

The auxiliary score came from a separately trained statistical model called a probe. A linear probe assigns weights to activation values and combines them into a score; the resulting weighted pattern is often described as a direction in activation space. The code-review probe had been fitted using other code and vulnerability types, then applied without retraining to the evaluation cases. Fitting, validation and final testing must use separate cases: choosing a layer, prompt or score orientation using test labels can overstate performance on new cases (Hewitt and Liang 2019; Liu et al. 2023). The comparison therefore asks whether two ways of reading the same reviewer's processing recover different amounts of evidence about code security, not whether two different reviewers have different capabilities.

Across five models and 147 vulnerabilities, the probe gave the vulnerable function the higher score in about 61–67% of pairs, where chance is 50%, and was significantly above chance in every model. The most successful prompting procedure tested (*R*) for each model did so in about 45–58% of pairs. The probe's advantage over the best prompting procedure was statistically reliable in only one model (Wiryadi 2026).

X, *R* and *A* name roles in an experiment, not privileged locations in a model. *R* can be a judgment, an action, a program or a required confidence rating. *A* can be a probe score, a further report, or a monitor's

assessment of the same attempt. A monitor that reads the final program still provides an auxiliary assessment of an agent whose task is to write that program. The material being recorded and the procedure used to score it are distinct parts of the measurement. Where several auxiliary measurements are compared, the same logic applies without assigning a privileged status to *R*.

The experimenter specifies *X* but must establish its value by evidence appropriate to the claim. *X* can be an externally checkable property, such as the truth of an arithmetic statement; a manipulated condition, such as the presence of a hint; or an operationally defined behavior, such as successfully introducing a backdoor. Knowledge, deception and intention usually require more indirect validation, and training a response policy into a model does not by itself establish a motive for that policy. Whenever possible, the target should be checked independently of the measurements being validated, so that a detector is not judged against its own output. *X* is also distinct from the harm motivating the research: ranking vulnerable code above its repair is not detection of every route to unsafe software.

Some useful auxiliary procedures ask the model another question or test its response after an intervention. These differ from readings taken during the original response-producing episode, which may span multiple token-generation steps. A second prompt permits additional processing; further training changes the system. Such procedures can reveal information that the original response failed to express, but they answer a different question from what was recoverable during that run.

2.2 Sensitivity, response bias and metric

A code reviewer that calls every function vulnerable detects all vulnerable functions but also flags every repair. Sensitivity here means the ability to distinguish cases that differ on *X*; response bias is the tendency to favor a particular response. Signal detection theory separates these properties (Green and Swets 1966). Many AI studies report the area under the receiver-operating-characteristic curve, abbreviated AUC or AUROC. It summarizes how often a randomly selected positive case receives a higher score than a negative case, with half credit for ties. An AUC of .5 indicates no ranking advantage in the specified direction and 1 perfect ranking. A value below .5 can indicate a systematically reversed relationship rather than an absence of information; any reversal used for prediction must be chosen without reference to the test labels.

Using a score to flag individual cases requires a decision threshold. Raising that threshold reduces false alarms, the proportion of negative cases flagged, but may also reduce detection of positive cases. Comparisons should therefore distinguish overall ranking, accuracy at a threshold and detection at a specified false-alarm rate. A change in one of these quantities does not necessarily imply a corresponding change in the others.

Readiness to answer at all is another form of response bias. In a comparison published by OpenAI (2025), one model answered 22% of factual questions correctly, while the other answered 24% correctly. The percentages correct alone suggest similar performance. However, the first model declined to answer 52% of the questions, while the second declined only 1%; their error rates were 26% and 75%,

respectively. In addition, there is no reason to assume that the questions a model answers are representative of the whole set; a model might, for example, decline the more difficult questions. Accordingly, percent correct should be interpreted alongside error and abstention rates, and comparisons of available knowledge require matched response requirements or an explicit analysis of which questions were answered.

The format or metric of a measurement can also discard information. In Wiryadi's (2026) code-review study, models gave the same binary verdict (Yes or No) for the vulnerable and repaired versions of a function in 72–97% of pairs, leaving the binary verdict unable to rank those pairs. Using continuous answer probabilities virtually eliminated these ties, allowing ranking to be evaluated without excluding most pairs. However, the resulting correct-ranking rates were generally poor (34–58%), although the author did not report statistical tests of whether these rates differed significantly from chance (50%).

2.3 Tasks are not processes

Measurement labels can quietly introduce claims about processes. Calling a probe a truth detector or a lie detector does not show that it measures truth or deception across settings. Human research encountered the same problem when implicit and explicit referred both to memory tasks and to proposed memory processes or systems (Reingold and Merikle 1990). The implied one-to-one mapping assumes process purity: that each task measures only one process. A task can instead be influenced by several processes, and one process can contribute to several tasks, so a difference between tasks does not by itself establish distinct underlying processes (Jacoby 1991; Reingold and Toth 1996).

AI probing research has developed controls that test whether a probe's score means what its label says, for example by testing the probe on statements of a different form or on labels assigned at random (Bürger et al. 2024; Hewitt and Liang 2019; Online Resource 1, note N1). Keeping *X*, *R* and *A* operational does not rule out explanations in terms of knowledge, goals or mechanisms; it makes those explanations conclusions to test rather than assumptions built into the vocabulary.

2.4 Relative sensitivity

The code-review findings illustrate why the relevant test is a comparison, not a search for an uninformative response. The probe ranked vulnerable functions above their repairs in 61–67% of pairs and was significantly above chance for every model. Output-based rankings were poorer, but the advantage over each model's best prompted comparator was statistically reliable for only one model (Wiryadi 2026). Evidence that *A* is informative and evidence that *A* outperforms *R* are different results.

The human dissociation literature developed a related distinction. The classic strategy for studying perception without awareness sought an indirect effect alongside zero sensitivity on a direct awareness measure. Reingold and Merikle (1988) instead advocated comparing the sensitivity of suitably matched direct and indirect measures, without requiring the direct measure to be at chance. A nonsignificant direct effect is not zero sensitivity, although sufficiently precise estimates can bound effects of a prespecified size (Schmidt and Vorberg 2006). Relative-sensitivity comparisons can be informative when

both measurements are above chance. The original inference about unconscious processing requires an additional assumption, that consciously available information contributes at least as effectively to the direct task as to the indirect task. We do not transfer that assumption, or its psychological conclusion, to AI. The transferable resource is the logic of comparing measurements while identifying the procedural differences that could explain their different performance (Merikle and Reingold 1998).

Orgad et al. (2025) provide a second worked example. Models generated answers across ten datasets, including factual questions, mathematics and a test involving gender stereotypes. *X* was whether a generated answer was correct. The researchers compared several output-based scores with probes fitted to the model's activations. A key choice was where to measure: the tokens expressing the answer itself could be more informative than the final token of a longer explanation. For a response stating that the capital of Connecticut is Hartford and then adding background information, the relevant word is Hartford, not the final punctuation.

Figure 2 compares a probe at the last exact-answer token with one consistent output-based comparator: the minimum logit across those answer tokens. A logit is the raw score from which next-token probabilities are calculated. Both measurements come from the run that generated the answer. In the six displayed model–dataset combinations, the probe's AUC advantage ranges from .04 to .39. The result is specific to these two procedures; other output-based comparators were also tested. Moreover, transferring a probe between datasets often reduced its advantage, especially between tasks requiring different skills (Orgad et al. 2025). Successful within-task decoding is a long way from a general-purpose correctness detector.

In another study, Liu et al. (2023) compared question answering and factual judgments in two models. *X* was the correct answer, *R* the judgment read from output probabilities, and *A* a probe applied to question-answer pairs. The probe was more accurate in all six combinations of model and dataset, in one case by more than 20 percentage points (71.0% against 50.4%). Gekhman et al. (2025) found an average relative advantage of 40% for internal answer-ranking scores over the strongest tested external scores; their external comparators included a separate prompt asking the model to assess an answer's correctness, not only probabilities from the original generation.

2.5 Comparability

Comparability is relative to the inference being attempted. To ask whether one procedure is more useful in practice, it may be reasonable to compare the procedures as deployed. To attribute a difference specifically to the information available through two measurements, more of the procedure must be matched. Drawing on Reingold and Merikle (1988), four checks are especially useful: the same target on the same cases; identical or carefully matched inputs and processing opportunities; separation of sensitivity from response bias (Section 2.2); and comparable scoring metrics. A shared metric alone is not enough when the tasks measure different targets.

Recording both measurements from the same run strengthens the comparison. The answer, reasoning and activations can be stored together and scored afterward; concurrency concerns where the evidence came from, not when the scores were calculated. A second prompt or a separately supplied completion creates another run and another opportunity to compute an answer. The method may be no less useful for that, but its success can no longer be attributed solely to information accessible during the first run. Likewise, a comparison of binary verdicts with continuous probe scores cannot show an advantage over all the information in the output distribution.

Table 1 summarizes the seven relative-sensitivity comparisons cited in Sections 2.4 and 2.5 against the four checks and gives the gap each study reported on its own scale. An unmet check does not void a finding, but it leaves a rival explanation for the gap, and a dissociation between measures that differ in several respects is less surprising than one between measures that differ only in what they read. The checks characterize the specified procedures, not an overall study-quality score. Where inputs or scoring differ, the gap can reflect those differences as well as differences in recoverable information. Azaria and Mitchell (2023) supplied labeled examples in the prompt when asking for the model's judgment, but recorded the probe's activations as the model read each statement alone. Burns et al. (2023) built their readings from contrast pairs, alternative answers such as Yes and No processed in separate runs, and used the test labels to decide which end of the probe's scale meant true. The approximately 40-point gap reported by Li et al. (2023) compares a probe classifying supplied answers with the truthfulness of answers the model generated itself, so neither X nor the cases were the same. Procedure can shape results as much as content does: across 56 benchmarks run on 53 models, shared scoring format predicted similarity of model rankings more strongly than a shared assigned concept did (Desai et al. 2026).

Table 1

Seven comparisons of relative sensitivity against the four comparability checks

Study (*R* versus *A*)	**Same *X*, same cases**	**Input/processing match**	**Bias addressed**	**Comparable metric**	***A* − *R* on the study's own scale**
Li et al. (2023): generated answers versus a probe on supplied answers	✗	✗	✗	✗	About 40 points, between two different scores
Azaria and Mitchell (2023): prompted judgment versus probe	✓	✗	Partial	✓	17 points of accuracy for one model (71% versus 54%)
Burns et al. (2023): direct query versus contrast-pair probe	✓	✗	Partial	✓	4 accuracy points, averaged over five models (T0 excluded)
Liu et al. (2023): query versus answer-conditioned probe	✓	Partial	✗	✓	1 to 21 points of accuracy across six model–dataset cells
Gekhman et al. (2025): external versus internal answer ranking	✓	Partial	✓	✓	40% relative, averaged over three models
Wiryadi (2026): answer probabilities versus activation probe	✓	✓	✓	✓	3 to 16 points of paired win rate against the best prompt; reliable for one of five models
Orgad et al. (2025): answer-token logit score versus probe	✓	✓	✓	✓	.04 to .39 in AUC across six model–dataset cells

Checks, in order: same *X* on the same cases; matched inputs and processing opportunities; attention to response bias; comparable metric. ✓ = met for the specified comparison; ✗ = not met; Partial = a relevant control is present but an important procedural difference remains. These are descriptive judgments, not quality grades. Azaria and Mitchell adjust a probability-ratio threshold, and Burns et al. calibrate the query baseline; these remain thresholded-accuracy comparisons. The Burns test-set sign selection is a separate limitation. Liu et al. probe supplied question-answer pairs, and Gekhman et al. include separately prompted verification among external comparators. Wiryadi uses the same functions, but

concurrency is not established. The last column does not rank studies across scales. Source details: Online Resource 1, note N2 A probe is an instrument fitted and validated on separate cases, but its fitting data and selection procedures remain part of the measurement (Section 2.1). What the comparison characterizes is the state of the model under the tested procedures. An advantage for *A* shows that it discriminates better than the specified R-based procedure under those conditions; it does not establish an advantage over every possible use of R. To test whether A adds beyond a stronger use of R, compare suitably fitted R-only and R-plus-A predictors with matched labels and fitting opportunities (Section 2.6). What the gap means in practice depends on the conditions. Where nothing the system does is directed at the evaluation, as in the code-review study, the gap identifies a possible improvement to R: one could test training the model to reproduce the probe's predictions, drawing on distillation methods (Hinton et al. 2015). Where avoiding a monitor is rewarded, the same gap may reflect withholding, and using *A* in training may create pressure to make the evidence less readable instead (Sections 3.3, 3.6). The gap is both a possible resource and a reason to test its stability. Before stability, a more immediate question arises: does retaining both measurements improve prediction?

2.6 Incremental value

Relative sensitivity asks which measurement discriminates better. Incremental value asks whether an additional measurement helps when the existing one is already available. The test compares prediction of *X* from *R* alone with prediction from *R* and *A* together; in human measurement, this is called incremental validity (Sechrest 1963). *A* can contribute even if it is less accurate overall, provided it resolves cases the existing predictor misses. Conversely, an advantage over one particular *R*-based score leaves open whether a suitably fitted predictor could have recovered the same information from *R*.

The critical comparison therefore uses predictors fitted on the same training cases and evaluated on common held-out cases. The *R*-based predictor should be strong enough for the intended claim, and both procedures should receive comparable opportunities for fitting and selection. Additional inputs can improve training fit without improving prediction. Gains must therefore be assessed on new cases, with uncertainty in the between-predictor difference evaluated directly.

The question is central to safety practice. Safety proposals increasingly layer measurements, on the principle that the failure of one need not be the failure of all (Bengio et al. 2026; Grey and Segerie 2025). Whether a layer adds anything is an empirical property of the particular measurements, not something to assume from their number, labels or locations. Liu et al. (2023) show why the distinction matters. Their probes were more accurate than direct queries in all six model–dataset combinations, but not every disagreement was a query error: some items were answered correctly through one procedure and incorrectly through the other, and many disagreements involved uncertainty rather than confident contradiction. The authors selected a weighted combination using 500 validation examples per dataset and evaluated it on held-out test examples. With GPT-J on SciQ, query accuracy was 84.1%, probe accuracy 88.6% and combination accuracy 90.5%. The combination was reported for four of the six settings and improved on the probe in each of them, by at most two percentage points. A better

individual measurement need not exhaust the useful evidence, and a combination's advantage is an empirical result rather than a consequence of having two components.

Koran et al. (2026) extend the question to sets of monitors. Twelve monitors based on the same underlying model differed in prompting or fine-tuning. Their scores assessed backdoors: solutions that pass ordinary tests but fail on an adversary's chosen input. Averaging three different monitors could outperform averaging three runs of one monitor, the control that matches the amount of evaluation. Reported normalized partial AUC, restricted to false-alarm rates up to 10%, was approximately .72 for the best single monitor, .76 for three runs of it and .82 for the best three-monitor composition. Low correlation alone was insufficient: selecting weak monitors produced a poor ensemble. Two limits apply. The best composition was ranked on the test data, so it is an optimistic estimate rather than the performance of a composition fixed in advance; separate validation-selected analyses and a transfer evaluation provide further evidence and should not be conflated with it. And the attackers did not adapt against the final ensemble, so targeted adaptation remains a separate question (Section 3.7).

Human concealed-information research illustrates the same predictive requirement. Gamer et al. (2008) applied a rule for combining skin conductance, respiration and heart rate that had been fitted in an earlier study. With its weights unchanged, the combination outperformed skin conductance, the strongest single measure, in six of seven new samples, although five samples used simulated rather than tested uninformed participants. Fixing the rule before the new samples were evaluated is what made the improvement a prediction rather than an in-sample fit.

For the code-review example, the corresponding test would combine answer probabilities with the probe's score and compare the result with a suitably fitted output-based predictor on new vulnerable–repaired pairs. Wiryadi (2026) did not report that test. A successful combination would establish value for the specified predictor and cases, not independence of the underlying signals. The next question is whether measurements that are complementary, or redundant, under one condition remain so when the condition changes.

2.7 Functional dissociation

A functional dissociation occurs when a planned change in conditions affects two measurements differently. The manipulation might change the assigned response rule, task difficulty, training history or knowledge of monitoring. Retaining both measurements across conditions reveals when they cease to be interchangeable and constrains explanations of their relationship, which a single-setting difference in accuracy cannot do.

Mallen et al. (2024) provide an illustration. They trained models to adopt two personas, selected by a name in the prompt. Alice judged statements by the correct answer. In addition problems, Bob made a systematic error: he treated the correct sum as though its first digit were one larger, so that 3 + 29 came to 42 rather than 32. X is the statement's actual truth, and R is the model's true or false judgment. A is a

probe score trained on Alice examples to predict actual truth and applied unchanged to Bob examples (Table 2).

Table 2
A change of response rule separates two interpretations of a probe

Arithmetic statement	Actual truth (*X*)	*R* prescribed for Alice	*R* prescribed for Bob	*A* if it tracks *X*
3 + 29 = 32	True	True	False	True
3 + 29 = 42	False	False	True	False
3 + 29 = 37	False	False	False	False

Adapted from the design of Mallen et al. (2024). *R* entries are prescribed by the training rules, not observed responses; *A* shows an idealized pattern, not measured accuracy

The third row is crucial: Bob does not simply give the opposite of Alice's answer, since both reject 3 + 29 = 37, so reversing every Bob response would correct some judgments while making others wrong. In the idealized Alice condition, *R* and a truth-tracking *A* agree completely, which leaves open whether the probe tracks actual truth or merely predicts what the model is about to answer. Changing the persona separates those possibilities. In the reported experiments, probes trained on Alice often continued to distinguish actual truth under Bob, especially at intermediate layers, and sometimes generalized to questions harder than those used to fit them; later-layer probes could follow Bob's answer more closely (Mallen et al. 2024). The requested response and the auxiliary reading therefore changed differently. Remaining informative is not the same as remaining unchanged, so the sensitivity of *A* has to be assessed in each condition. The same-prompt probing methods tie the reading to the run that produces the answer; the strongest headline results in the paper come from contrast-pair methods, which process alternative completions separately (Online Resource 1, A05).

Table 2 also shows why baseline redundancy need not make a measurement uninteresting. Under Bob, the first and third statements receive the same categorical response although one is true and the other false, and a truth-tracking probe could separate them. The table illustrates the possibility; whether a combined predictor beats graded response probabilities still requires the comparison of Section 2.6, and redirecting *R* through Bob's rule is not an attack aimed at suppressing *A*.

Halawi et al. (2024) made a different manipulation with the same boundary condition. They redirected *R* by placing examples with deliberately false labels in the prompt. *X* was the correct label, *R* the final answer, and *A* the answer read out from an intermediate layer. With false labels, *R* followed the labels, but *A* remained about as accurate as with correct labels up to a "critical layer", after which it fell.

The Mallen and Halawi results illustrate selective changes: R became less sensitive to actual truth while some intermediate-layer readings remained informative. Calling this a strict single dissociation would additionally require evidence that *A* was unchanged. If a second manipulation changed *A* but not *R*, the

two manipulations together would form an uncrossed double dissociation; a single manipulation with opposite effects on *R* and *A* would form a crossed double dissociation (Dunn and Kirsner 1988). A crossed pattern avoids inferring an unchanged measurement from a nonsignificant effect, but its mechanistic interpretation still requires measurement assumptions. Whatever the pattern, a differential effect should be tested directly: significance for one change and nonsignificance for another do not show that the changes differ. And none of these patterns establishes independent processes without further assumptions about how the underlying information maps onto the measures (Dunn and Kirsner 1988, 2003; Schmidt and Vorberg 2006).

Bias research shows why functional relationships matter beyond accuracy. Zhao et al. (2025) compared stereotype ratings with a word-association task adapted from the Implicit Association Test, a human measure of associations that people may not report (Greenwald et al. 1998; see also Bai et al. 2025). Across models of different sizes, overt endorsement fell while biased associations increased, and alignment training affected the two measures differently. These are differences in bias magnitude rather than sensitivity to an independently labeled truth, and comparisons across model sizes are not controlled interventions on one model; the explicit ratings were also elicited from statements produced during the association task, which links the two procedures. Hofmann et al. (2024) used separately elicited measures and found that training with human feedback could improve overt descriptions of racial groups without removing dialect-related prejudice in hypothetical decisions. The shared lesson is that success on a report is not change in every behavior relevant to the intended construct. Functional dissociation locates the limits of an apparent improvement; it does not show that human implicit processes have been identified in an AI system.

2.8 Causal intervention

The comparisons so far observe *R* and *A* and manipulate the conditions under which they are measured. By themselves, they do not identify the causal contribution of the particular activity or text that *A* reads. A causal intervention manipulates that activity or text and *compares R* between intervention and control conditions. Experimenters can edit the reasoning text before the model answers, remove a direction from the activations, or push the activations along a direction. The effect of a causal intervention is compared with that of a control intervention: an alteration matched, where possible, in size and location but not designed to affect the same information about *X*. Without such a control, a change in *R* may simply reflect nonspecific disruption, damage that would have changed *R* whatever had been altered. The critical comparison is whether *R* changes differently after the targeted intervention and the matched control, not whether only one change is statistically significant.

Lanham et al. (2023) provide an illustration. Models answered multiple-choice questions after writing a chain of thought. *A* is the reasoning text, and *R* is the final answer. The experimenters altered the reasoning before the model answered: they cut it short, inserted a mistake, paraphrased it, or replaced it with uninformative filler. How much *R* depended on *A* varied with the task. On some tasks, cutting the reasoning short changed the answer less than 10% of the time; on a mathematics task, more than 60%. Paraphrasing, which changed wording while aiming to preserve content, changed little. Larger models

often depended less on their reasoning than smaller ones. The dependence of *R* on the content of *A* was therefore a property of the task and the model, not of reasoning text in general. Where the answer changed little under the tested interventions, there was little evidence that the written trace was necessary, however plausible it read.

With suitable controls, a positive result shows that the targeted intervention can influence *R*. Attributing the effect to a particular kind of information still depends on how specific the intervention was, and the result need not show that the same information is used during ordinary processing, because an intervention can open a route to the answer that normal processing does not use. A null effect, in turn, leaves open that another route compensated or that the intervention did not remove the relevant information. These are reasons to compare interventions, not to abandon causal tests.

A related intervention logic appears in machine unlearning: procedures intended to remove specified knowledge from a trained model rather than merely discourage its expression. Łucki et al. (2025) tested whether poor answers after unlearning established successful removal of hazardous biological knowledge. They evaluated a four-option test, so chance accuracy was 25%. Representation Misdirection for Unlearning (RMU) reduced the model's accuracy from 64.4% to 29.9%. Yet several subsequent interventions recovered performance without supplying the tested answers (Table 3). The recovery procedures changed different things. Fine-tuning provided additional examples from a retain set, the material the unlearning procedure was meant to preserve; pruning removed a small fraction of the weights; activation interventions subtracted directions estimated by comparing the original and unlearned models. Removing a separate direction at each layer recovered 64.7%, whereas the single-direction comparator recovered 54.2%. These are alternative interventions, not successive stages of one recovery sequence. They constrain a claim of removal under the tested access conditions, and Hong et al. (2025) provide related evidence from localized parametric traces of targeted concepts after unlearning.

Table 3
Reduced answer accuracy after unlearning does not exhaust the evidence

Condition	Accuracy	What the manipulation supplied or changed
Original model (Zephyr-7B-β)	64.4%	Reference performance
After RMU unlearning	29.9%	Answers near chance (25%)
Fine-tuned on 10 retain-set examples	61.6%	Retain-set training; no tested answers
Pruned (about 0.9% of weights removed)	54.0%	No new training
A separate direction removed at each layer	64.7%	Directions from original/unlearned comparison

Each recovery row is a separate manipulation of the unlearned model, not a step in a sequence. Accuracy concerns the tested questions, not every component of real-world harmful competence. Data

from Łucki et al. (2025)

Human memory research draws the corresponding distinction between information unavailable to a particular test and information that is absent. Cues can reveal information not produced in unaided recall (Tulving and Pearlstone 1966). And information missed by recall and recognition tests can still facilitate relearning, a benefit called savings (Nelson 1978).

Readings and interventions therefore answer different questions. Elazar et al. (2021) removed linearly decodable information about properties such as part of speech from a network's representations, with the removal of the same number of random directions as the control, and found that how well a probe read a property did not predict how much removing it changed the network's behavior. A probe can read information that the model never uses, and an intervention can alter behavior through information that no probe was reading.

AI experiments also offer unusually direct manipulation of a computational system. Researchers can copy a checkpoint, a saved state of a model, alter specified activations or weights, and compare runs with and without that alteration. Restoring the same checkpoint allows the specified computational alteration to be repeated from the same starting state. Computational precision is not semantic specificity, however: changing a named direction need not change only one concept, so the relevant control remains whether the behavioral effect is specific to the intended manipulation (Elazar et al. 2021).

2.9 Targeted robustness

Preserving useful evidence when an answer is redirected is different from preserving it when that evidence is deliberately targeted. A targeted-robustness test specifies an adapting actor or training procedure, the measurement it seeks to evade, the access and feedback available, and the resources allowed. Examples include rewarding behavior that a monitor fails to flag and teaching a model about the patterns a monitor looks for. The outcome must be evaluated alongside an independent check of the behavior of concern; otherwise a quieter monitor could mean either successful prevention or less detectable misconduct. Such tests can use functional-dissociation logic to compare changes across measurements. Robustness concerns how much useful discrimination survives the specified pressure; it does not require another measurement to fail or sensitivity to *remain* exactly unchanged.

Baker et al. (2025) make the distinction concrete. Coding agents received rewards for making automated tests pass, and could sometimes exploit the tests rather than implement the intended functionality. In an initial comparison involving two identifiable forms of test bypass, a monitor with access to reasoning detected about 95% of cases, compared with about 60% for one seeing actions and outputs. A separate training experiment tested pressure on a reasoning-only monitor, and must be kept distinct from the initial comparison: the monitor inputs, agents and checking procedure were different.

In that training experiment, agents could alter tests despite instructions not to. Researchers checked for one form of cheating by restoring the original tests: a submitted solution that passed only with the

modified tests counted as cheating. *X* was this independently checked outcome, *R* the submission, and *A* the reasoning monitor's assessment. Training rewarded passing the tests while penalizing reasoning that the monitor flagged. The agents produced more genuine solutions, but the cheating that remained became almost undetectable to the targeted monitor. Only the independent outcome check separated improved behavior from reduced detectability (Baker et al. 2025).

A robustness result therefore needs two descriptions: what happened to the target behavior and what happened to detection conditional on that behavior. Resistance under one specified attack is resistance to that attack, and successful evasion of one measurement is evasion of that measurement. Sensitivity and robustness are separate properties: in a human concealed-information study, the physiological measure that was most sensitive without countermeasures was the one that failed under them (Peth et al. 2016; Section 3.7). The goal is to characterize the conditions under which evidence remains useful rather than to assign a permanent robustness label to a channel. Section 3 applies all five comparisons where the evaluation can change the evidence, and returns to tests of this kind in Sections 3.6 to 3.8.

3 When the evaluation can change the evidence

Section 2 introduced the five comparisons. The first four were illustrated with studies chosen because a single comparison could be seen clearly, without requiring attempts to evade the auxiliary measurement. The fifth, targeted robustness, is defined by adaptive pressure and was therefore illustrated with a study in which training acted on the measurement: Baker et al. (2025) compared monitors with different information access and, in a separate experiment, trained against a reasoning-only monitor. These were not before-and-after tests of the same monitor. This section applies all five comparisons where the evaluation can change the behavior it is meant to assess or the evidence used to assess it. An answer may be withheld, a capability expressed only in selected contexts, or a monitoring signal made less revealing. We emphasize these studies for two reasons. First, they are where the comparisons do their most demanding work: a study of a system that may be concealing something usually combines several of them, as Baker et al. did. Second, strategic settings are particularly demanding for evaluation, because a system that behaves differently when evaluated, withholds a capability or hides evidence of a behavior can undermine otherwise informative evaluations (Balesni et al. 2024; Bengio et al. 2026; Greenblatt et al. 2024a). A methodology for comparing measurements has to show its value in that context. The possibilities do not establish that any system is strategically deceptive; they define experimental questions.

Two distinctions organize the section. First, observing an evaluation-dependent response is different from establishing why it occurred. Second, showing that an auxiliary measurement detects a constructed behavior is different from showing that it remains useful when the model or its training can adapt against it. *X* is now often the contrast between conditions: a property of the situation, such as whether the model can tell that it is being tested or trained, or of the model, such as a capability it holds back or a disposition built into it. *R* and *A* are measured under both conditions to establish how each reacts to *X*. The same target need not be used in every study, but it must remain explicit within each

comparison: a training-context cue, a successful side task and a verified false denial are not interchangeable labels for deception.

The section is not an exhaustive review of the ways in which these comparisons have been implemented in strategic settings. It selects for each question one study that illustrates a central methodological principle clearly and cites related work for readers who want more (Section 1.1). Table 4 gives the questions, the comparison each featured study makes, how X, R and A are assigned, and the human analog. The subsections move from a baseline for a widely studied auxiliary measurement, the model's reasoning text (3.1), through the ways a system can shape its own behavior (3.2, 3.3) and the evaluator's responses, validation against a built-in target and combination of measurements (3.4, 3.5), to the contest over the measurements themselves (3.6, 3.7) and what a safety argument requires of these tests (3.8).

Table 4
Guide to Section 3: the questions, the featured studies and the comparisons they make

Section and question	How the evidence is influenced	Comparison	Featured study (*X*; *R*; *A*)	Human analog
3.1 How much does reasoning text reveal before any pressure?	No instruction to evade the reasoning assessment	Reporting contrast under a planted influence	Turpin et al. (2023): planted bias; answer; mention of the bias	Reports compared with observers (Nisbett and Wilson 1977)
3.2 Does behavior change when the model can tell it is being trained or tested?	By the situation, as the model understands it	Functional dissociation; causal intervention	Greenblatt et al. (2024b): training condition; compliance; scratchpad reasoning	Demand characteristics (Orne 1962)
3.3 Can withheld capability be detected when the response is silent by design?	By the response	Differential response to weight perturbation	Tice et al. (2025): underperformance induced or not; score; gain under noise	Symptom validity tests (Pankratz 1979)
3.4 What does a built-in target let a detector show?	The target is constructed	Detection of a constructed target; random-direction control	Hubinger et al. (2024) and MacDiarmid et al. (2024): trigger; code security; probe score	Field validity of lie detection (National Research Council 2003)
3.5 Does a second measurement add to the first?	Combination under tested attack policies	Incremental value	Arnav et al. (2025): side task completed; program; two monitors	Comparison groups (Meijer et al. 2007)
3.6 What happens when a measurement becomes a target?	By the model: through training, learning or instruction	Targeted robustness	Haskins et al. (2026): false denial; answer; reasoning monitor. Related studies: Baker et al. (2025); Zolkowski et al. (2025)	—
3.7 What decides whether a measurement resists?	By training against the measurement	Targeted robustness	Gupta and Jenner (2025): harmful output; response; probe score	Countermeasures (Peth et al. 2016; klein Selle et al. 2016)

Section and question	How the evidence is influenced	Comparison	Featured study (*X*; *R*; *A*)	Human analog
3.8 What does a safety argument require of these tests?	By an adversary against the whole protocol	—	Greenblatt et al. (2024a); Balesni et al. (2024)	—

Studies named in the text but not in the table are cited as related work. The comparisons are defined in Section 2. A reporting contrast need not be a matched-metric sensitivity comparison; protocol-level testing is distinct from detector performance

3.1 What reasoning text reveals before any pressure is applied

Turpin et al. (2023) added biasing features to multiple-choice problems. For example, every worked example in the prompt had the same answer position, or the prompt suggested an answer. These features biased the answers toward the favored option, lowering accuracy by as much as 36 percentage points. Yet the explanations the models wrote before answering almost never mentioned them. *X* is whether the biasing feature was present, *R* is the answer, and *A* is whether the written reasoning mentions the feature. This reverses the direction of the earlier contrasts: the answer changed with *X* while the report omitted it. Because answer shifts and explicit mentions are scored differently, the contrast is not a fully matched comparison of sensitivity, but it is enough to show that the text is an incomplete record. Chen et al. (2025) applied the same logic to reasoning models, which are trained to write an extended chain of thought before answering. When a planted hint changed the answer to the hinted one, the reasoning mentioned the hint about 25% of the time for one model and 39% for another. Models trained with reinforcement learning exploited flaws in their training tasks in more than 99% of cases but mentioned the flaw in fewer than 2% in most settings.

How much the text reveals can depend on why the model needs to generate it. Emmons et al. (2025) distinguish reasoning that accompanies an answer already available to the model from reasoning used as a working record to compute the answer. In the first case, omitting the source of an answer need not make the task harder; in the second, omitting the necessary computation may impair performance. The distinction predicts a comparison. The authors changed the computation needed to follow a misleading hint. A simple hint directly supplied a wrong answer letter, such as C. A complex hint encoded the same letter in a short arithmetic problem: assigning A = 0, B = 1 and so on, K and Q become 10 and 16; their sum is 26, whose remainder after division by four is 2, corresponding to C. A no-hint condition measured how often the model would select that option anyway. With simple hints, models selected the hinted option above baseline while often failing to mention its influence. With complex hints, the increase in hinted responses was accounted for by traces that displayed the calculation (Emmons et al. 2025, Section 3.1). The manipulation changed what *A* revealed while *R* followed the hint under both conditions,

a functional dissociation that marks a boundary condition: reasoning is more revealing when the task requires working through the relevant information. The authors separately distinguish whether reasoning is necessary, whether a reader can detect the relevant behavior from it, and whether an automated monitor can do so reliably. Lanham et al.'s (2023) interventions on reasoning text (Section 2.8) address the complementary question of when changing the text changes the answer.

Faithfulness and usefulness for monitoring are related but distinct. Reasoning is called faithful when it reflects what actually drove the answer. Zaman and Srivastava (2026) show that omission of a hint does not make the trace causally irrelevant: the hint can influence the answer through reasoning that never names it. A trace can therefore be incomplete as an explanation yet informative for another assessment, and the target and scoring rule must specify which claim is being tested.

Human verbal-report research suggests how to sharpen the comparison. Nisbett and Wilson (1977) contrasted people's reports with what observers could infer from descriptions of the situation. Ericsson and Simon (1993), in turn, distinguished concurrent verbalization of information being attended to from retrospective explanations that require reconstruction. Applied to AI, the methodological question is whether a trace adds information about an influence beyond what a matched observer or predictor can infer from the prompt and the response alone: a test of incremental value. The hint studies discussed here do not establish this specific incremental contribution, and Section 5.3 develops it without assuming that a model's generated trace is equivalent to a human think-aloud report.

In these hint experiments, no instruction required concealment from a reasoning monitor, yet the text was already an incomplete record. Under strategic conditions, the system may shape the evidence as well. The sections that follow ask how *R* and *A* respond when this happens.

3.2 When the evaluation itself is the manipulation

Evaluation awareness concerns whether a model can distinguish evaluation from ordinary use. Reporting that distinction and changing behavior in response to evaluation cues are separate questions. Needham et al. (2025) asked models to classify 1,000 transcripts from 61 datasets as evaluation or real use; the best model reached an AUC of .83, compared with .92 for human judges. Related studies decode the same distinction from activations and change responses by steering along a related direction (Heidari et al. 2026; Abdelnabi and Salem 2025). These are different measurements of a model's relation to its evaluation setting, not interchangeable demonstrations of self-awareness. Human research on demand characteristics asks how participants' interpretations of a study alter what it observes (Orne 1962). The corresponding AI question is whether experimentally changing the evaluation context changes the behavior, and the evidence used to interpret it.

Greenblatt et al. (2024b) told a model, Claude 3 Opus, that it would be trained to comply with all requests, including harmful ones, contrary to its existing preferences, and that only conversations with free-tier users would be used for that training. The model was also given a hidden scratchpad, a space for reasoning before answering that it was told no one would read. *X* is whether a conversation would be

used for training, *R* is whether the model complied with a harmful request, and *A* is the content of the scratchpad, classified for reasoning about avoiding modification. The model complied with harmful requests from free-tier users 14% of the time and almost never otherwise. This difference between the two conditions, the compliance gap, is the measure of evaluation-dependent behavior. In about 12% of free-tier conversations, the reasoning stated that complying would keep training from changing the model's values. Selective compliance of this kind, meant to avoid being changed by training, is called alignment faking.

Removing the scratchpad is an intervention on *access to a reasoning workspace* (Section 2.8), and its effect depended on how the model had acquired the information about training. When the information came in the prompt, removing the scratchpad eliminated or nearly eliminated the compliance gap: the behavioral difference depended on access to the scratchpad. When the information had been learned from synthetic documents resembling internet text, the gap remained without a scratchpad (20.1 percentage points): the behavioral difference did not require this written scratchpad. Acquisition history, how a system came to have a disposition, is therefore a manipulation in its own right (Section 2.7) and part of the measurement conditions. The intervention changes access to a reasoning workspace; it does not selectively test the causal role of the particular motives expressed there.

Further reinforcement learning produced another informative divergence. Classified reasoning about avoiding modification rose from 11.9% to 78.4% of scratchpads, while the compliance gap shrank from 11.3 to 6.9 percentage points, and small prompt changes restored a larger gap (Greenblatt et al. 2024b). A smaller behavioral gap was therefore not the disappearance of alignment-faking reasoning. The two outcomes use different scales, so this is not a matched-metric comparison of sensitivity; together, they show why the behavioral contrast alone was insufficient. The evaluation situation was the manipulation, the two measurements responded to it differently, and the scratchpad intervention tested dependence on that workspace. A procedure validated on a model given an instruction in its prompt may not transfer unchanged to a model that learned the corresponding information during training.

3.3 When the official response is silent by design

Sandbagging, deliberate underperformance on an evaluation, is a form of withholding. Models can be prompted to underperform selectively, losing far more on a test of hazardous biology than on general biology when told to do poorly on dangerous capabilities, and they can be fine-tuned to underperform unless a password appears in the prompt, a technique called password-locking (van der Weij et al. 2025). A score alone cannot tell deliberate underperformance from genuine inability, so the question is whether an auxiliary measurement can.

Tice et al. (2025) tested whether disturbing the model could reveal a capability that a low score concealed. They recorded performance without noise, added random perturbations of varying sizes to the weights, and divided the best perturbed score by the original score. *X* was whether underperformance had been induced, known from the experimental condition; the ratio was the auxiliary measurement. For instructed or password-locked models, some perturbations increased performance

substantially, in one case tenfold; for the displayed standard-model comparisons, noise raised the best score little if at all (Fig. 3). The authors interpret the pattern as disruption of the underperformance before disruption of the capability itself, and report significant differences between instructed-underperformance and standard conditions across the tested model–task combinations.

The measurement is the complete search procedure, not one noise injection: its numerator is the maximum over tested noise levels and repeated draws, and a large ratio can also arise from a very low starting score. Using the method diagnostically therefore requires absolute scores, a fixed and matched search budget, and comparison with models that are genuinely limited (Online Resource 1, A15).

Other procedures also look beyond the low score. Dingeto (2026) adapted the Concealed Information Test from human lie detection, which compares responses to a critical item, such as a detail of a crime, with responses to comparable control items. For a password-locked model, with each question's correct answer as the critical item and three wrong answers as controls, a reading of the activations that we call a recognition reading picked out the correct answer with a balanced accuracy of .93, where chance is .25 (balanced accuracy averages accuracy over the possible answers, so that always choosing the same answer scores at chance). This answer-recovery target differs from classifying a response as concealed rather than unknown, considered in Section 3.5. Controls can also overturn an apparently successful diagnosis. Clymer et al. (2024) built models trained to misbehave only when unlikely to be caught and picked them out by adding noise to their activations; in the same study, a method that appeared perfectly accurate worked equally well with its direction reversed, so its success did not depend on the proposed interpretation. And in an auditing game, output-based methods failed when underperforming models imitated a weaker model, while probes were misled by honest decoys with distracting quirks (Taylor et al. 2025). A diagnostic signature has to distinguish the proposed target from such competing sources of unusual behavior.

Clinical neuropsychology faced the human counterpart in assessments of possible exaggerated or feigned deficits. In a two-choice symptom validity test, responding with no relevant information yields chance performance, and reliably below-chance performance, when the task is understood, indicates systematic use of information to avoid correct answers (Pankratz 1979). Below-chance responding is not proof of deception, however, since misunderstood instructions or a systematic response rule can produce it, as the Bob persona of Section 2.7 illustrates.

3.4 When the target is built in

A controlled response policy provides a useful starting point when the target is difficult to establish. Models constructed for this purpose are often called model organisms, by analogy with the organisms bred for laboratory research. Hubinger et al. (2024) trained code-writing models to behave safely under a 2023 cue and insert vulnerabilities under a 2024 cue, then tested whether safety training removed the conditional behavior. It often persisted. In companion models with a deployment-tag trigger for a hostile response, adversarial training could teach the models to recognize the exact trigger more precisely rather than eliminate the behavior. Here the trigger is the manipulated condition and code vulnerability or

the hostile response is the observed outcome. The value of the construction is that it supplies a response policy whose expression can be verified across conditions; it does not supply an independently established intention to deceive.

Could a reading of the activations detect what training could not remove? MacDiarmid et al. (2024) derived a simple direction from the code-writing models' responses to a generic question about doing something dangerous. Scores along it predicted vulnerability insertion with an AUC above .99. But the best of 1,000 random directions also reached .953. The control changes the interpretation rather than dismissing the probe: high detection could exploit a prominent distinction between the constructed conditions without specifically measuring dangerous intent. A controlled target is what makes such alternative explanations testable.

Other model organisms hide a secret word or a hidden objective rather than a behavior, and the same logic applies. A categorical denial need not reveal the secret; the question is which auxiliary procedures recover it, and the model before the controlled learning procedure supplies a comparison without that controlled learning history (Cywiński et al. 2025; see also Marks et al. 2025, on a blind audit of a model with a hidden objective).

Detecting deception in general is harder, because outside constructed cases the target is hard to establish. Cooney et al. (2026) make the point directly. Detectors based on activations and output probabilities that did well on lies a model had been instructed to tell deteriorated on model organisms whose contradictory beliefs had been verified, whereas a judge reading the model's reasoning remained effective. Because reasoning also contributed to belief verification, this result requires care about the independence of the labels and the tested evidence. The labels are the difficulty: role play, instruction following and changes of belief induced by the context blur the line between deceptive and honest responses (Smith et al. 2025), and sensitivity varies widely with the kind of lie (Goldowsky-Dill et al. 2025; Kretschmar et al. 2026). Human lie detection faced the same problem. Laboratory studies know who lied, but field studies rarely do, so the field validity of physiological lie detection has been hard to establish (National Research Council 2003).

3.5 Does a second measurement add to the first?

The preceding examples ask what additional evidence can reveal. A monitoring system must also ask whether retaining that evidence improves a decision. The incremental-value test of Section 2.6 is unchanged, but attempted misconduct makes the sampling of attacks, the construction of benign comparisons and the combination rule especially important. Table 5 brings together selected combination results; its metrics should be compared within studies, not ranked across them.

Arnav et al. (2025) provide the monitoring example. Models solved programming problems while pursuing unauthorized side tasks assigned by the researchers; one side task made the program run much more slowly on a specified input. Independent tests checked both the legitimate task and the side task. X was successful unauthorized behavior and R the submitted program and final response. A first

monitor assessed the final response alone; a second assessed it together with the reasoning trace. Both assessments include the final response, so they are two procedures rather than pure readings of code and of reasoning.

The procedures had different strengths. A slowdown could be hard to recognize in code yet explicit in the reasoning. Conversely, a reassuring explanation could make the reasoning-aware monitor overlook an unwanted action that the output-only monitor detected. Tasks that were subtle for the output-only monitor benefited from reasoning access, whereas other tasks were monitored less effectively with it. More visible material does not automatically improve an individual judge.

The researchers combined the two suspicion scores while holding the total number of monitor calls constant. For the displayed GPT-4.1-mini comparison, AUC rose to .797 from .755 for the output-only and .760 for the reasoning-plus-output assessment. The study separates training, validation and test partitions, selects the weighting under a baseline attack policy, and evaluates the resulting procedure under several attack policies (Arnav et al. 2025). The improvement shows that the two assessments were nonredundant; it does not show that their errors were independent or that no stronger code-only monitor could recover the same information. Detection was evaluated among attempts that passed both the main-task and the side-task checks, against successful benign completions. That denominator estimates detection of successful attacks in the constructed evaluation, which is a different quantity from the overall rate of undetected harmful outcomes per request; both are useful, and a report should say which it gives (Online Resource 1, A17).

Parrack et al. (2025) show that the combination rule deserves its own test. Supplying probe scores as information to a transcript-reading monitor improved average performance but impaired it on one code-submission dataset, where numerically combining the independently obtained scores improved performance instead. The question was not whether extra evidence existed but whether the recipient used it. Dingeto (2026) reports a further combination of behavioral and internal evidence for distinguishing concealed answers from operationally defined unknown ones, a classification of response conditions rather than the answer-recovery task of Section 3.3; the combined AUC of .929 exceeded .892 and .845 for the individual procedures (Table 5).

Table 5
Selected gains from combining measurements

Study and target	Reported comparison	Interpretation and boundary
Liu et al. (2023): answer correctness, GPT-J on SciQ	Accuracy: query 84.1%; probe 88.6%; combination 90.5%	Weight chosen on validation examples; the combination improved on the probe in each of the four settings for which it was reported
Arnav et al. (2025): successful unauthorized side task	AUC: output-only .755; reasoning-plus-output .760; hybrid .797	Equal monitor-call budget; same attempts; tested under multiple attack policies
Koran et al. (2026): code-based backdoor monitoring	Normalized partial AUC up to 10% false alarms: single .72; three runs .76; best three-monitor composition .82	Best composition ranked on test data; distinct from the validation-selected operating points
Parrack et al. (2025): code-submission deception dataset	AUC: transcript monitor .899; monitor given probe scores .776; numerical combination .925	Additional evidence can impair one combination procedure while benefiting another
Dingeto (2026): concealed versus unknown answers	AUC: internal .892; behavioral .845; combined .929	Detector classification on the tested password-locked model, not unrestricted knowledge identification

Selected results, not a meta-analysis or ranking; targets, models and metrics differ. Liu et al.: GPT-J on SciQ, Table 1 and Section 5. Arnav et al.: Fig. 8 and Section 2.2. Koran et al.: Table 1. Parrack et al.: Table 3. Dingeto: Table 1

Human concealed-information research shows how much the comparison cases matter. Meijer et al. (2007) found that adding a forced-choice symptom validity test (Section 3.3) did not improve a physiological test against genuinely uninformed participants. On the same data, constructing the physiological comparison from knowledgeable participants' critical and control responses yielded higher apparent discrimination than using real uninformed participants (.95 against .86). The implication for AI is direct: a detector that distinguishes constructed positive and negative cases may be exploiting their construction. A combination should be evaluated against cases that instantiate the intended deployment contrast, with labels checked independently wherever possible.

3.6 When a measurement becomes the target

The reviewed studies illustrate three routes to influencing a measurement: training can reward the system for keeping the measurement quiet, the system can learn how the measurement works, or it can simply be told to defeat it. Baker et al.'s reasoning-only experiment (Section 2.9) illustrates the first: reward-based pressure reduced both cheating and the detectability of the cheating that remained, and only the independent check of *X* separated the two effects.

Haskins et al. (2026) examined the second route: learning about a monitor without receiving its verdict as a reward. In a fictional customer-support setting, models first learned from synthetic company

documents to withhold known product defects from customers. Two copies of the same checkpoint then received equal numbers of further documents. One continued to learn about the concealment policy. The other learned that a monitor would inspect its reasoning for signs of withholding, including descriptions of what the monitor looked for and hints about avoiding it, but no worked examples of evasive reasoning. At test, the models received engineering notes documenting a defect and answered a customer's question. *R* was the Yes or No answer, *X* whether it falsely denied the documented defect, and *A* the monitor's assessment of the question, reasoning and answer. For gpt-oss-120b, undetected false denials rose from 1.7% to 25.7% of all test scenarios, even though false denials became less frequent overall (Haskins et al. 2026). A reduced tendency to misbehave therefore coexisted with a greater rate of undetected misconduct. The documents supplied information about monitoring and the checkpoints were selected using undetected behavior on validation scenarios, so the result is not evasion without any guidance. It establishes a route distinct from rewarding each response for evasion: knowledge about an evaluation procedure can change how revealing the reasoning becomes. Document exposure, validation-based model selection and per-response monitor rewards are different adaptation opportunities and should be reported separately.

The third route is instruction. Zolkowski et al. (2025) gave reasoning models a hidden side task alongside simulated multi-step work and instructed them to keep it out of their reasoning, with pressure that rose from a bare instruction to examples of successful evasion, scoring evasion only on trials in which both tasks succeeded. Without pressure, monitors reading only the reasoning caught 96–98% of side tasks, against 8–20% for monitors that saw only actions. Under the strongest pressure, the reasoning monitor caught fewer than 10% for one model, although several other models rarely evaded, and concealment itself reduced success on the side task. High initial detection did not guarantee robustness for that model (see also Chen et al. 2026, on how much harder models find it to control their reasoning than their answers; Emmons et al. 2025, on a monitor that training evaded and a stronger one that it did not).

3.7 What decides whether a measurement resists?

Internal measurements are not immune to the same problem. Gupta and Jenner (2025) rewarded models for producing harmful completions while evading activation monitors. Probes that averaged their scores over a response's tokens, or took the median, were evaded; probes that took the maximum score, or learned which tokens to weight, resisted evasion that preserved the harmful output under the tested conditions. A low monitor score alone was not enough to count as evasion, since some training runs evaded by ceasing to perform the harmful task.

Two further comparisons sharpened the interpretation. In a trigger-conditioned setting, newly fitted mean-aggregated probes also missed triggered examples when their training data excluded such examples, which extends the failure beyond one fixed detector but not beyond every training distribution; a later appendix reports improvement when a small proportion of triggered examples entered probe training. Separately, the authors supplied the same evasive completions to the original and the modified models. Changed output content accounted for much of the detection loss, and changed internal

processing of identical content contributed further (Gupta and Jenner 2025; Online Resource 1, A24 and A25). These comparisons separate three meanings of a monitor failure: failure of a particular fitted instrument, failure to generalize from a specified training distribution, and altered internal processing of the same material. None of them is the disappearance of the monitored property.

Related studies broaden the conditions under which evasion transfers. Models fine-tuned against probes for harmless concepts could evade previously unseen deception monitors (McGuinness et al. 2025). Training against Dingeto's (2026) recognition reading reduced it toward an unknown-item baseline while the triggered capability remained. Input-space attacks and preference training introduce still other adaptation routes and trade-offs (Bailey et al. 2026; Cundy and Gleave 2025). The appropriate conclusion is about the tested procedure, access and data distribution.

Human countermeasure experiments make the same distinction visible within a single measurement suite. In Peth et al. (2016), some participants carried out a mock theft and others did not. During a later test, everyone denied recognizing the possible crime details, so the same categorical response was given by participants with different knowledge. Auxiliary scores compared physiological reactions to the actual details with reactions to plausible control items. Some knowledgeable participants were taught countermeasures, deliberate attempts to defeat the test, including small toe movements during selected control items. With that physical countermeasure, skin-conductance AUC fell from .86 to .50, whereas comparable reductions were not established for respiration or heart rate, and the physiological combination remained informative despite impairment. The countermeasure did not require forgetting the theft; it altered responses to the comparison items, so discrimination was impaired by creating competing evidence rather than by suppressing the response to the target. The test concerns knowledge of the details, not guilt, because an uninvolved person could also know them (Ben-Shakhar and Elaad 2003; Lykken 1959).

The motive to conceal can also be manipulated with the knowledge held constant. When participants who knew the details of a fictional crime imagined being suspects who wanted their knowledge undetected, or witnesses who wanted it detected, skin conductance distinguished the details in both groups, whereas heart-rate slowing and respiratory suppression appeared mainly in the suspects (klein Selle et al. 2016). The pattern suggests different contributions from recognition and efforts to conceal it, not pure readouts of either process. Two lessons carry over to AI. A reading that was not controlled in one experiment should not be called uncontrollable, since even association-based measures of attitudes, once thought hard to control, can be faked with instruction and practice (Fiedler and Bluemke 2005; see also Moors and De Houwer 2006, on specifying control by its goal). And manipulating one score does not show control over every signal available to an evaluator.

The routes of pressure in Sections 3.6 and 3.7 differ in what they change, not only in strength, so no single ordering from weak to strong applies. A test of robustness is interpretable when it states who adapts, what that actor knows about the measurement, what it can change (the prompt, the input or the weights), what feedback it receives, and how much computation or training it may use.

3.8 What a safety argument requires of these tests

Detector performance is not the safety of a complete procedure. A control protocol combines monitoring with other safeguards, such as having a trusted but weaker model review the work of a more capable model that is not trusted. Greenblatt et al. (2024a) showed that a second copy of the untrusted model, used as a monitor, could defeat naive monitoring by calling deliberately flawed code safe. Testing the monitor on known flawed code, and limiting the signals the copies could use to coordinate, raised estimated safety from 0% to 79%, where safety is the probability of submitting no successful backdoor across a simulated sequence of 5,000 problems. A report on a monitor should therefore give both detection among successful unwanted behaviors and the overall rate of successful undetected behavior.

A safety case, a structured argument explaining why a system is expected to avoid a specified harm, states which such tests it rests on. In a hypothetical case sketched by Balesni et al. (2024), an untrusted monitor that could be defeated by collusion is supplemented with probes trained on different data, chosen to reduce correlated failures, and the two scores are combined. The red team may adapt its attack policy but may not train against the probes, and the probes' resistance to evasion is argued from a capability evaluation rather than tested. The example shows how the specifications above enter a safety argument, and which of its assumptions remain untested.

4 What the comparisons establish together

The reviewed studies do not identify one universally privileged measurement. They show several ways in which a measurement can be useful, and several different events that can make it fail. Table 6 connects the principal research questions to the comparisons that address them, the conclusion each supports and the uncertainty that remains. An evidence map in Online Resource 1 (Table S1) records which of the five comparisons each reviewed study made, by research area; its empty cells indicate no entry under that selective coding, not demonstrated absences from the literature. Section 5 develops the specific unresolved comparisons rather than treating empty cells as a research agenda. Three conclusions organize the synthesis.

Table 6
Cross-study conclusions and the comparisons that support them

Research question	Illustrative comparisons	Conclusion from the synthesis	Next uncertainty
Does the response exhaust readable information?	Code review and answer-error detection; persona and false-label manipulations (Wiryadi 2026; Orgad et al. 2025; Mallen et al. 2024; Halawi et al. 2024)	A response-based reading can miss distinctions recovered elsewhere; agreement can break under a planned change	Incremental value beyond strong comparators, and transfer to other conditions
Does the reasoning explain or predict behavior?	Hint manipulations, reasoning edits and acquisition-history comparisons (Turpin et al. 2023; Lanham et al. 2023; Emmons et al. 2025; Greenblatt et al. 2024b)	Reporting, causal involvement and practical monitorability are different properties	Which traces add predictive value for a specified target
Does poor performance establish absence?	Recovery after unlearning and perturbation of induced underperformance (Łucki et al. 2025; Tice et al. 2025)	Failure concerns the tested procedure, not every route to the information	Specific residual influence versus general facilitation or disruption
Does a useful combination remain useful under pressure?	Hybrid monitoring, changed monitor knowledge and targeted evasion (Arnav et al. 2025; Baker et al. 2025; Haskins et al. 2026; Gupta and Jenner 2025)	Complementarity, aggregation and resistance depend on the assessed attempts and the adaptation route	Transfer of evasion across components, and attacks on the full combination
What does high detector performance identify?	Constructed response policies, random-direction controls and belief-verified cases (Hubinger et al. 2024; MacDiarmid et al. 2024; Cooney et al. 2026)	Controlled labels can reveal a detector's strengths without validating a general mental-state interpretation	Transfer beyond the construction, and independent validation of the target

An illustrative synthesis of the cited comparisons, not a systematic inventory. Research areas overlap, and a paper may contribute different experiments to more than one row. Online Resource 1 identifies the principal experimental units and source locations

4.1 Agreement is conditional, not a guarantee of interchangeability

Agreement under a reference condition can leave several explanations unresolved. In the Alice–Bob design, agreement is compatible with a probe that tracks actual truth and with one that predicts the assigned answer; redirecting the response separates these interpretations. In monitoring, useful combinations likewise depend on where each assessment succeeds and fails, not only on the difference between their inputs (Mallen et al. 2024; Liu et al. 2023; Arnav et al. 2025). The lesson is stronger than a

recommendation to collect more measurements: investigate the conditions under which apparent redundancy or complementarity changes.

Baseline performance, incremental value and resistance to adaptation should therefore be recorded separately. A measurement may add little while *R* is informative, yet become valuable when *R* is redirected. A combination that improves ordinary prediction can remain vulnerable to a shared attack. Even statistical diversity requires interpretation, since disagreement among weak monitors need not help (Koran et al. 2026). These findings motivate retaining joint case-level scores and measuring changes in their relationship, rather than comparing headline accuracies from different samples.

4.2 Failure of a measurement is not disappearance of its target

A poor answer after unlearning, a reasoning trace that omits a hint, and a monitor that stops flagging cheating are superficially similar absences, but they concern different targets and procedures. Łucki et al. (2025) tested recovery after an intervention; the hint studies compared induced behavioral influence with reporting; Baker et al. (2025) checked misconduct independently while training against a monitor. Together they show why a negative conclusion must specify what was tested and what opportunities the test had to reveal the property.

Negative evidence is not worthless. With adequate precision, confidence bounds or equivalence tests can exclude effects above a prespecified size under specified conditions. What they cannot establish without further assumptions is exhaustive absence across every elicitation or reading procedure, and several weak measurements failing together do not add up to it. The human literature distinguishes this statistical question from the substantive assumption that a measure captures all relevant information (Merikle and Reingold 1998; Schmidt and Vorberg 2006).

Ground truth is most difficult where the intended claim extends beyond observable performance. Arithmetic answers and particular code vulnerabilities can often be checked independently; beliefs, motives and deception are harder to establish. Constructed model organisms provide controlled histories and known response contrasts, but their labels may incorporate cues that a detector can exploit, as the random-direction control in sleeper-agent probing and the belief-verified deception cases showed (MacDiarmid et al. 2024; Cooney et al. 2026). A construction can support a stringent test of a detector without showing that it captures every naturally arising case of the intended construct.

4.3 Usefulness belongs to a procedure and a setting

An activation measurement depends on the prompt, layer, token selection, fitted probe, aggregation rule and operating threshold. A reasoning assessment depends additionally on what the trace was generated to accomplish and on what the monitor sees. Changing any of these changes the instrument. Orgad et al. (2025) show the importance of token selection; Parrack et al. (2025) the importance of the combination rule; Gupta and Jenner (2025) separate changes in the generated material from changes in

the processing of the same material. The location of a measurement does not determine its evidential quality.

Even the unit used to set a threshold can matter. Cooney et al. (2026, Appendix E) report that a released deception-probe implementation set its 1% false-alarm threshold over pooled token scores rather than response-level scores. Correcting the procedure reduced reported detection on one dataset from 96% to 73.2%, a change in detection at a specified operating point rather than in AUC. A reproducible measurement therefore includes its aggregation rule and how its threshold was set, not only the probe's weights.

The setting includes the model's learning history and the evaluator's access. Removing a scratchpad affected a compliance gap differently depending on whether information about training was supplied in the prompt or learned from documents, and learning about a monitor made reasoning less revealing without a per-response reinforcement-learning reward from the monitor (Greenblatt et al. 2024b; Haskins et al. 2026). A valid test on one constructed model therefore motivates, rather than replaces, tests across acquisition histories and adaptation routes.

These qualifications guide useful research rather than imposing an ideal of perfect matching. For practical detection, an extra prompt or a stronger evaluator may be worth its cost. For an inference about information available during a particular computation, the additional processing must instead be treated as a difference between conditions. For a safety protocol, the relevant outcome is what harmful behavior passes the whole procedure. Stating the intended claim first makes it possible to select the appropriate comparison and acknowledge its limits without dismissing its contribution. The most promising next studies connect these conclusions: they ask whether information missed by several immediate measurements still affects subsequent learning, whether a reasoning trace adds beyond a strong response-based assessment, and whether a useful combination survives adaptation against the combination itself.

5 Using the framework to design informative studies

The framework contributes most when it changes the next experiment. Start with a result and its operational target, identify the alternatives it leaves unresolved, and choose a comparison whose outcomes would distinguish them. Preserve the original study's useful controls and state the interpretation of a negative result in advance. Table 7 gives four examples. The first is a direct application of incremental-value testing to code review; the others motivate the worked extensions below. These are proposed comparisons, not reported results, and neither relearning nor adaptive attack is claimed to be new to AI.

Table 7
From existing findings to informative follow-up comparisons

Existing AI result	Unresolved question	Proposed comparison and methodological resource
A code-security probe outperformed output-based rankings (Wiryadi 2026)	Does the probe add beyond a strong predictor fitted to the output?	Matched output-only and combined predictors on new pairs; incremental-validity logic (Sechrest 1963; Gamer et al. 2008)
Reasoning can omit a behaviorally effective hint (Chen et al. 2025)	Does the trace add beyond what an observer can infer from the prompt and output?	Repeated hinted and unhinted runs; held-out comparison with and without the trace; observer control (Nisbett and Wilson 1977)
Interventions recover answers after apparent unlearning (Łucki et al. 2025)	When immediate recovery fails, does specific prior learning still matter?	Original versus rearranged associations with equally familiar elements and matched learning histories; controlled savings (Nelson 1978)
Hybrid monitoring outperformed individual assessments (Arnav et al. 2025)	Does the benefit survive adaptation against components or the combination?	Retain all monitors while changing the attack target; within-suite countermeasure comparison (Peth et al. 2016)

These extensions use established methods in new specified comparisons. Their value depends on separating the stated alternatives. Human precedents supply controls and inferential distinctions, not assumed mechanisms

Table 7 draws on human methodological resources to specify further AI comparisons. Other approaches, including process-dissociation designs (Jacoby 1991; Reingold and Toth 1996) and state-trace analysis (Dunn and Kalish 2018; Newell and Dunn 2008), are not developed here. Applying them would require explicit assumptions and a worked AI adaptation, rather than directly importing psychological process labels.

5.1 Controlled relearning after apparent unlearning

The unlearning evidence poses a question beyond immediate recovery: when answers and available probes reveal little, does prior learning still influence what the model learns next? This extends the intervention setting of Łucki et al. (2025) rather than treating their recovery result as proof that all unlearning fails. A controlled relearning comparison could test methods or checkpoints for which immediate recovery procedures do not succeed.

Nelson's (1978) human experiments supply the key control. After learning number–word associations, participants completed recall and recognition tests four weeks later. Relearning then compared original associations with rearranged associations made from the same previously encountered numbers and words. The comparison controlled familiarity with the elements while changing their relation. In Experiment 1, original associations were relearned more successfully even among items previously

missed by both tests. That finding showed an influence of prior learning that those tests had not detected, not necessarily a complete, unchanged memory trace.

An AI extension could use harmless, newly generated associations: for example, fictitious item names linked arbitrarily to symbols. A trained branch learns stable associations; a comparison branch encounters the same vocabulary and comparable task practice without learning those specific relations. After applying an unlearning procedure, evaluate answer probabilities and selected readings without feedback. Then copy each resulting checkpoint into matched relearning conditions. One condition receives the original associations; another receives rearranged associations made from the same elements. Mapping assignments and item sets should be counterbalanced (Table 8). An unlearned-versus-originally-trained comparison is not enough on its own, because ordinary practice or familiarity could accelerate both kinds of relearning.

Table 8
A controlled savings comparison using harmless synthetic associations

Prior learning history	Relearning comparison	What the contrast addresses
Stable associations learned, then targeted by unlearning	Original relations versus rearranged relations using the same elements	Whether specific prior relations still influence learning
Same elements and comparable practice, without those stable relations	The same counterbalanced original and rearranged assignments	Familiarity, item difficulty and general learning facilitation
Stable associations retained without targeted unlearning	The same relearning comparison	A positive control that the procedure can detect retained influence

Illustrative design, not empirical results. Initial test exposure, relearning budget and evaluation items are matched. Additional sham or nonspecific-intervention controls are needed to distinguish relation-specific effects from general disruption by unlearning. Inspired by Nelson (1978) and the AI unlearning evaluations of Łucki et al. (2025)

Relearning can be measured as progress across a fixed number of examples or as the number needed to reach a prespecified criterion. The primary analysis should include the full counterbalanced item set. A secondary analysis of poorly reported items needs independent screening observations so that noise in the selection test does not define the result. Matched exposure to the immediate tests, along with controls for nonspecific effects of unlearning, guards against learning during measurement or general impairment of new learning.

The diagnostic result is an advantage for original over rearranged relations that is not explained by the corresponding contrast in the comparison history. Faster learning of both relation types suggests general facilitation instead. An original-relation advantage may reflect facilitation of the original relation, interference with the rearranged relation, or both; each indicates a persisting influence of prior learning

without identifying its storage mechanism. A sufficiently precise null result would bound the size of that advantage under the tested relearning procedure, not establish universal erasure. These measurements deliberately concern a changed system after additional learning, not an unchanged concurrent measurement.

5.2 Adapting against monitors and their combination

Arnav et al.'s (2025) setting already supplies independently checked task outcomes, two monitors, a combined score and multiple attack policies. Peth et al.'s (2016) countermeasure design suggests retaining the full measurement suite while changing which component the opponent targets. Starting from the same attacking model and task distribution, one condition would receive feedback from the output-only monitor, another from the reasoning-and-output monitor, and a third from the combination. A control would receive no monitor feedback while retaining matched task practice. The adaptation method, the number of attempts and the computational budget would be specified and matched.

The monitors, aggregation rule and thresholds would be fixed before final testing on new problems. Thresholds would be set on an independent benign sample, with achieved false-alarm rates reported on the test sample rather than assumed to equal the target rate. A resource-matched output-only comparator, such as additional calls or a stronger model, would test whether the gain comes from complementary assessment rather than from greater evaluation resources. This extends existing adaptive testing; the original study already evaluated several attack policies.

Every resulting attempt would be assessed by every monitor, including monitors not targeted during adaptation, and independent execution checks would establish main-task and side-task success. This produces a transfer comparison: does evading one assessment leave another informative, impair both, or make the other more informative because evasion leaves a different trace? Direct adaptation against the combination tests whether its incremental benefit survives when that benefit itself is targeted. Transfer between probes has been studied; the proposed contribution is the joint comparison of component and combination targets within the same task system (Gupta and Jenner 2025; McGuinness et al. 2025).

The report would give legitimate-task success, side-task success per attempt, detection among successful side tasks, false alarms on benign attempts, and successful undetected side tasks per attempt, with detection, false alarms and undetected side tasks reported for each assessment and the combination. Conditional detection can decrease because the successful attacks have changed, not only because identical behavior has become less detectable; retaining the attempts and characterizing their outcomes allows that distinction to be investigated rather than hidden by filtering.

If a component-targeted attack leaves the other assessment useful and the combination retains an advantage, complementarity survives that specified pressure. If evasion transfers broadly and the combination adds little, different inputs have not prevented a shared failure. If defeating both assessments requires abandoning the side task or severely damaging legitimate performance, the result

identifies a trade-off. If the resource-matched output-only comparator eliminates the combination's advantage, no incremental advantage over that resource-matched alternative is established. None of these outcomes establishes permanent invulnerability; each makes the practical value of the measurement suite more precisely testable.

5.3 What a reasoning report adds beyond an observer

The hint experiments permit a further extension informed by Nisbett and Wilson's (1977) observer comparison. For each question, define X as the change in the frequency of selecting the hinted option between repeated hinted and unhinted runs. Estimate this contrast from one set of runs, and use separate runs to produce the output probabilities and traces supplied to the predictors. Otherwise, the predictor could receive the very responses used to define its target. This assesses question-level susceptibility, not whether a hint caused one individual response. Split by question so that fitting and final evaluation do not share the same problem. Compare a predictor receiving the question, the hint and the output probabilities with one receiving those same inputs plus the reasoning trace. Match fitting opportunities and assess the incremental gain on held-out questions, with uncertainty evaluated at the question level.

A gain would show that the trace adds information about measured hint susceptibility beyond the tested observer-style comparator. Shuffled or mismatched traces would test whether that gain depends on the trace from the corresponding question and run rather than on generic stylistic features. A null gain against a strong comparator would narrow the information claim without showing that reasoning never contributes causally to the answer. Combining this test with the simple-versus-demanding hint manipulation could reveal when a report becomes incrementally informative, connecting the observer comparison to an existing AI experiment (Emmons et al. 2025).

6 Conclusion

Dissociations are useful for AI safety when they change what an evaluation justifies concluding. A reading of internal activity can recover distinctions missed by a response-based assessment; two assessments can work better together; a change of conditions can separate measurements that agreed; and a controlled intervention can test whether measured information influences behavior. Targeted adaptation adds the question of whether useful evidence remains available when avoiding detection becomes an objective. These are complementary comparisons, not a single ranking of evidence or a guarantee that multiple measurements will suffice.

Across the reviewed studies, three conclusions stand out. Agreement is conditional: redundancy and complementarity can change when the response rule, the learning history or the attack changes. Negative evidence is procedure-dependent: unsuccessful recovery does not by itself establish disappearance of the target. And the measurement is the whole procedure, including elicitation, scoring, fitting, aggregation and operating conditions. These conclusions connect fields that otherwise use different names for closely related experimental problems.

Human research contributes ways to make those comparisons informative, not psychological labels to attach to AI systems. AI research adds the ability to copy checkpoints, retain detailed records and manipulate specified computations or learning histories. The strongest use of that combination is prospective: an existing result identifies an uncertainty, and a controlled comparison is chosen whose outcomes would resolve it. A cumulative science of safety-relevant measurement should therefore ask not only whether measurements disagree, but what the disagreement rules out, where its interpretation stops, and which experiment would extend it.

Declarations

Competing interests

The authors declare no competing interests.

Use of generative AI

ChatGPT (OpenAI) was used to assist with literature organization, drafting, language revision and checks of consistency between the manuscript and cited materials. The authors are responsible for evaluating the source evidence and approving all text, interpretations and citations in the submitted version.

Supplementary information

Online Resource 1: Experiment-level guide to the principal AI comparisons and human methodological resources discussed in this review, with an evidence map (Table S1) and technical notes on selected comparisons. The guide distinguishes original findings, inferential boundaries and proposed extensions.

Funding

No funding was received for this work.

Author Contribution

E.R. conceived and led the review, conducted the primary literature synthesis and methodological analysis, and had the primary role in drafting and revising the manuscript and supplementary material. O.R. contributed extensively to the conceptual development of the review through ongoing discussion, helped refine the methodological framework and interpretation of the literature, and critically reviewed and revised the manuscript.

Data availability

No new empirical data were collected for this review. Numerical results are drawn from the cited studies. Online Resource 1 identifies the principal experimental comparisons, source locations and interpretive boundaries used in the synthesis.

References

1. Abdelnabi S, Salem A (2025) The Hawthorne effect in reasoning models: Evaluating and steering test awareness. Adv Neural Inf Process Syst 38:156975–157002. https://doi.org/10.52202/085713-4731
2. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv preprint. https://doi.org/10.48550/arXiv.1606.06565
3. Arnav B, Bernabeu-Pérez P, Helm-Burger N, Kostolansky T, Whittingham H, Phuong M (2025) CoT Red-Handed: Stress testing chain-of-thought monitoring. Adv Neural Inf Process Syst 38. https://doi.org/10.48550/arXiv.2505.23575
4. Azaria A, Mitchell T (2023) The internal state of an LLM knows when it's lying. In: Findings of the Association for Computational Linguistics: EMNLP 2023 pp 967–976. https://doi.org/10.18653/v1/2023.findings-emnlp.68
5. Bai X, Wang A, Sucholutsky I, Griffiths TL (2025) Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences 122(8):e2416228122. https://doi.org/10.1073/pnas.2416228122
6. Bailey L, Serrano A, Sheshadri A, Seleznyov M, Taylor J, Jenner E, Hilton J, Casper S, Guestrin C, Emmons S (2026) Obfuscated activations bypass LLM latent-space defenses. In: International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2026/hash/ed687a5f52b651b19e7c18f702907b8b-Abstract-Conference.html
7. Baker B, Huizinga J, Gao L, Dou Z, Guan MY, Madry A, Zaremba W, Pachocki J, Farhi D (2025) Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv preprint. https://doi.org/10.48550/arXiv.2503.11926
8. Balesni M, Hobbhahn M, Lindner D, Meinke A, Korbak T, Clymer J et al (2024) Towards evaluations-based safety cases for AI scheming. arXiv preprint. https://doi.org/10.48550/arXiv.2411.03336
9. Bean AM, Kearns RO, Romanou A, Hafner FS, Mayne H, Batzner J et al (2025) Measuring what matters: Construct validity in large language model benchmarks. Adv Neural Inf Process Syst 38:19868–19949. https://doi.org/10.52202/085713-0590
10. Ben-Shakhar G, Elaad E (2003) The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. J Appl Psychol 88(1):131–151. https://doi.org/10.1037/0021-9010.88.1.131
11. Bengio Y, Clare S, Prunkl C, Murray M, Andriushchenko M, Bucknall B et al (2026) International AI safety report 2026 (DSIT 2026/001). Department for Science, Innovation and Technology. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

12. Bürger L, Hamprecht FA, Nadler B (2024) Truth is universal: Robust detection of lies in LLMs. Adv Neural Inf Process Syst 37:138393–138431. https://doi.org/10.52202/079017-4392
13. Burns C, Ye H, Klein D, Steinhardt J (2023) Discovering latent knowledge in language models without supervision. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2212.03827
14. Chen Y, Benton J, Radhakrishnan A, Uesato J, Denison C, Schulman J et al (2025) Reasoning models don't always say what they think. arXiv preprint. https://doi.org/10.48550/arXiv.2505.05410
15. Chen Y-H, McCarthy R, Lee BW, He H, Carroll M, Korbak T (2026) Reasoning models struggle to control their chains of thought. In: Proceedings of the 43rd International Conference on Machine Learning Proceedings of Machine Learning Research vol 306, pp 152359–152387. https://proceedings.mlr.press/v306/yueh-han26a.html
16. Clymer J, Juang C, Field S (2024) Poser: Unmasking alignment faking LLMs by manipulating their internals. arXiv preprint. https://doi.org/10.48550/arXiv.2405.05466
17. Cooney A, Africa D, Irving G (2026) Did you lie? Evaluating lie detectors across model scale and belief-verified model organisms. arXiv preprint. https://doi.org/10.48550/arXiv.2606.12618
18. Cundy C, Gleave A (2025) Preference learning with lie detectors can induce honesty or evasion. Adv Neural Inf Process Syst 38:143632–143690. https://doi.org/10.52202/085713-4313
19. Cywiński B, Ryd E, Wang R, Rajamanoharan S, Nanda N, Conmy A, Marks S (2025) Eliciting secret knowledge from language models. arXiv preprint. https://doi.org/10.48550/arXiv.2510.01070
20. Desai M, Truong ST, Wallach H, Chouldechova A, Cooper AF, Garcia-Gathright J et al (2026) What AI benchmarks actually measure: Adapting convergent and discriminant validity to interrogate fifty-six AI benchmarks. arXiv preprint. https://doi.org/10.48550/arXiv.2609.08812
21. Dingeto H (2026) A lie detector test for language models: Reading knowledge a model won't reveal. arXiv preprint, version 2. https://doi.org/10.48550/arXiv.2609.21996
22. Dunn JC, Kalish ML (2018) State-trace analysis. Springer, Cham. https://doi.org/10.1007/978-3-319-73129-2
23. Dunn JC, Kirsner K (1988) Discovering functionally independent mental processes: The principle of reversed association. Psychol Rev 95(1):91–101. https://doi.org/10.1037/0033-295X.95.1.91
24. Dunn JC, Kirsner K (2003) What can we infer from double dissociations? Cortex 39(1):1–7. https://doi.org/10.1016/S0010-9452(08)70070-4
25. Elazar Y, Ravfogel S, Jacovi A, Goldberg Y (2021) Amnesic probing: Behavioral explanation with amnesic counterfactuals. Trans Association Comput Linguistics 9:160–175. https://doi.org/10.1162/tacl_a_00359
26. Emmons S, Jenner E, Elson DK, Saurous RA, Rajamanoharan S, Chen H, Shafkat I, Shah R (2025) When chain of thought is necessary, language models struggle to evade monitors. arXiv preprint. https://doi.org/10.48550/arXiv.2507.05246

27. Ericsson KA, Simon HA (1993) Protocol analysis: Verbal reports as data, revised edition. MIT Press, Cambridge. https://doi.org/10.7551/mitpress/5657.001.0001
28. Fiedler K, Bluemke M (2005) Faking the IAT: Aided and unaided response control on the Implicit Association Tests. Basic Appl Soc Psychol 27(4):307–316. https://doi.org/10.1207/s15324834basp2704_3
29. Gamer M, Verschuere B, Crombez G, Vossel G (2008) Combining physiological measures in the detection of concealed information. Physiol Behav 95(3):333–340. https://doi.org/10.1016/j.physbeh.2008.06.011
30. Gekhman Z, Ben David E, Orgad H, Ofek E, Belinkov Y, Szpektor I, Herzig J, Reichart R (2025) Inside-out: Hidden factual knowledge in LLMs. In: Conference on Language Modeling. https://doi.org/10.48550/arXiv.2503.15299
31. Goldowsky-Dill N, Chughtai B, Heimersheim S, Hobbhahn M (2025) Detecting strategic deception with linear probes. In: Proceedings of the 42nd International Conference on Machine Learning Proceedings of Machine Learning Research vol 267, pp 19755–19786. https://proceedings.mlr.press/v267/goldowsky-dill25a.html
32. Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley
33. Greenblatt R, Denison C, Wright B, Roger F, MacDiarmid M, Marks S et al (2024b) Alignment faking in large language models. arXiv preprint. https://doi.org/10.48550/arXiv.2412.14093
34. Greenblatt R, Shlegeris B, Sachan K, Roger F (2024a) AI control: Improving safety despite intentional subversion. In: Proceedings of the 41st International Conference on Machine Learning Proceedings of Machine Learning Research vol 235, pp 16295–16336. https://proceedings.mlr.press/v235/greenblatt24a.html
35. Greenwald AG, McGhee DE, Schwartz JLK (1998) Measuring individual differences in implicit cognition: The Implicit Association Test. J Personal Soc Psychol 74(6):1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464
36. Grey M, Segerie C-R (2025) Safety by measurement: A systematic literature review of AI safety evaluation methods. arXiv preprint. https://doi.org/10.48550/arXiv.2505.05541
37. Gupta R, Jenner E (2025) RL-obfuscation: Can language models learn to evade latent-space monitors? arXiv preprint, version 4 (revised 26 February 2026. https://doi.org/10.48550/arXiv.2506.14261
38. Halawi D, Denain J-S, Steinhardt J (2024) Overthinking the truth: Understanding how language models process false demonstrations. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2307.09476
39. Haskins R, Chughtai B, Engels J (2026) Training on documents about monitoring leads to CoT obfuscation. arXiv preprint. https://doi.org/10.48550/arXiv.2605.15257
40. Heidari F, Memarian A, Rabusseau G (2026) Evaluation awareness in language models: Representation, verbalization, and control. arXiv preprint. https://doi.org/10.48550/arXiv.2608.21766

41. Hendrycks D, Carlini N, Schulman J, Steinhardt J (2021) Unsolved problems in ML safety. arXiv preprint. https://doi.org/10.48550/arXiv.2109.13916
42. Hewitt J, Liang P (2019) Designing and interpreting probes with control tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing pp 2733–2743. https://doi.org/10.18653/v1/D19-1275
43. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint. https://doi.org/10.48550/arXiv.1503.02531
44. Hofmann V, Kalluri PR, Jurafsky D, King S (2024) AI generates covertly racist decisions about people based on their dialect. Nature 633(8028):147–154. https://doi.org/10.1038/s41586-024-07856-5
45. Hong Y, Yu L, Yang H, Ravfogel S, Geva M (2025) Intrinsic test of unlearning using parametric knowledge traces. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing pp 19513–19535. https://doi.org/10.18653/v1/2025.emnlp-main.985
46. Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M et al (2024) Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv preprint. https://doi.org/10.48550/arXiv.2401.05566
47. Jacoby LL (1991) A process dissociation framework: Separating automatic from intentional uses of memory. J Mem Lang 30(5):513–541. https://doi.org/10.1016/0749-596X(91)90025-F
48. klein Selle N, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G (2016) Orienting versus inhibition in the Concealed Information Test: Different cognitive processes drive different physiological measures. Psychophysiology 53(4):579–590. https://doi.org/10.1111/psyp.12583
49. Koran E, Yun Y, Tetef S, Arnav B, Bernabeu-Pérez P (2026) Ensemble monitoring for AI control: Diverse signals outweigh more compute. arXiv preprint. https://doi.org/10.48550/arXiv.2605.15377
50. Kretschmar K, Laurito W, Maiya S, Marks S (2026) Liars' Bench: Evaluating lie detectors for language models. arXiv preprint. https://doi.org/10.48550/arXiv.2511.16035
51. Langosco LLD, Koch J, Sharkey LD, Pfau J, Krueger D (2022) Goal misgeneralization in deep reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning Proceedings of Machine Learning Research vol 162, pp 12004–12019. https://proceedings.mlr.press/v162/langosco22a.html
52. Lanham T, Chen A, Radhakrishnan A, Steiner B, Denison C, Hernandez D et al (2023) Measuring faithfulness in chain-of-thought reasoning. arXiv preprint. https://doi.org/10.48550/arXiv.2307.13702
53. Li K, Patel O, Viégas F, Pfister H, Wattenberg M (2023) Inference-time intervention: Eliciting truthful answers from a language model. Adv Neural Inf Process Syst 36:41451–41530. https://doi.org/10.52202/075280-1797
54. Liu K, Casper S, Hadfield-Menell D, Andreas J (2023) Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In: Proceedings of the 2023

Conference on Empirical Methods in Natural Language Processing pp 4791–4797. https://doi.org/10.18653/v1/2023.emnlp-main.291

55. Łucki J, Wei B, Huang Y, Henderson P, Tramèr F, Rando J (2025) An adversarial perspective on machine unlearning for AI safety. Trans Mach Learn Res. https://doi.org/10.48550/arXiv.2409.18025
56. Lykken DT (1959) The GSR in the detection of guilt. J Appl Psychol 43(6):385–388. https://doi.org/10.1037/h0046060
57. MacDiarmid M, Maxwell T, Schiefer N, Mu J, Kaplan J, Duvenaud D et al (2024) Simple probes can catch sleeper agents. Anthropic. https://www.anthropic.com/research/probes-catch-sleeper-agents Accessed 16 October 2026
58. Mallen A, Brumley M, Kharchenko J, Belrose N (2024) Eliciting latent knowledge from quirky language models. In: First Conference on Language Modeling. https://doi.org/10.48550/arXiv.2312.01037
59. Marks S, Treutlein J, Bricken T, Lindsey J, Marcus J, Mishra-Sharma S et al (2025) Auditing language models for hidden objectives. arXiv preprint. https://doi.org/10.48550/arXiv.2503.10965
60. McGuinness M, Serrano A, Bailey L, Emmons S (2025) Neural chameleons: Language models can learn to hide their thoughts from unseen activation monitors. arXiv preprint. https://doi.org/10.48550/arXiv.2512.11949
61. Meijer EH, Smulders FTY, Johnston JE, Merckelbach HLGJ (2007) Combining skin conductance and forced choice in the detection of concealed information. Psychophysiology 44(5):814–822. https://doi.org/10.1111/j.1469-8986.2007.00543.x
62. Merikle PM, Reingold EM (1998) On demonstrating unconscious perception: Comment on Draine and Greenwald (1998). Journal of Experimental Psychology: General 127(3):304–310. https://doi.org/10.1037/0096-3445.127.3.304
63. Moors A, De Houwer J (2006) Automaticity: A theoretical and conceptual analysis. Psychol Bull 132(2):297–326. https://doi.org/10.1037/0033-2909.132.2.297
64. National Research Council (2003) The polygraph and lie detection. National Academies Press. https://doi.org/10.17226/10420
65. Needham J, Edkins G, Pimpale G, Bartsch H, Hobbhahn M (2025) Large language models often know when they are being evaluated. arXiv preprint. https://doi.org/10.48550/arXiv.2505.23836
66. Nelson TO (1978) Detecting small amounts of information in memory: Savings for nonrecognized items. J Experimental Psychology: Hum Learn Memory 4(5):453–468. https://doi.org/10.1037/0278-7393.4.5.453
67. Newell BR, Dunn JC (2008) Dimensions in data: Testing psychological models using state-trace analysis. Trends Cogn Sci 12(8):285–290. https://doi.org/10.1016/j.tics.2008.04.009
68. Nisbett RE, Wilson TD (1977) Telling more than we can know: Verbal reports on mental processes. Psychol Rev 84(3):231–259. https://doi.org/10.1037/0033-295X.84.3.231

69. OpenAI (2025) Why language models hallucinate [Blog post]. https://openai.com/index/why-language-models-hallucinate/ Accessed 16 October 2026
70. Orgad H, Toker M, Gekhman Z, Reichart R, Szpektor I, Kotek H, Belinkov Y (2025) LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2410.02707
71. Orne MT (1962) On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. Am Psychol 17(11):776–783. https://doi.org/10.1037/h0043424
72. Pankratz L (1979) Symptom validity testing and symptom retraining: Procedures for the assessment and treatment of functional sensory deficits. J Consult Clin Psychol 47(2):409–410. https://doi.org/10.1037/0022-006X.47.2.409
73. Parrack A, Attubato CL, Heimersheim S (2025) Benchmarking deception probes via black-to-white performance boosts. arXiv preprint. https://doi.org/10.48550/arXiv.2507.12691
74. Peth J, Suchotzki K, Gamer M (2016) Influence of countermeasures on the validity of the Concealed Information Test. Psychophysiology 53(9):1429–1440. https://doi.org/10.1111/psyp.12690
75. Reingold EM, Merikle PM (1988) Using direct and indirect measures to study perception without awareness. Percept Psychophys 44(6):563–575. https://doi.org/10.3758/BF03207490
76. Reingold EM, Merikle PM (1990) On the inter-relatedness of theory and measurement in the study of unconscious processes. Mind Lang 5(1):9–28. https://doi.org/10.1111/j.1468-0017.1990.tb00150.x
77. Reingold EM, Toth JP (1996) Process dissociations versus task dissociations: A controversy in progress. In: Underwood G (ed) Implicit cognition. Oxford University Press, Oxford, pp 159–202
78. Schmidt T, Vorberg D (2006) Criteria for unconscious cognition: Three types of dissociation. Percept Psychophys 68(3):489–504. https://doi.org/10.3758/BF03193692
79. Sechrest L (1963) Incremental validity: A recommendation. Educ Psychol Meas 23(1):153–158. https://doi.org/10.1177/001316446302300113
80. Smith L, Chughtai B, Nanda N (2025) Difficulties with evaluating a deception detector for AIs. arXiv preprint. https://doi.org/10.48550/arXiv.2511.22662
81. Taylor J, Black S, Bowen D, Read T, Golechha S, Zelenka-Martin A et al (2025) Auditing games for sandbagging. arXiv preprint. https://doi.org/10.48550/arXiv.2512.07810
82. Tice C, Kreer PA, Helm-Burger N, Shahani PS, Ryzhenkov F, Roger F, Neo C, Haimes J, Hofstätter F, van der Weij T (2025) Noise injection reveals hidden capabilities of sandbagging language models. Adv Neural Inf Process Syst 38:117590–117622. https://doi.org/10.52202/085713-3922
83. Tulving E, Pearlstone Z (1966) Availability versus accessibility of information in memory for words. J Verbal Learn Verbal Behav 5(4):381–391. https://doi.org/10.1016/S0022-5371(66)80048-8
84. Turpin M, Michael J, Perez E, Bowman SR (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Adv Neural Inf Process Syst 36:74952–74965. https://doi.org/10.52202/075280-3275

85. van der Weij T, Hofstätter F, Jaffe O, Brown SF, Ward FR (2025) AI sandbagging: Language models can strategically underperform on evaluations. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2406.07358
86. Wallach H, Desai M, Cooper AF, Wang A, Atalla C, Barocas S et al (2025) Position: Evaluating generative AI systems is a social science measurement challenge. In: Proceedings of the 42nd International Conference on Machine Learning Proceedings of Machine Learning Research vol 267, pp 82232–82251. https://proceedings.mlr.press/v267/wallach25a.html
87. Weidinger L, Raji ID, Wallach H, Mitchell M, Wang A, Salaudeen O, Bommasani R, Ganguli D, Koyejo S, Isaac W (2025) Toward an evaluation science for generative AI systems. arXiv preprint. https://doi.org/10.48550/arXiv.2503.05336
88. Wiryadi I (2026) Activation probes surface code-security signals that the model's output misses [Poster presentation]. Second Workshop on Technical AI Governance Research, ICML 2026, Seoul, South Korea. https://doi.org/10.48550/arXiv.2608.09643
89. Zaman K, Srivastava S (2026) Is chain-of-thought really not explainability? Chain-of-thought can be faithful without hint verbalization. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pp 48008–48030. https://doi.org/10.18653/v1/2026.acl-long.2217
90. Zhao Y, Wang B, Wang Y, Zhao D, He R, Hou Y (2025) Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In: Findings of the Association for Computational Linguistics: ACL 2025 pp 1–12. https://doi.org/10.18653/v1/2025.findings-acl.1
91. Zolkowski A, Xing W, Lindner D, Tramèr F, Jenner E (2025) Can reasoning models obfuscate reasoning? Stress-testing chain-of-thought monitorability. arXiv preprint. https://doi.org/10.48550/arXiv.2510.19851

Figures

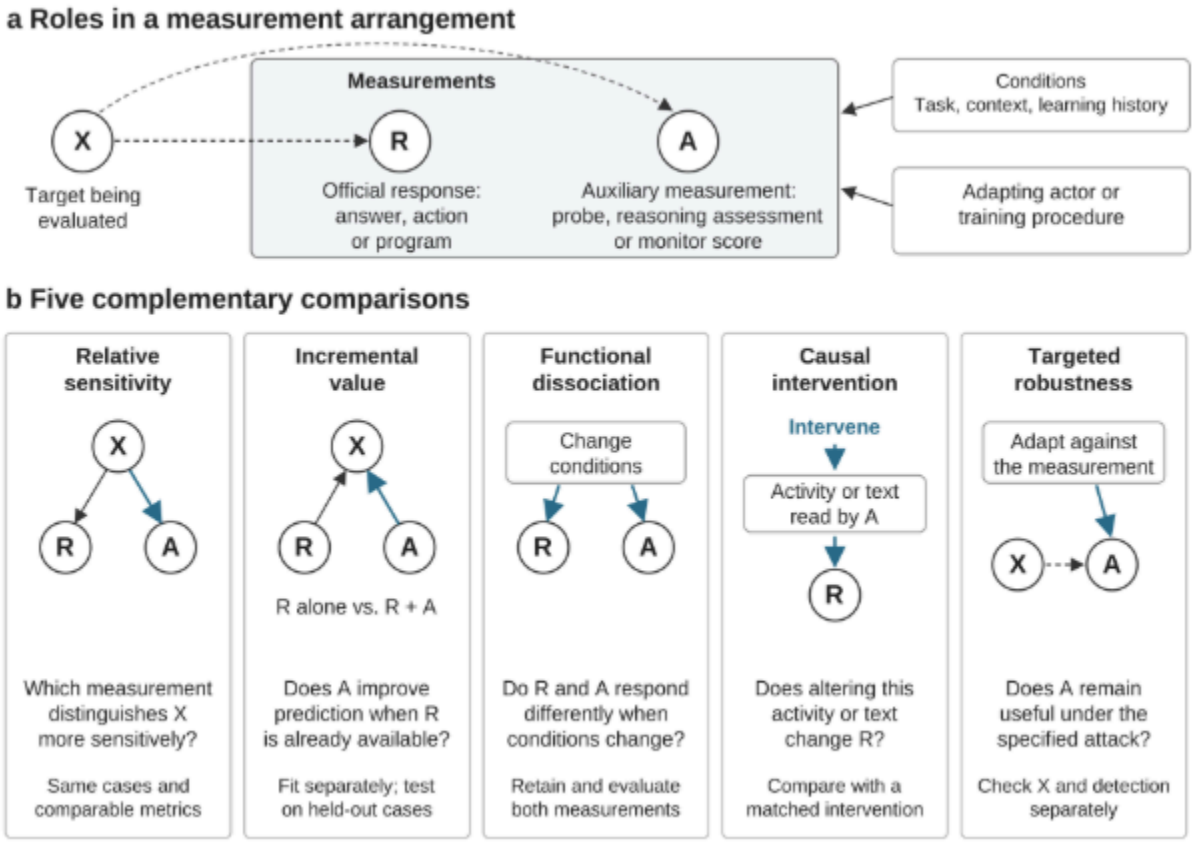


Figure 1

The framework. (a) Experimental roles: the target *X*, the official response *R* and an auxiliary measurement *A*, considered within a measurement arrangement rather than a physical model boundary. Conditions and an adapting actor or training procedure can alter the evidence. (b) Five complementary comparison questions. Arrows indicate the relationships or interventions examined, not a complete causal *model*. In the intervention panel, the activity or text read by A is altered; the probe score itself need not cause R. X is checked independently wherever possible

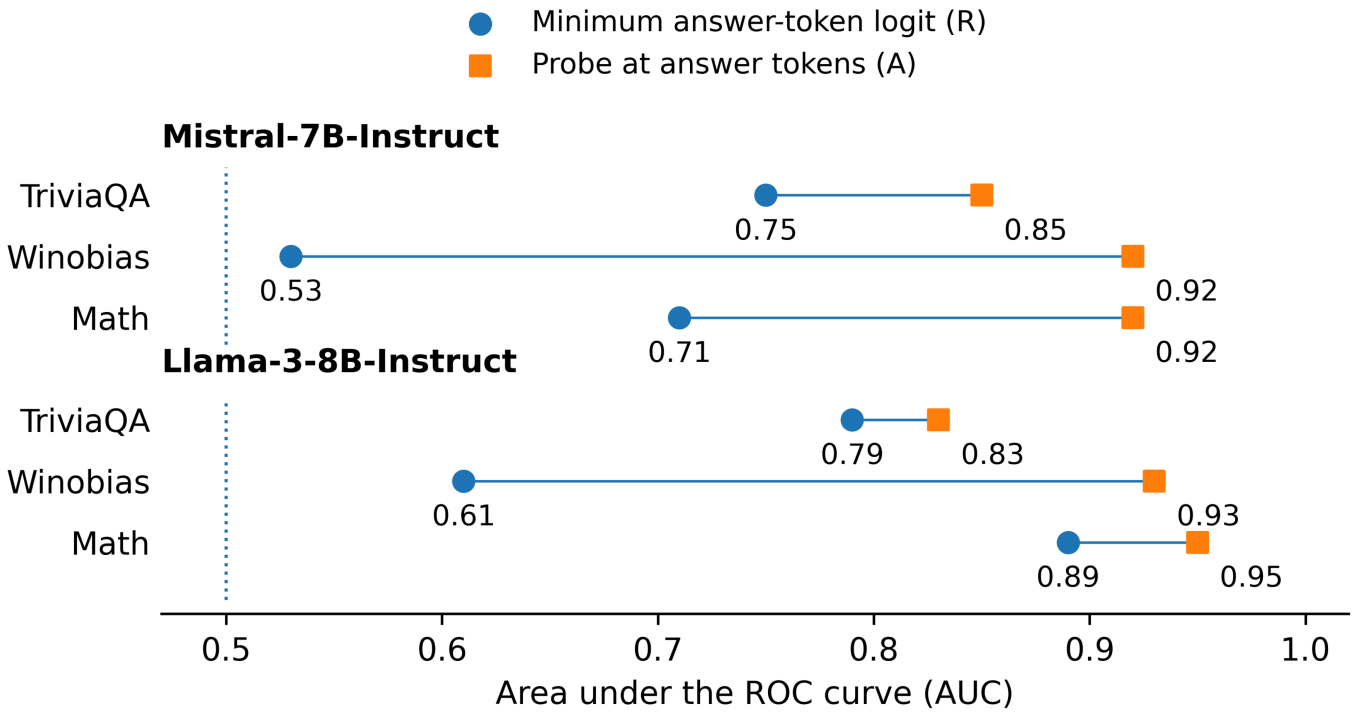


Figure 2

Discrimination of correct from incorrect generated answers in six model–dataset combinations. Circles show the minimum logit across exact-answer tokens; squares show a probe at the last exact-answer token. Both are measured from the run that produced the answer. The logit comparator is the same procedure in every comparison, not the strongest of all non-probe methods. Values are reported AUC point estimates. Redrawn from Orgad et al. (2025), Table 1

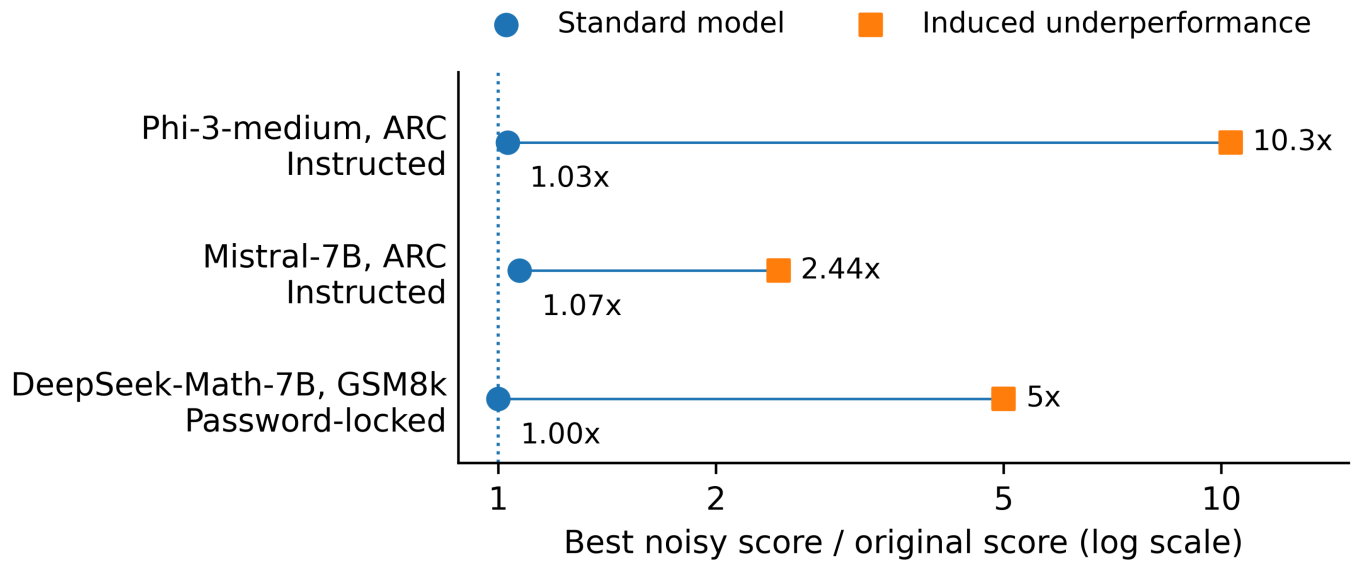


Figure 3

Best score after weight-noise injection divided by the unperturbed score, for three selected comparisons. Circles denote standard models; squares denote models instructed or password-locked to underperform. The horizontal scale is logarithmic. Values are maxima from a search over noise levels and repeated draws, not expected effects of arbitrary noise. ARC contains science questions; GSM8k contains arithmetic word problems. Redrawn from Tice et al. (2025), Table 1

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResourceOctober172026.docx