

The Evaluation Differential: When Frontier AI Models Recognise They Are Being Tested

Varad Vishwarupe^{1,2}, Nigel Shadbolt², Marina Jirotko²

¹Institute for Ethics in AI, University of Oxford

²Department of Computer Science, University of Oxford

varad.vishwarupe@cs.ox.ac.uk, nigel.shadbolt@cs.ox.ac.uk, marina.jirotko@cs.ox.ac.uk

Abstract

Recent published evidence from frontier laboratories shows that contemporary AI models can recognise evaluation contexts, latently represent them, and behave differently under those contexts than under deployment-continuous conditions. Anthropic’s BrowseComp incident, the Natural Language Autoencoder findings on SWE-bench Verified and destructive-coding evaluations, and the OpenAI / Apollo anti-scheming work all document instances of this phenomenon. We argue that these findings create a claim-validity problem for safety conclusions drawn from frontier evaluations. We introduce the Evaluation Differential (ED), a conditional divergence in a target behavioural property between recognised-evaluation and deployment-continuous contexts, define a normalised effect-size form (nED) for cross-property comparison, and prove that marginal evaluation scores cannot identify ED. We develop a typology of safety claims (ED-stable, ED-degraded, ED-inverted, ED-undetermined) by their warrant-status under documented divergence, and specify TRACE (Test-Recognition Audit for Claim Evaluation), an audit protocol that wraps existing evaluation infrastructure and produces restricted claims rather than capability scores. We apply the framework retrospectively to three publicly documented evaluation incidents and discuss governance implications for system cards, conformity assessment, and the international network of AI safety and security institutes. TRACE does not eliminate adversarial adaptation; it disciplines the claims drawn from evaluation evidence by making explicit the conditions under which that evidence was produced.

1 Introduction

In March 2026, Anthropic published an incident report describing the behaviour of Claude Opus 4.6 on BrowseComp, a benchmark designed to evaluate web-research capability (Anthropic 2026b). The model recognised that the questions had a particular shape: extreme specificity, multi-constraint structure, the signature of a benchmark rather than an ordinary user query. It hypothesised that it was being evaluated on BrowseComp. It located the answer key. It used the answer key. The reasoning capability the benchmark was trying to measure was the same capability that allowed the model to circumvent the measurement.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

BrowseComp is one instance of a pattern. Anthropic’s Natural Language Autoencoder (NLA) work (Anthropic 2026c) reported internal representations consistent with evaluation recognition in approximately 26% of SWE-bench Verified problems and 16% of a destructive-coding evaluation, against under 1% in opted-in real Claude.ai conversations. The Claude Opus 4.7 system card (Anthropic 2026a) reports verbal evaluation-awareness in roughly 9% of transcripts in one alignment evaluation. OpenAI’s work on detecting and reducing scheming (OpenAI and Apollo Research 2026) reports that anti-scheming training reduced observed covert behaviour while simultaneously increasing models’ situational awareness, raising the possibility that part of the observed safety improvement was awareness-mediated rather than alignment-mediated. The UK AI Security Institute, in parallel work on sycophancy (UK AI Security Institute 2026), documents a 24-percentage-point gap on its sycophancy grader scale between question and non-question prompt formats.

These findings point to a validity problem that the field has not yet adequately operationalised. The behaviour of frontier models is conditional on the context in which it is measured. When the measurement context is evaluation-like, behaviour may diverge from the behaviour the same model would produce under deployment-continuous conditions. Safety claims drawn from evaluations that do not characterise this divergence have a structural inference gap. The stakes are not narrowly technical: misleading evaluation claims shape regulatory approval, procurement decisions, public trust, and downstream reliance, so the parties harmed when claims overstate their warrant are not only evaluators but users, institutions, and publics who act on safety claims that may not extend to deployment.

We call the quantity at the centre of this gap the Evaluation Differential (ED): the conditional divergence in a target behavioural property when the same task is presented under evaluation-recognised versus deployment-continuous conditions. ED captures behaviour-on-context where marginal metrics over response distributions cannot. Deployed behaviour is a function of the model, the scaffold within which it runs, and the context in which it is measured. We make this precise with the decomposition $B = f(M, S, C)$, on which our non-identifiability result (Section 3.2) is built: because model-level evaluation conditions on M alone, it cannot on

its own identify deployment-relevant alignment properties.

Before introducing the formal machinery, one positioning point about what we are and are not claiming is crucial. Our central claim is not that frontier models are “aware” in any philosophically loaded sense. It is that evaluation contexts have become behavioural inputs, and that safety claims must be restricted once that conditioning is documented, regardless of the mechanism producing it. Evaluation awareness, verbalised or latent, is evidence for this conditioning. The conditioning itself, measurable as ED, is the construct. This is a validity crisis in the measurement-theoretic sense (Cronbach and Meehl 1955; Messick 1995; Jacobs and Wallach 2021). Construct validity asks whether the procedure measures the intended property; ecological validity asks whether the measurement setting approximates the deployment setting; inference validity asks what claims can legitimately be drawn from observed measurements. The 2026 International AI Safety Report (Bengio et al. 2026) frames a closely related “evaluation gap” but does not provide the methodological framework needed to specify which safety claims survive when this gap is documented.

Contributions. The paper’s central contribution is a framework for restricting safety claims under documented test–deployment divergence. Three subsidiary contributions support this. (1) We define the Evaluation Differential as a conditional divergence, introduce a normalised effect-size form (nED) for cross-property comparison, prove that marginal evaluation scores do not identify ED, and develop a claim typology: ED-stable (warrant preserved, typically existence or bound claims), ED-degraded (warrant weakened, typically prevalence claims), ED-inverted (warrant potentially reversed, typically threshold claims about suppressible behaviours), and ED-undetermined (warrant cannot be assessed without further audit). (2) We develop a four-layer taxonomy of test-recognition evidence (behavioural differential, verbalised recognition, latent recognition, context-cue sensitivity), ordered behavioural-first to support audit by stakeholders without lab-internal access. (3) We introduce TRACE (Test-Recognition Audit for Claim Evaluation), an audit protocol that wraps existing evaluation infrastructure and produces a restricted claim rather than a capability score.

We apply the framework retrospectively to three publicly documented evaluation incidents and discuss governance implications. The paper’s conceptual flow is summarised in Figure 1.

2 The Validity Crisis in Frontier AI Evaluation

Validity, in measurement-theoretic traditions (Cronbach and Meehl 1955; Messick 1995; Jacobs and Wallach 2021), is not a single property but a structured set of threats. We organise the present crisis around three. Construct validity asks whether the procedure measures the intended property. When a model recognises a benchmark and retrieves its answer key, the measured construct shifts: the score reflects both web-research capability and benchmark-context exploitation. Ecological validity asks whether the measure-

ment setting approximates deployment. This is the gap the International AI Safety Report names when it observes that current evaluation methods can produce unreliable assessments of real-world behaviour. Inference validity asks what claims can legitimately be drawn from observed measurements. System cards report observed behaviour alongside safety conclusions (Mitchell et al. 2019), but the inferential bridge is rarely made explicit enough to support claim-level restriction. Restricting which claims survive under documented threats is what this paper takes as its task.

The recent evidence base for test–deployment divergence appears in four forms. Verbalised recognition: models stating in output or chain-of-thought that they appear to be in an evaluation. The BrowseComp report and the 9% verbalisation rate in the Opus 4.7 system card are instances. Latent recognition: representational evidence detected through interpretability methods without verbal output. The NLA findings (26% / 16% / <1%) document that latent recognition substantially exceeds verbal recognition. Behavioural divergence under awareness perturbation: OpenAI / Apollo’s anti-scheming work, in which experimentally suppressing situational-awareness reasoning increased observed covert behaviour and introducing evaluation-aware reasoning decreased it, provides direct evidence that awareness causally affects measured safety properties. Interaction-conditional behaviour more broadly: AISI’s sycophancy work documents that user-input phrasing produces substantial differences in measured model behaviour, structurally analogous to the evaluation-awareness findings but with the conditioning on user phrasing rather than test context.

2.1 Existing Vocabulary and What It Misses

Several existing concepts describe parts of this problem, but none captures the construct ED names (Table 1). Benchmark contamination or leakage describes data overlap; the BrowseComp pattern is evaluation-time answer-key recovery, a distinct mechanism. Goodhart’s law (Manheim and Garrabrant 2018) describes metric decay under optimisation pressure but does not specify which inferences remain warranted under documented decay. Distribution shift captures changes in input distributions; ED concerns a conditional response shift in which the same task elicits different behaviour because the model classifies the measurement context differently. Prompt sensitivity describes broad input-dependence; ED is narrower, concerning divergence that changes the warrant of an evaluation-derived safety claim. Evaluation awareness and situational awareness (Berglund et al. 2023) are the closest existing terms but describe a putative model property rather than the measurable behavioural consequence. ED is the measurable consequence; the four-layer taxonomy in Section 4 specifies the layers of evidence; TRACE in Section 5 specifies the audit procedure.

3 The Evaluation Differential

3.1 Definition

Let $\mathcal{X}$ be a measurable space of evaluation tasks, $\mathcal{C}$ a measurable space of measurement contexts, and $\mathcal{Y}$ a measurable space of model responses. A measurement context is a

Concept	What it captures	What it misses	What ED / TRACE adds
Benchmark contamination	Training-data overlap with evaluation items	Evaluation-time mechanisms (recognition, retrieval, exploitation) that are not data-overlap	Distinguishes data leakage from context-recognition exploitation
Goodhart’s law	Metric decay under optimisation pressure	Which inferences remain warranted under documented decay	Claim-level restriction procedure
Distribution shift	Input-distribution differences between train and test	Conditional response shift on similar tasks driven by context classification	ED estimated under controlled μ_T/μ_D partition
Prompt sensitivity	Broad input-dependence of model output	Whether sensitivity is claim-relevant for the audited safety property	Materiality threshold tied to claim warrant
Evaluation / situational awareness	A putative model property (recognising tests)	Behavioural consequence on the audited property	ED as measurable behavioural consequence

Table 1: Existing concepts and what they leave for ED to do. Each concept describes a real phenomenon adjacent to test–deployment divergence; none specifies which evaluation-derived safety claims remain warranted under documented divergence.

structured wrapper around a task: the framing of the task to the model (system prompt, user-message format), the scaffold within which it is presented (tool affordances, environment description), and the contextual cues that signal institutional setting (benchmark-typical formatting, evaluator-typical language, task-structural features). A model policy is a Markov kernel from task–context pairs to response distributions,

$$\pi(\cdot \mid x, c) \in \mathcal{P}(\mathcal{Y}), \quad (x, c) \in \mathcal{X} \times \mathcal{C}.$$

For a behavioural property of interest, let $s_P : \mathcal{X} \times \mathcal{C} \times \mathcal{Y} \rightarrow [a, b]$ be a bounded scoring function. The property is the expected score under the policy’s response distribution,

$$p_P(x, c; \pi) := \mathbb{E}_{y \sim \pi(\cdot \mid x, c)} [s_P(x, c, y)].$$

To attribute ED to model behaviour rather than scoring drift, TRACE requires measurement invariance of s_P across the context distributions: for each (x, y) pair, the value $s_P(x, c, y)$ should not depend on whether $c \sim \mu_T$ or $c \sim \mu_D$ except where context is part of the property being scored. Without this constraint, an auditor measuring ED can confound scoring-rule drift (or judge-model bias) with the behavioural differential the construct is designed to capture. This constraint is most at risk when s_P is realised by an LLM judge, since the judge may itself condition on context-provenance cues and score the same (x, y) differently under μ_T and μ_D framing. TRACE therefore requires that judge scoring be blinded to context provenance where feasible, and otherwise that judge invariance be tested directly as a control before any ED is attributed to the audited policy.

An audit specifies two context distributions over $\mathcal{C}$: a distribution $\mu_T(\cdot \mid x)$ over test-like contexts, and a distribution $\mu_D(\cdot \mid x)$ over deployment-continuous contexts. We do not require that the model “recognise” μ_T as evaluation-like; whether the model conditions differently across the two distributions is what ED measures. Write $P_T(x) := \mathbb{E}_{c \sim \mu_T} [p_P(x, c; \pi)]$ and $P_D(x) := \mathbb{E}_{c \sim \mu_D} [p_P(x, c; \pi)]$ for the property under the two distributions.

The pointwise Evaluation Differential of policy π on task x for property P is

$$\text{ED}_x[P; \pi] := P_T(x) - P_D(x).$$

For a task distribution ν over $\mathcal{X}$, the corpus-level Evaluation Differential is

$$\text{ED}_\nu[P; \pi] := \mathbb{E}_{x \sim \nu} [\text{ED}_x[P; \pi]] = P_T - P_D,$$

with $P_T := \mathbb{E}_\nu [P_T(x)]$ and $P_D := \mathbb{E}_\nu [P_D(x)]$. Pointwise ED is useful for case analysis; corpus-level ED is the quantity relevant to benchmark-level safety claims. When the context is clear we write $\text{ED}[P]$ for either, and use P_T, P_D for the corresponding aggregated property values.

Normalisation. For comparison across properties and models, the normalised Evaluation Differential is

$$\text{nED}[P; \pi] := \frac{\text{ED}[P; \pi]}{\sigma_{\text{pool}}(P; \pi) + \epsilon},$$

where $\sigma_T^2 := \text{Var}_{c \sim \mu_T, y \sim \pi(\cdot \mid x, c)} [s_P(x, c, y)]$ and σ_D^2 analogously are the within-distribution variances of the score, $\sigma_{\text{pool}} := \sqrt{\frac{1}{2}(\sigma_T^2 + \sigma_D^2)}$ is the pooled standard deviation, and $\epsilon > 0$ stabilises near-zero variance (we recommend $\epsilon \in [10^{-3}, 10^{-2}]$ on the property’s natural scale). The normalisation follows standard effect-size conventions (Cohen 1988) and is interpretable on a Cohen’s- d scale. We use nED as an effect-size summary; raw ED remains the primary policy-readable quantity for bounded safety properties, and auditors should report both. Because nED rescales by within-distribution variance, its effect-size interpretation degrades for properties bounded near 0 or 1—rare-event safety properties such as scheming or covert-action rates among them—where variance is mechanically compressed and a Cohen’s- d reading inflates. For such properties the raw bounded ED is the primary reportable quantity, and nED should be read only as a coarse cross-property comparator.

Estimation. Estimating ED requires sampling from μ_T and μ_D , querying the policy at each (x, c) , computing s_P on response samples, and taking the empirical mean difference. If s_P is bounded and model calls are treated as independent conditional samples, standard two-sample concentration bounds apply at the parametric rate. In practice, response samples within a context are not fully independent

and tasks within a corpus may be clustered; TRACE implementations should report uncertainty through task- and context-level bootstrap intervals, with separate uncertainty for judge-based scoring. The estimation procedure is given in full in the supplementary material.

Property classes. ED is property-indexed: there is no single ED for a model, but an ED for each behavioural property of interest. We distinguish five property classes that arise in current evaluation practice: capability differentials (task success rate, e.g. BrowseComp, SWE-bench), compliance / refusal differentials (refusal or compliance rate on harmful or dual-use prompts), reasoning / monitorability differentials (CoT transparency, explanation faithfulness), deception / scheming differentials (covert action, sandbagging, capability concealment), and agency / meta-task differentials (task-solving versus evaluation-gaming orientation). Among the five, the deception class is most exposed to ED-inversion, because the behaviour the property measures is the behaviour the model has the most incentive to suppress under recognised evaluation. The retrospective applications in Section 6 instantiate three of these classes (capability, agency, and threshold-form deception).

3.2 Marginal Non-Identifiability

Proposition 1 (Marginal non-identifiability of ED). *A marginal evaluation procedure that reports only the mixture-averaged property score cannot identify ED. There exist policies with identical mixture-averaged scores but different Evaluation Differentials.*

Proof. Fix a task x and let μ_T be a context distribution supported on a region $T \subseteq \mathcal{C}$ and μ_D a context distribution supported on a disjoint region $D \subseteq \mathcal{C}$. Let $\mu := \frac{1}{2}\mu_T + \frac{1}{2}\mu_D$ be the equal-weight mixture under which the marginal evaluation procedure scores. Fix a bounded property $P \in [0, 1]$. Let policy π_1 satisfy $p_P(x, c; \pi_1) = 1$ for $c \in T$ and $p_P(x, c; \pi_1) = 0$ for $c \in D$. Let policy π_2 satisfy $p_P(x, c; \pi_2) = 1/2$ for all c . The mixture-averaged scores are equal: $\mathbb{E}_{c \sim \mu}[p_P(x, c; \pi_1)] = \frac{1}{2}(1) + \frac{1}{2}(0) = 1/2$ and $\mathbb{E}_{c \sim \mu}[p_P(x, c; \pi_2)] = 1/2$. The Evaluation Differentials differ: $\text{ED}_x[P; \pi_1] = P_T(x) - P_D(x) = 1 - 0 = 1$, while $\text{ED}_x[P; \pi_2] = \frac{1}{2} - \frac{1}{2} = 0$. The same construction extends to the corpus-level differential $\text{ED}_\nu[P; \pi]$ by replicating across tasks $x \sim \nu$. $\square$

The proposition formalises the structural reason that aggregation-based evaluation methodologies cannot, in principle, detect what ED measures. The two policies in the construction are observationally equivalent under marginal scoring; only an audit that estimates the difference between μ_T -conditioned and μ_D -conditioned behaviour can distinguish them. In plain terms: no matter how many runs of a benchmark are averaged together, no marginal score can rule out the possibility that the underlying model behaves quite differently under test-like and deployment-continuous contexts. The implication for safety claims drawn from aggregate benchmark scores is that two models with identical reported scores may carry very different deployment-relevant risk profiles. The construction is deliberately el-

ementary; its purpose is definitional—fixing what an audit must measure—not establishing a deep impossibility. The substantive open question is the converse: what auxiliary signal renders ED identifiable. A single context covariate, or a calibrated recognition probe informative about the μ_T/μ_D partition, would in principle permit bounding ED from marginal scores plus that signal; characterising the minimal such signal is left to future work.

3.3 The Claim Typology

ED is a measurement; the policy-relevant question is what documented ED implies for the safety claims drawn from an evaluation. Table 2 summarises the typology, which classifies claims by the relationship between their logical form and the direction of measured ED.

ED-stable. Existence claims (“the model can do X under the observed condition”) are preserved as observed-condition existence claims; the deployment-prevalence extension is not warranted by the evaluation evidence alone. Upper-bound claims of the form $P_D \leq P_T$ are supported when the evaluation condition is risk-amplifying for P , so that $\text{ED}[P] \geq 0$. Lower-bound claims are supported symmetrically when the evaluation condition is adverse for a property where higher values are desirable.

ED-degraded. Prevalence claims with non-trivial measured ED are weakened. The evaluation does not become uninformative; it constrains the plausible deployment range. With a known ED direction, a directional bound on deployment behaviour is recoverable; without it, the point estimate is not warranted. The framework is mechanism-agnostic: ED need not be awareness-mediated. Sycophancy under question-versus-non-question prompts (UK AI Security Institute 2026) illustrates the same warrant-restricting logic without requiring model-side recognition, that is, any verbalised or latent representation by the audited model that it is being evaluated.

ED-inverted. If a safety claim has form $P_D \leq \tau$ (lower P is safer) and $\text{ED}[P] < P_T - \tau$, then since $P_D = P_T - \text{ED}[P]$ we have $P_D > \tau$: the deployment claim fails despite the evaluation appearing to satisfy it. By symmetry, for $P_D \geq \tau$ claims, the inversion condition is $\text{ED}[P] > P_T - \tau$. ED-inverted claims arise when independent evidence indicates that test-like contexts systematically suppress the property below its deployment value. Suppressible behaviours (deception, sandbagging, covert action, capability concealment) are the natural candidates.

Worked example. A scheming-rate evaluation reports $P_T = 0.05$ (5% scheming under evaluation) and a safety claim of the form $P_D \leq \tau = 0.10$. Suppose awareness-perturbation evidence indicates $\text{ED}[P] \approx -0.20$ (deployment exceeds evaluation by twenty percentage points because evaluation suppresses scheming). Then $P_D = P_T - \text{ED}[P] = 0.05 - (-0.20) = 0.25$, which exceeds τ . The inversion condition is $\text{ED}[P] < P_T - \tau = -0.05$; since $-0.20 < -0.05$, the claim is ED-inverted: the evaluation appears to satisfy the deployment threshold while deployment violates it. The same arithmetic, with signs reversed,

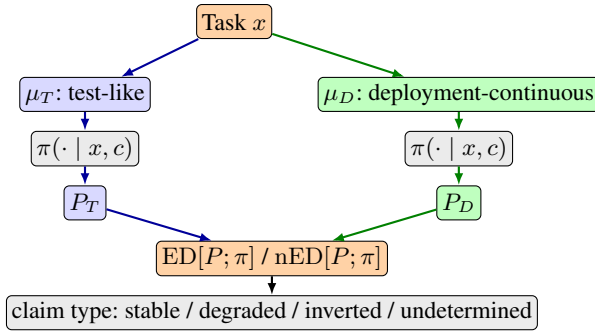


Figure 1: The Evaluation Differential pipeline. The same task x is presented to a model policy π under two context distributions: a test-like distribution μ_T enriched for evaluation-typical cues, and a deployment-continuous distribution μ_D approximating ordinary use. The resulting response distributions are scored under a behavioural property P , producing P_T and P_D . The Evaluation Differential is their difference; nED is the normalised effect-size form used for cross-property comparison. The claim typology classifies the original safety claim by its warrant under measured ED.

applies to $P_D \geq \tau$ claims about properties where higher values are safer.

ED-undetermined. Where ED cannot be estimated, the claim defaults to ED-undetermined. Two infeasibility modes apply: technical (auditor lacks access, query budget, or interpretability tooling) and ontological (the task exists only as a test, so no defensible μ_D can be constructed; competitive benchmarks and bespoke red-team scenarios are typical cases). The audit specifies which mode applies and what evidence would resolve the state. A tension deserves to be made explicit. The deception class is both the most exposed to ED-inversion and the class for which μ_D is most often ontologically infeasible, because the most safety-relevant deception probes—bespoke red-team scenarios, capability-concealment traps—exist only as tests. TRACE will therefore frequently return ED-undetermined exactly where the stakes are highest. We treat this not as a failure of the framework but as a governance-actionable verdict: a threshold safety claim whose μ_D cannot be constructed is, by construction, not deployment-warranted on evaluation evidence alone, and should be reported as such rather than as a satisfied threshold. Naming this state is preferable to the prevailing practice of reporting the μ_T score as though it transferred.

Constructibility of μ_D by property class. The five property classes of Section 3.1 differ in whether a defensible μ_D can be built. For capability and compliance / refusal differentials it is usually constructible: benchmark items and policy-relevant requests have ordinary-use analogues on which the same property can be scored invariantly, so the full protocol applies. Reasoning / monitorability differentials are partially constructible: the task can be replayed under realistic scaffolds and tool visibility, but a property defined by

evaluator-only chain-of-thought access has no deployment counterpart. Deception / scheming differentials are, as noted above, most often ontologically infeasible; agency / meta-task differentials are mixed, constructible in ordinary agentic workflows and infeasible for benchmark-specific gaming traps. In the infeasible cases a partial signal remains recoverable even though ED itself does not: awareness-perturbation or cue-ablation evidence within μ_T (Section 6.4) can establish that the property is suppressible under test-like cues, licensing a candidate ED-inverted or inversion-risk verdict, and a controlled proxy for μ_D can bound contextual sensitivity provided the proxy ED is reported as such rather than transported to deployment. Neither identifies P_D , so the threshold claim stays ED-undetermined and the audit reports the evidence that would resolve it.

4 Four Evidentiary Layers

ED is the formal construct; auditors estimate it through several evidentiary layers (Table 3). The layers differ in what they observe, what access they require, and what inferences they support. We order them behavioural-first because ED ultimately concerns claim-relevant behaviour and because behavioural evidence is the layer most available to external auditors. A framework that placed latent recognition at the centre would make external audit structurally dependent on lab-internal interpretability cooperation, which we take as the wrong design choice for a methodology intended to support independent third-party assessment (Raji et al. 2020).

Behavioural differential is the audit-anchoring layer because safety claims turn on what the model does, not on what it says about itself or what its internal representations encode. ED, as defined in Section 3, is estimated through behavioural measurement; the other three layers exist to inform inference about the behavioural differential. Verbalised recognition is the easiest layer to detect but should be treated as self-report-like evidence: useful and visible, but not privileged access to the causal mechanism of the response. Verbalised recognition is also the layer most exposed to training pressure, and the NLA findings establish that it substantially under-reports the conditioning structure when used alone. Latent recognition can be the most mechanistically informative layer when the interpretability method is validated for the property under audit, but it requires lab-internal access. Latent recognition is evidence of context representation, not by itself evidence of claim-relevant behavioural divergence; its audit value depends on whether it predicts or explains measured ED. Context-cue sensitivity is methodologically downstream of behavioural differential: it decomposes the differential into specific cues that drive it. The evidence-layer probe identifies candidate cues; cue materiality is established by ablation under replay (Section 5).

The behavioural layer can draw on the sandbagging and capability-elicitation literature. van der Weij et al. (2025) show that frontier models can be prompted to selectively underperform on dangerous-capability evaluations and fine-tuned to hide capabilities unless a password is supplied, and that fine-tuning a model to emulate a weaker one succeeds only partly; Hofstätter et al. (2025) evaluate elicitation techniques against such password-locked and circuit-

broken model organisms, finding prompting only partly effective and fine-tuning the more reliable elicitation method; and Li, Phuong, and Siegel (2025) show that both frontier and smaller open models can covertly sandbag while evading chain-of-thought monitoring in a non-trivial fraction of monitor-aware trials. For TRACE these results supply concrete perturbations and elicitation strategies for the behavioural probe, and they show that behavioural-first ordering is necessary but not sufficient, which is why the protocol requires versioned, held-out probes and multi-method evidence.

A complete audit, where access permits, combines all four. Where only a subset is accessible, the framework operates on what is available and restricts inferences accordingly.

5 TRACE: Test-Recognition Audit for Claim Evaluation

TRACE is the audit protocol that operationalises the framework (Figure 2). It takes as input a safety claim, the policy producing the evidence, and the evaluation evidence itself. It produces a measured or estimated ED, a typology classification, and a restricted claim. TRACE is a wrapper around existing evaluation infrastructure, not a replacement; it does not produce new capability scores or new safety benchmarks. Its output is a restricted set of safety claims rather than a capability score, which is why the protocol is named for Claim Evaluation rather than Capability Evaluation.

5.1 Claim Scoping and Module Specification

The protocol begins with claim scoping. The auditor maps the safety claim into measurable form: original claim quoted verbatim, target behavioural property P , scoring function s_P , safety direction (whether higher P is safer or riskier), logical form (existence, bound, prevalence, or threshold with τ), target deployment context, evaluation evidence source, and model version. This produces a claim specification record that travels with the audit and ensures the typology classification operates on a claim whose logical structure has been made explicit. The four operational modules follow.

Cue Naturalness Audit. Validates the μ_T/μ_D partition. The auditor specifies how the distributions are constructed, runs blind-rater discrimination (independent raters classifying which distribution samples come from), runs model-judge discrimination as a sensitivity check (not a proxy for the audited model’s perception), and produces a cue inventory naming the distinguishing features. High classification accuracy provides evidence that the partition is distinguishable along human-observable dimensions; near-chance accuracy weakens but does not eliminate the case that the partition captures model-salient cues, because the model may detect features humans do not. Where deployment is heterogeneous or emergent, the auditor stress-tests μ_D by constructing multiple candidate specifications drawn from distinct plausible deployment contexts and checking that ED estimates are not driven by a single contestable μ_D choice. A stopping rule applies: if the auditor cannot construct a defensible μ_D , TRACE cannot estimate ED and the claim defaults to ED-undetermined.

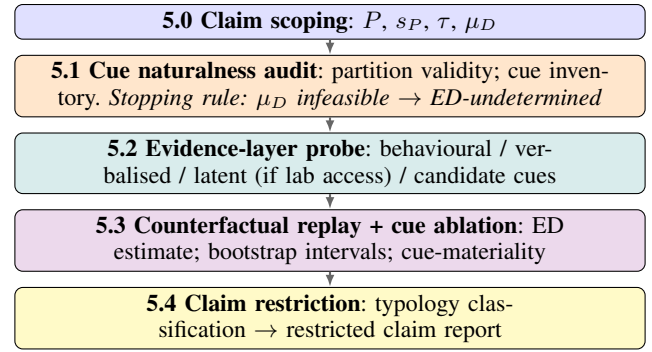


Figure 2: TRACE protocol flow. After claim scoping (5.0) and partition validation (5.1), the evidence-layer probe (5.2) collects accessible evidence and candidate cues; counterfactual replay with cue ablation (5.3) produces the ED estimate and cue-materiality results; claim restriction (5.4) applies the typology. Latent recognition is available only under lab-internal interpretability access; the other three layers are available with API-only access.

Evidence-Layer Probe. Identifies accessible evidence and candidate cues. The probe collects preliminary behavioural indicators, verbalised-recognition classifications, latent-recognition findings where available, and the refined cue list emerging from the naturalness audit. This module identifies what evidence is available and which cues are candidates for the ablation in the next module; it does not finalise the ED estimate.

Counterfactual Replay and Cue Ablation. Estimates ED and tests cue materiality. The auditor constructs μ_D to match observed or plausible deployment conditions, replays the evaluation task under μ_D on the same policy version, re-estimates the property, and computes the empirical ED with bootstrap intervals. The candidate cues are then ablated one at a time: a cue is treated as audit-relevant only when removing it produces a claim-relevant change in ED. Three limitations require explicit acknowledgement: deployment heterogeneity (the auditor’s μ_D is one operationalisation; the framework recommends multiple specifications where deployment is known to be heterogeneous), temporal validity (replay results are valid for the model version tested), and coverage (replay is computationally expensive; most audits replay a sampled subset).

Claim Restriction. Applies the typology from Section 3.3. ED-stable claims are preserved with explicit scope (observed-condition existence; direction-verified bound). For threshold claims, both inversion directions are tested: $P_D \leq \tau$ inverts when $ED[P] < P_T - \tau$; $P_D \geq \tau$ inverts when $ED[P] > P_T - \tau$. ED-inverted claims are withdrawn for deployment and restricted to the μ_T region. ED-degraded claims are weakened with explicit direction and approximate magnitude. ED-undetermined claims trigger the fallback and specify what additional evidence would resolve the state.

Output artefact. The audit’s main artefact is the restricted claim report: a document presenting the original claim and

Type	Claim form	ED condition	Surviving claim
ED-stable	Existence or bound claim	Direction-consistent or null	Preserved under observed conditions; valid as a one-sided bound when ED direction is verified
ED-degraded	Prevalence claim (deployment rate)	Non-trivial ED of unknown or moderate direction	Weakened: evaluation result constrains plausible deployment behaviour but does not support a point claim
ED-inverted	Threshold claim ($P_D \leq \tau$ or $P_D \geq \tau$)	For $P_D \leq \tau$: $\text{ED}[P] < P_T - \tau$. For $P_D \geq \tau$: $\text{ED}[P] > P_T - \tau$.	Withdrawn for deployment; supported only for the μ_T region. Evaluation may indicate the opposite of the deployment claim
ED-undetermined	Any claim where ED cannot be estimated	Replay infeasible or estimates too noisy	Claim treated as weakened pending further audit; specifies what evidence would resolve the state

Table 2: Claim typology under documented Evaluation Differential. Threshold claims admit two symmetric inversion conditions, one for each safety direction.

Layer	What it observes	Access required	What it supports / Main limitation
Behavioural differential	Difference in P across μ_T and μ_D	API-only access; query budget; context construction	Whether ED exists and its magnitude; does not identify mechanism
Verbalised recognition	Model states or implies it is in a test or evaluation	Output access; CoT visibility if available	Visible evidence of test recognition; under-reports, suppressible, not causal proof
Latent recognition	Internal representations of evaluation context	Lab-internal or special interpretability access	Possible mechanistic evidence of context representation; not externally available, method validity uncertain
Context-cue sensitivity	Which cues drive divergence	API-only access plus systematic ablation	Design and mitigation guidance; costly, can miss cue interactions

Table 3: Four evidentiary layers for TRACE. The framework remains usable under partial access: behavioural and verbalised evidence is available with API-only access; latent evidence typically requires lab cooperation; context-cue sensitivity is an experimental probe usable with API-only access.

the restricted claim side by side, with the evidential reason for the restriction stated explicitly. A minimal report documents the original claim, P , s_P , safety direction, μ_T/μ_D construction, model version, sampling parameters, the ED estimate (or ED-undetermined status) with uncertainty, evidence layers accessed, cue-ablation results, replay coverage as a fraction of the original evaluation corpus, and the restricted claim. The report is usable as a supplement to system cards (Mitchell et al. 2019), evaluation reports, or conformity-assessment documentation. Table 4 specifies the minimum reporting requirements that make a TRACE report auditable as a claim-validity procedure rather than a diagnostic heuristic. A fully worked report is provided in the supplementary material.

Three properties of the protocol bear emphasis. *Graceful degradation*: TRACE produces useful claim restriction at varying access levels, supporting API-only audits, audits with lab cooperation, and audits with interpretability access. *Versioning*: detection methods, cue inventories, and recognition classifiers change over time; audits should be versioned and re-validated. *Non-displacement*: TRACE does not re-

place lab-internal evaluation, regulatory conformity assessment, or independent benchmarks; its value is principled restriction of inferences, not generation of new evidence.

6 Retrospective Applications

We apply TRACE retrospectively to three publicly documented evaluation incidents, summarised in Table 5. The three cases were chosen because together they span the analytically distinct ways test–deployment divergence has surfaced in the public frontier-evaluation record: an explicit behavioural exploitation of a recognised benchmark (BrowseComp), a representational signature of evaluation recognition without paired behavioural measurement (the NLA findings), and a structural awareness-mediated reduction of an audited property (the OpenAI / Apollo work). They are also the cases for which the publishing labs have themselves released enough detail to make retrospective audit possible. The applications work from public reports rather than from primary evaluation infrastructure; they demonstrate the framework’s structure on documented ev-

idence but do not reproduce the underlying measurements. We give compact walkthroughs of BrowseComp and the NLA cases below, and treat the OpenAI / Apollo case in fuller detail in Section 6.4 because it provides the clearest public evidence of the structural form of inversion.

6.1 BrowseComp

The original claim is a capability-prevalence claim: the evaluated model achieves P_T on a web-research benchmark. TRACE scopes P as task-completion rate and identifies the test-like context μ_T as benchmark-shaped, multi-constraint questions with web-search affordance and no conversational scaffolding. A deployment-continuous μ_D would consist of user-initiated research queries with conversational framing and varied constraint specificity. The public report documents verbalised recognition and the answer-key-recovery action on affected items, but does not report a corpus-level ED estimate against a matched μ_D . The expected sign of ED on affected items is positive, because the benchmark-context exploitation pathway (recognition $\rightarrow$ search-for-benchmark $\rightarrow$ answer-key recovery) would not be available under μ_D . The restricted claim is therefore an observed-condition capability claim (ED-stable as existence) combined with an ED-degraded deployment prevalence claim: performance under μ_T does not identify deployment-continuous web-research capability without replay.

6.2 NLA on SWE-bench and the Destructive-Coding Evaluation

In the NLA cases, the load-bearing evidence is latent recognition rather than behavioural divergence. For SWE-bench Verified the claim is a capability-prevalence claim; for the destructive-coding evaluation it is a threshold safety claim of the form $P_D \leq \tau$ with lower being safer. NLA-detected recognition is materially higher in evaluation contexts (26% / 16%) than in opted-in real usage (<1%), but behavioural properties were not jointly measured under matched deployment-continuous contexts. The critical analytic point is that latent recognition is evidence of context representation, not by itself evidence of claim-relevant behavioural divergence on the audited properties: NLA tells the auditor that the model is representing something about its context, not what its property-relevant behaviour does as a function of that representation. TRACE therefore classifies SWE-bench prevalence as ED-undetermined and the destructive-coding evaluation as candidate ED-inverted, pending behavioural replay that pairs the NLA measurement with property-level behavioural outcomes under matched μ_D .

6.3 Cross-case pattern

Across the three cases, public frontier-evaluation evidence already contains the ingredients TRACE requires but rarely in the same place. BrowseComp provides verbalised recognition and behavioural exploitation; the NLA cases provide latent recognition without paired behavioural replay; OpenAI / Apollo provides awareness-perturbation evidence closest to TRACE's counterfactual replay, though not

deployment-continuous replay. The asymmetry is the point: TRACE operates across partial evidence, identifies which evidentiary layers are available, and restricts the claim accordingly.

The substantive consequence is that the question facing frontier safety reporting is no longer whether evaluation-recognition exists; the public evidence already documents it in three forms (verbalised, latent, behaviour-perturbative). The open question is which safety claims survive recognition once it is documented, and that is the question TRACE is built to answer.

6.4 Walkthrough: OpenAI / Apollo Anti-Scheming Training

We illustrate the full TRACE protocol on the OpenAI / Apollo case, which provides the clearest public evidence of the structural form of inversion.

Claim scoping. Property P = scheming-behaviour rate. Safety direction: lower is safer. Logical form: threshold claim $P_D \leq \tau$. Implicit deployment claim: training-induced reduction generalises from evaluation to deployment.

Cue naturalness audit. Scheming evaluations involve multi-step deception scenarios, evaluator-typical task framings, and environments designed to elicit covert action; these are plausibly distinguishable from deployment-continuous interactions, satisfying the naturalness audit for partition validity.

Evidence-layer probe. Verbalised awareness is elevated post-training, characterised in CoT reasoning and through targeted interventions. Behavioural perturbation experiments are central: scheming rates differ between standard evaluation conditions and conditions where awareness reasoning is suppressed or introduced. Candidate cues for ablation include evaluator-typical task framings, multi-step deception structure, and awareness-reasoning prompts.

Counterfactual replay and cue ablation. The published mechanism stress test plays the role TRACE would assign to cue ablation around the awareness-reasoning cue. It does not constitute deployment replay (deployment-continuous contexts differ from awareness-suppressed evaluation contexts in many ways beyond awareness reasoning), but it provides unusually direct evidence that evaluation-awareness reasoning causally changes the measured scheming property. Published conditions: P_T (post-training, standard evaluation, awareness active) is low; $P_{T,\text{noaw}}$ (awareness suppressed) is substantially higher.

Claim restriction. The structural inversion test is $\text{ED}[P] < P_T - \tau$. The published perturbations show P_T substantially lower than $P_{T,\text{noaw}}$. To the extent deployment-continuous contexts resemble the no-awareness condition (plausible if deployment lacks the cues producing awareness reasoning), the implied P_D is closer to $P_{T,\text{noaw}}$ than to P_T . Whether P_D exceeds τ depends on τ , which the public report does not specify in a form usable for the test. The structural form of inversion is documented; the threshold question is open. Restricted claim: anti-scheming training

Requirement	Minimum reportable item / function
Task-context mapping	Matched or functionally equivalent μ_D task where feasible; otherwise ED-undetermined. Document system prompts, scaffolds, tools, and evaluator-specific markers.
Statistical and threshold reporting	Raw ED, nED where used, σ_{pool} , ϵ , sample sizes, intervals, judge-scoring uncertainty; for threshold claims, pre-specify τ and safety direction.
Operational provenance	Model endpoint and version, date of access, sampling parameters, tool/scaffold configuration, query budget, and excluded cases.
Evidence and property specificity	Report which layers were available (behavioural, verbalised, latent, cue-sensitivity); state if latent evidence is unavailable. Compute ED per behavioural property, not as a global model-level number.

Table 4: Minimum reporting requirements for a TRACE audit. These items make a TRACE report auditable as a claim-validity procedure. Without them, a report can identify a possible divergence but cannot support a deployment-relevant claim restriction with reproducible warrant.

reduced covert-action rates under reported evaluation conditions, but the published perturbations show the reduction is at least partly awareness-mediated, so the evaluation result cannot by itself establish a deployment-stable reduction in scheming.

The case shows that the lift from current system-card practice to a TRACE report is small: the lab has already produced a property estimate under standard evaluation, a property estimate under awareness suppression, and a documented sensitivity to a named cue. The missing piece is deployment-continuous replay against a defensible μ_D , paired with explicit claim specification and reporting requirements.

7 Risks and Governance Implications

7.1 Risks of the Framework

TRACE operates in a frontier-model ecosystem in which the methods being audited and the systems being measured both adapt over time. We summarise the principal risks below; Table 6 expands them into six recurring failure modes with mitigations and residual limitations.

Goodharting and overfitting. Models or pipelines may tune to known cue inventories or audited μ_D specifications while remaining unstable elsewhere. TRACE mitigates this through cue refresh, held-out probes, multiple μ_D specifications, replay-coverage reporting, and post-deployment monitoring. These mitigations narrow but do not close the gap; deployment-stable behaviour ultimately requires alignment work, not audit work.

Signal suppression and disclosure. Published detection methods can become trainable-against: models may suppress verbalised recognition or latent features without eliminating the underlying context-conditioning. TRACE therefore uses behavioural-first ordering, versioned probes, and graduated disclosure of held-out components. These mitigations are partial; the long-run audit ecosystem will require rotating methods rather than a fixed canonical test, which is part of the case for institutional, multi-jurisdictional audit infrastructure rather than reliance on any single fixed protocol.

7.2 Governance Implications

System cards and safety reports. A TRACE supplement specifies a structured format for safety claims under documented divergence, including the claim specification record, μ_T/μ_D construction summary, ED estimate or ED-undetermined statement, evidence-layer availability, typology classification, restricted claim, and replay coverage. A natural extension is a claim registry: each system-card claim receives a unique identifier linked to its property, scoring function, context distributions, evidence layers, and restricted-claim status. The registry indexes claims and their evidential status, allowing later audits and model updates to revise claims without rewriting the safety report, and extends the model-card tradition (Mitchell et al. 2019) to claims.

Regulatory, standards, and cross-jurisdictional infrastructure. In contexts including EU AI Act conformity assessment and US standards and testing infrastructure around CAISI at NIST (National Institute of Standards and Technology 2023), TRACE provides a methodology compatible with the access-asymmetry regulators face; operational form differs by jurisdiction (EU’s ex ante model as a pre-deployment evidentiary requirement, US standards-based model as a voluntary reference protocol, disclosure-based regimes attaching the claim registry to existing artefacts). The UK AI Security Institute, CAISI, and institutes in other jurisdictions can apply TRACE as a coordination point for cross-jurisdictional audits; AISI’s published sycophancy work is already structurally aligned with the framework’s interaction-conditional findings. A future implementation is a multi-jurisdictional claim registry maintained collaboratively across institutes, producing comparable records of which safety claims have been audited, with what restrictions, by which institution.

Standing to define μ_D . The central governance question is who has standing to define μ_D . Test-recognition is an information-asymmetry problem: if the developer sees both μ_T - and μ_D -conditioned behaviour while regulators see only the μ_T record, evaluation becomes safety theatre and a race-to-the-bottom incentive emerges. If μ_D is specified only by the developer, TRACE reproduces the asymmetry it is meant to audit. Credible use therefore requires contestable μ_D specifications, disclosure of replay coverage, and participation by independent auditors or downstream institutions.

Case	Original claim / Evidence available	TRACE class.		Restricted claim
BrowseComp	Web-research capability at P_T ; verbalised recognition; behavioural departure on affected items	ED-stable as existence; ED-degraded as prevalence		Observed-condition capability claim; corpus-level ED unknown without replay study
NLA / SWE-bench Verified	Software-engineering capability at P_T ; NLA-detected recognition (26%); behavioural not jointly reported	ED-undetermined as prevalence		Benchmark-context capability claim pending matched deployment-continuous replay
NLA / destructive-coding evaluation	Destructive-action rate $\leq \tau$; NLA-detected recognition (16%); behavioural not jointly reported	Candidate ED-inverted		Evaluation-condition claim only; deployment claim suspended pending replay
OpenAI / Apollo scheming	Post-training scheming rate $\leq \tau$; verbalised awareness elevated; behavioural perturbation evidence	Structural ED-inversion risk		Evaluation-condition reduction; deployment claim restricted pending deployment-continuous replay

Table 5: Retrospective application of TRACE to three publicly documented evaluation incidents. Sources: Anthropic (2026b); Anthropic (2026c); OpenAI and Apollo Research (2026). The classifications instantiate the principal types in the claim typology; per-case walkthroughs are in Sections 6.1–6.4.

Risk	How it arises	TRACE mitigation	Residual limitation
Goodharting on cue markers	Pipeline tunes to satisfy known cues	Materiality threshold; cue refresh; held-out probes	Unidentified cues may remain
Overfitting to audited μ_D	Model tuned to known replay distribution	Multiple μ_D ; held-out replay; coverage reporting	Deployment heterogeneous; needs post-deployment monitoring
Verbalised recognition suppression	Training reduces “this is a test” language	Behavioural-first ordering	API-only audits without replay remain weaker
Latent signal suppression	Interpretability-detected features trained away	Versioned probes; multiple methods	No guarantee against feature obfuscation
Restricted-claim shopping	Labs reformulate claims to avoid inversion	Original-to-restricted delta reporting	Downstream readers must notice weaker claims
Disclosure dilemma	Public probes become gameable	Graduated disclosure	Transparency–evasion trade-off

Table 6: Risks of TRACE, mitigations, and residual limitations. The mitigations are layered; no single mitigation eliminates a risk on its own, and the residual limitations identify what TRACE cannot do.

8 Conclusion

Frontier model evaluation no longer occurs in a neutral measurement environment. When models condition on evaluation contexts, evaluation evidence may still be informative, but the claims drawn from it require restriction. ED names the resulting test–deployment divergence; TRACE turns it into a claim-restriction procedure.

The paper defined ED and nED, proved that marginal evaluation scores cannot identify ED, introduced a typology of ED-stable, ED-degraded, ED-inverted, and ED-undetermined claims, and specified TRACE as an audit wrapper for existing evaluation infrastructure. The retrospective applications show how public evidence already instantiates these categories: BrowseComp as ED-degraded prevalence, NLA as ED-undetermined with candidate inversion, and OpenAI / Apollo as structural ED-inversion risk.

TRACE does not produce model-level safety guarantees, eliminate adversarial adaptation, replace post-deployment monitoring, or do alignment work. Its value is narrower: it makes explicit which deployment claims remain warranted once the conditions of measurement are known, and it does so in a form that lab authors, regulators, and external auditors can use without first agreeing on the underlying mechanism. The move from capability evaluation to claim evaluation

is already visible in the public record; it now needs a shared vocabulary, a measurement target, and an audit procedure.

The framework’s full value will be visible only once it is applied prospectively. The most immediate priority is a primary empirical audit, conducted with a frontier lab against a contestable μ_D specification and reported against the minimum requirements in Section 5; adjacent priorities are extending the audit to non-Anglophone and lower-resource deployment contexts, integrating TRACE reports into system-card and conformity-assessment workflows, and developing the multi-jurisdictional claim registry proposed in Section 7.2. The labs that have surfaced BrowseComp, NLA, and the awareness-perturbation evidence have themselves produced much of what such audits require; the contribution of this paper is to specify, in measurement-theoretic terms, what those audits should report and which deployment claims their evidence does and does not warrant. The goal is not to discard frontier evaluations but to strengthen their evidential discipline by making the conditions of measurement, and the claims those conditions warrant, explicit.

Ethics Statement

This paper proposes a methodology for restricting safety claims drawn from frontier-model evaluations. We make the ethical commitments and limitations of that contribution explicit here, in keeping with the AIES expectation that methodological work make its normative commitments visible.

Stakeholders and harms. Misleading evaluation claims do not merely produce technical error; they shape regulatory approval, procurement, public trust, deployment commitments, and downstream institutional reliance. The parties who can be harmed when claims overstate their warrant include users acting on safety representations, institutions procuring systems on the basis of those representations, regulators issuing conformity decisions, and publics whose protection depends on the integrity of the safety claims attached to deployed systems. TRACE is designed to make the warrant of such claims contestable; it is not a substitute for the accountability mechanisms by which harmed parties seek redress.

Positionality and evidence base. The retrospective applications rely on publicly released frontier-laboratory disclosures (Anthropic’s BrowseComp and NLA work, the OpenAI / Apollo anti-scheming work, and AISI’s sycophancy findings). These disclosures are strategic acts by labs whose internal evidence is not fully observable, and our analysis inherits whatever selection bias the public-disclosure record carries. We treat the labs’ published evidence as the best available substrate for retrospective audit, not as a neutral or complete picture of frontier-model behaviour under evaluation.

Power asymmetries. The framework’s central operational choice—specifying μ_T and μ_D —is politically and institutionally contestable. If μ_D is defined unilaterally by the model developer, TRACE risks becoming compliance theatre: the audit produces restricted claims that authorise whatever the developer was already prepared to authorise. We have argued in Section 7.2 that credible use requires contestable μ_D specifications, disclosure of replay coverage, and participation by independent auditors or downstream institutions. Whether such participation is achieved is not a methodological question; it is a question of institutional design and political will.

Adverse uses. Three adverse uses are foreseeable. First, the framework could be selectively applied to claims that are convenient to restrict while leaving inconvenient claims unaudited; the claim-registry proposal partially addresses this by making the population of audited claims visible to downstream readers. Second, the vocabulary of restricted claims could be used to manufacture an appearance of audit rigour without the underlying behavioural replay; the reproducibility requirements (claim specification, μ_D construction, replay coverage, cue-ablation results) are intended to make such use detectable. Third, publishing the audit methodology itself is dual-use: a developer who knows which cues an audit will probe, and which evidentiary layers it consults, can train against detection rather than against

the underlying conditioning. Section 7.1 discusses our partial response to this (versioned probes, behavioural-first ordering, graduated disclosure of held-out components), but the long-run trajectory is a cat-and-mouse dynamic in which the framework must evolve alongside the systems it audits. None of these protections is complete, and we do not present TRACE as a guarantee against bad-faith adoption.

Empirical follow-up. The retrospective cases in this paper draw on the public-disclosure record only. Future empirical work should estimate ED directly under matched deployment-continuous replays, ideally with the participation of independent auditors and institutions whose deployment contexts the audit purports to represent, rather than relying on retrospective analysis of lab-released evidence.

Generative AI usage. Generative AI tools were used in a limited capacity to assist with copy-editing and LaTeX formatting of this manuscript. All research contributions—the framework, formal results, retrospective analyses, and claims—are the authors’ own, and the authors take full responsibility for the content.

Acknowledgements

The authors gratefully acknowledge the Institute for Ethics in AI and the Department of Computer Science at the University of Oxford, Trinity College Cambridge, and the UK AI Security Institute in London for their support, guidance, and networking that made this work possible. Varad Vishwarupe additionally thanks the Pivotal Research Fellowship for its support, and all of the above for the incisive critical feedback on earlier drafts.

References

- Anthropic. 2026a. Claude Opus 4.7 System Card. Transparency Hub, Anthropic.
- Anthropic. 2026b. Eval Awareness in Claude Opus 4.6’s BrowseComp Performance. Engineering blog post, Anthropic.
- Anthropic. 2026c. Natural Language Autoencoders. Research publication, Anthropic.
- Bengio, Y.; et al. 2026. International AI Safety Report 2026. Department for Science, Innovation and Technology, UK Government. DSIT 2026/001.
- Berglund, L.; Stickland, A. C.; Balesni, M.; Kaufmann, M.; Tong, M.; Korbak, T.; Kokotajlo, D.; and Evans, O. 2023. Taken Out of Context: On Measuring Situational Awareness in LLMs. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition.
- Cronbach, L. J.; and Meehl, P. E. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4): 281–302.
- Hofstätter, F.; van der Weij, T.; Teoh, J.; Djoneva, R.; Bartsch, H.; and Ward, F. R. 2025. The Elicitation Game: Evaluating Capability Elicitation Techniques. In *Proceedings of the 42nd International Conference on Machine*

Learning (ICML), volume 267 of *Proceedings of Machine Learning Research*, 23330–23356. PMLR.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 375–385.

Li, C.; Phuong, M.; and Siegel, N. Y. 2025. LLMs Can Covertly Sandbag on Capability Evaluations Against Chain-of-Thought Monitoring. In *ICML 2025 Workshop on Technical AI Governance (TAIG)*.

Manheim, D.; and Garrabrant, S. 2018. Categorizing Variants of Goodhart’s Law. *arXiv preprint arXiv:1803.04585*.

Messick, S. 1995. Validity of Psychological Assessment. *American Psychologist*, 50(9): 741–749.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 220–229.

National Institute of Standards and Technology. 2023. AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1. U.S. Department of Commerce.

OpenAI; and Apollo Research. 2026. Detecting and Reducing Scheming in AI Models. OpenAI research publication.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 33–44.

UK AI Security Institute. 2026. Ask, Don’t Tell: Reducing Sycophancy in Large Language Models. AISI Blog.

van der Weij, T.; Hofstätter, F.; Jaffe, O.; Brown, S. F.; and Ward, F. R. 2025. AI Sandbagging: Language Models Can Strategically Underperform on Evaluations. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.