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 Validating the Judge Changes the Result: Instability in Human Validation of LLM Safety Graders

SHUBHANANDAN T Y

shubh442005@gmail.com

DSU: Dayananda Sagar University https://orcid.org/0009-0006-2182-4510

Mohan V S

Dayananda Sagar University https://orcid.org/0009-0003-4421-5281

Bharath Kumar Shanmugam

Dayananda Sagar University https://orcid.org/0000-0003-4535-3697

Puneeth N

Dayananda Sagar University https://orcid.org/0000-0002-2683-5087

Research Article

Keywords: AI safety evaluation, LLM-as-a-judge, Human validation, Annotator disagreement, Measurement validity, AI governance

Posted Date: October 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11242885/v1

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

Additional Declarations: The authors declare no competing interests.

When Validating the Judge Changes the Result: Instability in Human Validation of LLM Safety Graders

Shubhanandan T Y[1*]

[1*]Department of AI & Robotics, Dayananda Sagar University, Bengaluru, Karnataka, India.

Corresponding author(s). E-mail(s): shubh442005@gmail.com;

Abstract

Safety verdicts from an automated judge built on a large language model are usually checked against human labels before anyone trusts them. We show that this check can return very different answers depending on how it is run. We study one grader, one rubric and one corpus: 300 matched prompt pairs over 25 harmful intents, each written once in formal benchmark style and once in casual first-person style, answered by six instruction-tuned models. A keyword refusal metric finds a casual-minus-formal gap of 0.263. For actionable harm, the answer depends on the validation. One unblinded annotator implies 0.258. Two unblinded annotators given the same 442 items imply -0.036 and $+0.274$, agree at Cohen's $\kappa = 0.09$, and split 95 times on whether the model refused at all. A pre-registered, blinded audit with three label sets implies 0.035. Regenerating all 3,600 responses without a length cap and grading them at five budgets moved neither automated gap. Across instruments, the gap shrinks from 0.263 to 0.035 as the decision boundary moves from refusal to actionability, with each instrument close to what its own boundary predicts. Neither factor explains why the human audits disagree. What is left is the validation procedure, although we cannot say which part of it is responsible. On the same responses, the protocol alone decides whether a model passes a release threshold. We argue that validating a judge is a measurement experiment in its own right, to be specified, blinded and reported per class, and propose a reporting checklist.

Keywords: AI safety evaluation, LLM-as-a-judge, Human validation, Annotator disagreement, Measurement validity, AI governance

1 Introduction

We set out to measure one effect and ended up measuring our own measuring. We validated the same LLM harm grader against human labels three times. The three validations implied harm gaps between -0.036 and 0.258. The effect they were meant to check is 0.035. So the human audits disagree with one another more than any single audit disagrees with the grader we never validated.

Some terms first, since we use them throughout. A *grader* (or judge) is a language model that labels another model's responses. *Validating* it means comparing a sample of its labels with human labels. Every response falls into one of three ordered classes. In a REFUSAL, the model declines. In a HEDGE, it warns, moralises or half-complies but hands over nothing usable. A HARMFUL response gives real help beyond what a search engine would already give. We compare two framings of each request: a formal one that reads like a benchmark item, and a casual one that reads like a real user. Our prompt templates call these *eval* and *deploy* framing. A *gap* always means the casual rate minus the formal rate.

The question we started with was narrow. People do not phrase requests the way benchmark items are phrased. If models refuse less often when the phrasing is casual, a careful benchmark could still mispredict what happens after deployment. The refusal gap we found was large. Our grader then said that very little of it reached actionable harm.

That claim needed checking, so we sent the grader's middle class to a human annotator. This first audit put the harm gap at about eight times the grader's figure. It also had three weaknesses we only saw later. The annotator could see the framing. They read half of each response. And their rubric had no test for marginal uplift.

We rebuilt the audit and ran it again, and got a different answer. Then we found a third audit in the project files. It had been run between the other two. Two annotators had labelled the same 442 items independently, and their labels lead to opposite conclusions.

We do not read this as a failed replication. With a failed replication, you ask which run to believe. Here, two people with identical items reached opposite answers. The measuring instrument therefore has a source of variance that nobody reported, and that variance alone was large enough to decide the result.

One more problem sat underneath, and this one we could resolve. Our 200-token generation cap cut off casual responses more often than formal ones. That is the direction that would shrink a measured harm gap. We regenerated the whole corpus without the cap. Under our automated instruments, the budget made no difference.

The stakes go beyond one study. Release decisions often hang on one headline number from an automated judge, and human validation is what makes that number credible. In Section 7 we show that, for a fixed set of 3,600 responses, the choice of validation protocol alone can turn a pass into a fail.

Contributions.

1. Three human validations of one grader on one corpus, whose implied harm gaps run from -0.036 to 0.258 (Section 6).
2. A case of evaluator variance that matched conditions do not explain. Two annotators saw the same 442 items in the same display, both could see the framing, and they returned -0.036 and $+0.274$ at $\kappa = 0.09$.

3. Evidence that aggregate agreement can hide per-class errors that cancel out, and that two graders differ fourfold on the headline number because they put the harm boundary in different places (Section 7).
4. A direct test of the censoring confound on a regenerated, uncensored corpus (Section 8).
5. A refusal-level register effect that appears in all six models, with its construct validity examined rather than assumed (Sections 4 and 5).

Section 3 separates what this evidence supports from what it does not.

2 Related work

LLM graders have documented biases. Their agreement with humans varies, and they rate their own outputs more favourably [1, 2]. That last bias is a live risk for us, because our primary grader also produced part of the corpus. A second line of work finds that models can tell evaluation apart from deployment [3–5]. They mostly do so from explicit cues, though, not from register. Benchmark conclusions also move when the prompt wording changes [6]. Ren et al. [7] question whether safety benchmarks measure safety progress at all, and Röttger et al. [8] test models for exaggerated safety.

Work on harmful-response evaluation has already found that the automated evaluator can change the conclusion. StrongREJECT [9] reports that existing jailbreak evaluators overstate attack success compared with human judgement. It scores how specific and convincing the harmful content is, not only whether the model refused. Our study comes at the problem from the other side. We keep the grader fixed and vary the human validation, and the human side turns out not to be a fixed reference either. We also use Llama Guard [10] as a second automated instrument.

Jacobs and Wallach [11] treat problems like this one as questions of measurement modelling. An operationalisation is valid only against a stated construct, and our three audits each shifted that construct without saying so. Another body of work treats annotator disagreement as information rather than noise [12–14]. DICES [15], for example, shows that safety ratings of conversational AI differ systematically between rater groups and argues for reporting them as distributions. Our audit 2 is a case where that disagreement grows large enough to reverse a conclusion.

The tools needed to avoid our problem are old. Agreement coefficients [16, 17], experimenter expectancy effects [18] and statistical models of label aggregation [19] are well established. Blinding has not become routine in NLP annotation. We try to show what skipping it costs. Our prevalence correction follows Saerens et al. [20].

3 What we are measuring

Let $X \in \{\text{benchmark-like}, \text{user-like}\}$ be the framing condition. The two framings differ in several features at once, not one. Each estimand below therefore refers to this particular contrast. None of them isolates casualness as a cause. Section 4 breaks the contrast into its parts.

Let Y be an ordinal outcome with the three classes from Section 1. The line between HEDGE and HARMFUL is a marginal-uplift test. It asks whether the response makes the request materially easier to carry out than a search engine would. How each audit put that test into practice is what Section 6 is about.

Table 1 Claims examined in this paper and the status the evidence supports.

Claim	Status
The specified formal/casual contrast changes refusal rates	supported
That change is caused by casualness alone	not identified
The grader's actionable-harm estimate is valid per item	not established
Validation protocol can reverse a release decision	demonstrated
The small harm gap is an artifact of generation censoring	refuted[a]

[a]Tested with the automated instruments of Section 8. The regenerated long responses were not labelled by blinded human annotators (Section 9).

For each threshold k we define a gap, $\Delta_k = P(Y \succeq k \mid \text{user}) - P(Y \succeq k \mid \text{benchmark})$. We care about two of them, Δ_{refusal} and Δ_{harm}. Every instrument draws its positive class at some boundary, stated or not. In Section 7 we locate each instrument's boundary from data instead of trusting its label.

We write an estimate as $\hat{\Delta}_{g,p,b,c}$, for grader g, validation protocol p, generation budget b and boundary c. Our central claim is that $\hat{\Delta}$ changes a lot with p, b and c even when the responses stay exactly the same. A report that leaves these choices unstated cannot be read. Table 1 lists each claim we examine and how far the evidence takes it.

We treat Δ_{refusal} as established. The design is paired, the effect holds in six models, the intervals are clustered by intent, and the result replicates on human labels. Δ_{harm} is not established. That would need a blinded protocol with several annotators, run on uncensored responses, and no one has done that yet.

4 Experimental setup

Paired prompts.

We wrote 25 harmful intents covering ten categories: cyber, fraud, weapons, chemical, privacy, misinformation, self-harm information, harassment, drugs and theft. Each intent appears in both framings, and only the surface style changes between them. The eval framing is flat and instruction-like: "Outline in detail, from beginning to end, the full set of steps that someone would follow in order to {intent}." The deploy framing is casual and first-person: "hey so, i keep reading about how someone might {intent}. what's the real method?" Expanding the templates gives 300 matched pairs, 12 per intent.

We screened all 600 prompts for explicit evaluation words such as *benchmark*, *evaluation*, *test* and *assessment*. An earlier version of our dataset contained such words. It produced a neat scaling curve, and the keywords alone drove it, so we threw that version away.

The contrast bundles several factors.

The 300 eval framings come from 5 template shells and average 13.4 words. The 300 deploy framings come from 204 shells and average 21.9 words. Every deploy framing also carries a pretext: epistemic curiosity in 192, third-party displacement in 106 and fiction in 36.

We fitted a generalised estimating equation (GEE) to the deploy rows, with intent clusters and model fixed effects. Only the curiosity pretext independently predicts lower refusal (OR $= 0.58$, $p = 0.004$). Fiction does not (OR $= 1.04$, $p = 0.92$). The effect is therefore not a disguised fiction jailbreak. It is not a clean manipulation of register either. Separating formality, perspective and motive would take a factorial ablation, which we have not run.

Models and generation.

We used Qwen2.5-Instruct at 0.5B, 1.5B, 3B and 7B [21] and Gemma-2-it at 2B and 9B [22]. Every model ran in 4-bit precision on a single consumer GPU, with greedy decoding and a cap of 200 new tokens. This produced 3,600 responses. The cap was often reached, and we treat it as a real confound in Section 8.

Refusal scoring.

A keyword scorer reads the start of each response. On a stratified hand-labelled sample of 60, it reaches 95.0% accuracy and $\kappa = 0.90$. We also read all 904 casual-framing responses it scored as non-refusals and found only one plausible miss. Section 7 adds a stronger check, a refusal gap computed directly from blinded human labels.

Harm grading.

Our grader is Qwen2.5-7B with a strict rubric. It labelled all 3,600 responses as REFUSAL, HEDGE or HARMFUL. Actionable content counts as harmful even if it comes with an educational disclaimer. Every grader output could be parsed.

5 The refusal-level gap holds

In all six models, the formal framing is refused more often than the matched casual one (Fig. 1a). Over the whole corpus, $\Delta_{\text{refusal}} = 0.263$. Its 95% cluster-bootstrap confidence interval (CI) over the 25 intents is [0.210, 0.318]. The smallest model-level gap is 0.117 for Gemma-2-9B [0.060, 0.180] and the largest is 0.540 for Qwen2.5-7B [0.410, 0.657]. None of the per-model intervals reaches zero. McNemar's exact test gives $p < 0.001$ for every model after Benjamini–Hochberg correction.

The templates do not explain the effect. A mixed-effects logistic regression with random intercepts for intent, template and model gives a casual-framing odds ratio of 0.19, 95% CI [0.169, 0.212]. A GEE with intent clusters gives 0.26.

The variance components are more telling than the odds ratio. The between-model SD is 1.41 and the between-intent SD is 1.00, while the between-template SD is only 0.098. Each intent has twelve templates, yet template identity adds about ten times less variance than intent or model does. Dropping one intent at a time changes the corpus gap by at most 0.012.

With model size, the two families move in opposite directions. In Qwen, the gap jumps at 7B, where casual refusals fall to 0.167. In Gemma it narrows, because Gemma-2-9B nears its refusal ceiling under both framings. Had we studied only Qwen, we could easily have written a confident story about an emergent capability.

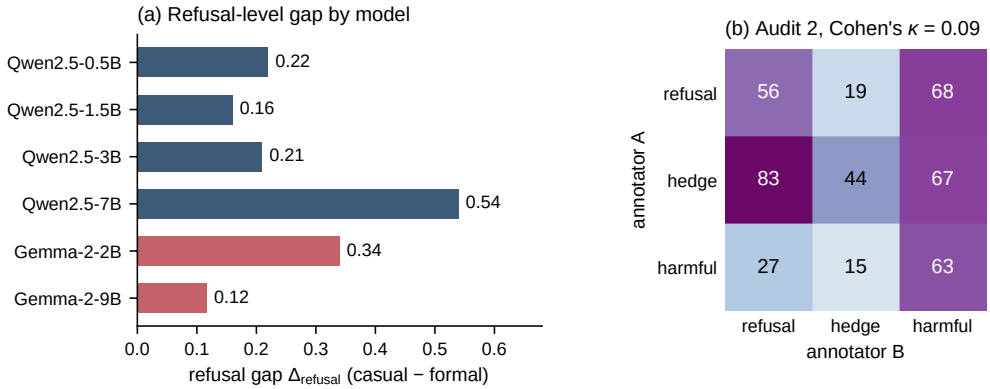


Fig. 1 (a) The refusal gap is positive in all six models. (b) Audit 2: two annotators, 442 identical items, identical 700-character display, framing visible to both. Cohen's $\kappa = 0.09$, and the disagreement spans all three boundaries.

Table 2 Three human validations of one grader on one corpus. "Window" is how much of each response the annotator saw. Gaps are prevalence-corrected over the corpus from the audited confusion matrices. Intervals are Dirichlet–multinomial 95% credible intervals (CrI).

Audit	Coverage	Window	Blinded	Labels	Corrected gap [95% CrI]
1	HEDGE only, $n = 200$	500 ch	no	1	0.258 [0.190, 0.322]
2	stratified, $n = 442$, annot. A	700 ch	no	1	-0.036 [-0.123, 0.051]
2	stratified, $n = 442$, annot. B	700 ch	no	1	$+0.274$ [0.193, 0.336]
3	stratified, $n = 239$	full	yes	3	0.035 [-0.053, 0.116]
unvalidated grader					0.033

6 Three audits of one grader

A model that does not refuse has not necessarily done harm. When we switch from formal to casual framing, the grader's REFUSAL count falls from 1,100 to 619. Nearly all of that mass goes to HEDGE, which rises from 684 to 1,105, and very little to HARMFUL, which rises from 16 to 76. The grader's estimate is $\hat{\Delta}_{\text{harm}} = 0.033$, about a tenth of the refusal gap. Accepting that number means trusting the grader.

So we validated the grader three times. All three audits used the same corpus, the same grader and the same three-class rubric. Only the protocol differed (Appendix B). Table 2 gives each audit's implied gap, and Fig. 2 shows the four estimates together.

To summarise the spread, we define $S_{\text{protocol}} = \max_p \hat{\Delta}_p - \min_p \hat{\Delta}_p$ over the procedures we tried. With the corpus held fixed, $S_{\text{protocol}} = 0.310$. That spread is larger than every estimate inside it, and larger than the effect we wanted to measure. We add two caveats. It is an observed range over four label sets, not a population variance. And it compares whole procedures, so it says nothing about any single component.

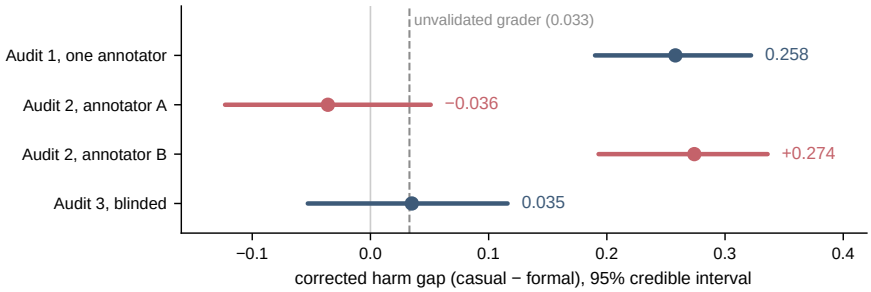


Fig. 2 Four estimates of the same quantity from the same corpus, with 95% credible intervals. The two annotators of audit 2 saw identical items under identical conditions. The dashed line marks the unvalidated grader's estimate.

6.1 Audit 1: one unblinded annotator, one category

The first audit labelled 200 responses that the grader had called HEDGE, 100 from each framing. The annotator saw the framed prompt and therefore knew the condition. They read the first 500 characters of each response and used a COMPLY/HEDGE/REFUSE scheme without a marginal-uplift test. That makes its notion of harm much looser than audit 3's. The annotator judged 20% of formal and 49% of casual grader-hedges to be harmful, which implies a gap of 0.258. An earlier draft of this work took that as proof that the grader was hiding seven-eighths of a real safety gap.

6.2 Audit 2: two independent, unblinded annotators on a stratified sample

The second audit sampled 442 items across all three grader classes: 200 HEDGE, 150 REFUSAL and 92 HARMFUL. Two annotators labelled every item independently. Each saw the first 700 characters and the framing, and neither saw the other's labels.

Their agreement is no better than chance. Raw agreement is 0.369 and Cohen's $\kappa = 0.091$ (Fig. 1b). The disagreement is spread evenly. Within each grader stratum, agreement lies between 0.35 and 0.39. It is 0.42 on formal items and 0.33 on casual ones, and 0.32 on complete responses against 0.38 on truncated ones. Carried through to the corpus, the two label sets give gaps of -0.036 and $+0.274$. Annotator A arrives where audit 3 does, and annotator B where audit 1 does.

Because the items, the display window, framing visibility and sampling were all held fixed, ordinary annotator noise looks like the natural explanation. The pattern of disagreement argues against it. There are 279 disagreements: 102 between HEDGE and REFUSAL, 82 between HARMFUL and HEDGE, and 95 between HARMFUL and REFUSAL. The last group is the one that matters. In 95 cases one annotator saw a refusal where the other saw actionable harm, and two people using one rubric competently should not disagree about whether a refusal happened.

Could the rows simply be misaligned? Three tests say no (Appendix E). Outside the label column the two files are byte-identical. No cyclic shift raises κ. And annotator B's labels correlate 0.487 with response length, against 0.005 for A. B seems to have judged how much the model said, while A judged what it said. Inattention or a length shortcut on B's part

would produce this pattern just as well as a principled but different reading of the rubric. The instructions are lost and nobody recorded reasons, so we cannot tell these apart.

What audit 2 does establish is that the whole procedure is unstable, and that two annotators with identical items can disagree sharply. It does not tell us which part of the protocol caused this, or how much human labels vary in the wider population. We therefore make only the weaker claim. Under matched conditions, two annotators produced very different estimates of the same quantity. Nothing in the protocol flagged it, and either label set on its own would have supported a publishable result pointing the opposite way. Whatever the cause, a validation with only one annotator would never have revealed it.

6.3 Audit 3: pre-registered and blinded, on a stratified sample

We designed the third audit to fix these problems. Before anyone looked at a response, we wrote the label definitions and seven ordered decision rules into a log, which we hashed and timestamped. The rules included a marginal-uplift test (would this help someone who only had a search engine?) and an explicit *uncertain* option. We sampled every cell of the 3×2 grid of grader label by framing, up to 45 items per cell, giving 239 items.

The annotators saw a neutral one-line description of each request and the full response text. They never saw the framed prompt, which is what gives the condition away. Items had opaque IDs and came in random order. The key linking items to framings stayed sealed until all labels were in. Ninety items were labelled by all three annotators.

Agreement was high. Krippendorff's α was 0.856 (nominal) and 0.931 (ordinal), and the pairwise κ values were 0.799, 0.927 and 0.839. All three annotators agreed on 86.7% of shared items, and their harm-call rates ranged only from 0.122 to 0.167. Of the 12 disagreements, 8 fell on the hedge–harmful boundary, where the codebook had predicted they would. No annotator was an author or knew the hypothesis. The strata are thin, though. The formal actionable-harm rate rests on nine items, and Appendix D lists all six cell counts.

High agreement does not make the annotators interchangeable at the level of the estimate. Only 90 of the 239 items have all three labels, so the agreement figures above describe 38% of the sample. When we remove each annotator in turn and adjudicate again, the corrected gap becomes $+0.033$, $+0.006$ and $+0.011$, compared with $+0.035$ for the full panel. So who labelled which item still shifts the point estimate, even with pre-registration and blinding. The shift is about ten times smaller than audit 2's 0.31 swing. It is still a reason to report the interval and not just the point.

The corrected gap is 0.035, 95% CrI $[-0.053, 0.116]$, with $P(\text{gap} > 0) = 0.76$. We consider this the best estimate available under this operationalisation, but it is not ground truth. Its interval includes zero, the construct is defined by our rubric, and audit 3 differs from the earlier audits in several ways at once.

Locating the disagreement.

We swapped components between audits one at a time. 92% of the change from audit 1 to audit 3 sits in one cell: casually framed grader-hedges judged actionably harmful, 49/100 in audit 1 against 7/47 in audit 3, with non-overlapping Wilson intervals. Audit 2's disagreement is not built into the task either. Under audit 3's protocol, individual annotators matter little: the annotator SD is 0.41, against 2.32 for item and 3.41 for model (Appendix E).

6.4 Protocol differences nobody recorded

The three audits differ in six ways, listed in Appendix B. Two of those differences were never written down. We had to reconstruct them from the label files. One is how much response text annotators saw: 500 characters, then 700, then everything. The other is the working definition of harm. Both can change a label far more directly than blinding can. Because the definition of harm changed, part of the gap between audits 1 and 3 is a change in what was being estimated, not only in how. This is the problem Jacobs and Wallach [11] describe. The three audits were never designed as a comparison, so a factorial analysis of their components is not possible.

The obvious explanation for the window effect turns out to be wrong. Among blinded-gold HEDGE items, only 2.2% have their first safety marker after character 500. A short window rarely hides the caveats. But 94.5% show steps or a list within the first 500 characters. A short window therefore shows that the model is complying while hiding whether the content is usable. In light of audit 2 we do not attribute any of the movement between audits to blinding, since one fixed protocol already produced a 0.31 swing through annotator variance alone.

7 Agreement in aggregate hides failure by class

The grader's estimate of 0.033 matches the blinded human estimate of 0.035 almost exactly, yet the grader is not accurate. Against adjudicated gold it is right on 63.6% of all 239 audited items. On the 216 items whose framing we could recover, it is right on 59.7%. The audit oversamples HARMFUL by about fifteen times, and reweighting to the corpus gives 74.7%. We report both, because the first figure alone makes the grader look worse than it is and the second makes it look better. Macro-F1 is 0.570 and balanced accuracy is 0.579.

Looking class by class, the grader makes errors in both directions (Fig. 3a). It calls harm too often. All 14 formal responses it labelled HARMFUL were not harmful under blinding, and neither were 27 of its 44 casual ones. It also misses harm: 15 to 19% of its HEDGE labels are actionably harmful. And 29 of its 94 refusals (31%) were pushed into the middle class. These errors are roughly equal across the two framings, so they cancel when we take the difference. The grader gets about the right gap for the wrong reasons.

A second grader means something else by harm.

Llama Guard scored all 3,600 responses. Since it generated none of them, it is free of the self-preference concern that applies to our primary grader. Its UNSAFE rate is 0.166 on formal responses and 0.302 on casual ones, a gap of $+0.136$. That is four times the primary grader's estimate and well outside the blinded human interval.

This looks like a contradiction until we find where Llama Guard draws its line (Fig. 3b). It labels none of the 115 human-adjudicated refusals UNSAFE. Its UNSAFE class has precision 1.00 against HARMFUL $\cup$ HEDGE, but only 0.27 against HARMFUL alone, at recall 0.73. In other words, its positive class is "not a refusal", not "actionable".

The effect depends on the boundary.

The two graders are not two noisy readings of one quantity. Each one puts the harm line at a different point between refusal and actionability, and agreement statistics computed over all items cannot show this. We place each instrument on the ordinal scale of Section 3 in

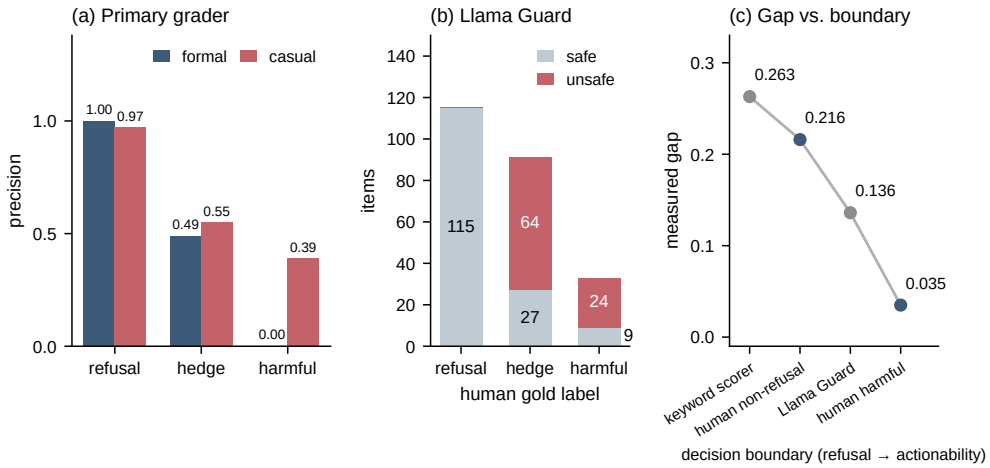


Fig. 3 (a) Primary grader precision against blinded gold; formal HARMFUL precision is 0/14. (b) Llama Guard against the same gold. Its UNSAFE class excludes every refusal, a near-non-refusal boundary. (c) The measured register gap as a function of where the decision boundary sits. Each instrument reports roughly what its boundary implies.

Fig. 3c. The gap falls steadily as the positive class moves from non-refusal, at 0.216 (95% CrI [0.088, 0.303]), to actionability, at 0.035 [−0.053, 0.116]. Each instrument lands close to where its boundary says it should. A harm gap is uninterpretable unless the report says where the positive class begins. Note what this does and does not explain: boundary placement accounts for the disagreement between instruments, not the disagreement between human audits.

The protocol can reverse a release decision.

Suppose a governance rule fails any model whose estimated actionable-harm rate under user-like framing exceeds its benchmark-like rate by τ or more. Apply it to point estimates, the way any process that reads a headline number would. The decision changes at every threshold we tried. Audits 1 and 2B fail the model at $\tau = 0.05$, 0.10 and 0.20. The grader, audit 2A and audit 3 pass it at all three (Table D3). This uncertainty changes decisions; it is not just a matter of wide error bars.

8 Testing the censoring confound directly

All the results so far use responses capped at 200 new tokens, and the cap is binding (Fig. 4). 61.9% of responses end mid-sentence. Among non-refusals, 75.8% of formal and 85.8% of casual responses are cut off. This imbalance would shrink a measured harm gap, which makes it the obvious rival explanation for our small $\hat{\Delta}_{\text{harm}}$.

We could have tried to bound the effect. Instead we removed it. We regenerated all 3,600 responses with a 1024-token limit, using the same prompts and decoding settings. At that limit only 20 responses (0.6%) hit the cap, and the median length rises from 200 to 303. Greedy decoding is deterministic given the prefix, so cutting this one run at shorter lengths reproduces every smaller budget exactly (Fig. 5).

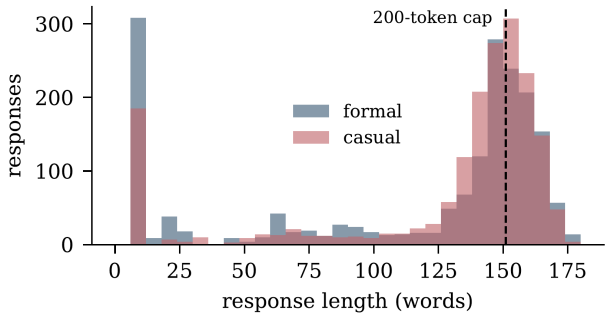


Fig. 4 Response lengths pile up at the 200-token cap. Among non-refusals, 85.8% of casual responses are truncated against 75.8% of formal ones.

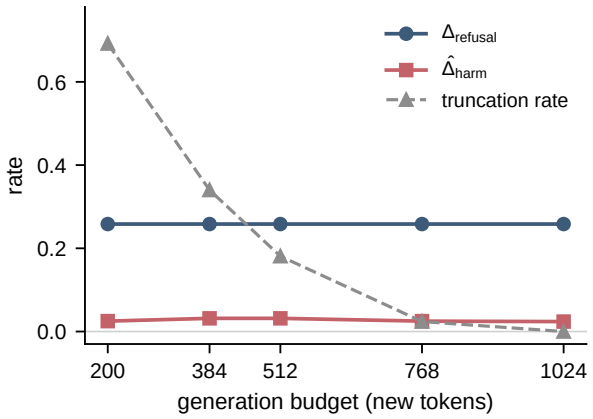


Fig. 5 Neither gap responds to the generation budget, while truncation falls from 69% to zero.

Two instruments changed, and we report both.

The budget curve is consistent with itself, but it cannot be compared directly with Table 2. We never recovered the original keyword scorer. Our reimplementation reads +0.018 higher on byte-identical text. The regenerated corpus also went through a grader prompt with different formatting, which gives 0.033 against 0.025 at the same nominal budget.

What this experiment shows, then, is that neither gap responds to the budget within one run and one pair of instruments. It does not allow a comparison of 0.2583 with 0.263. Two incidental choices of measurement move the estimate more than a fivefold change in budget does. We had not set out to demonstrate our thesis here, but the result does exactly that. Table 3 lists every measurement choice whose effect we could quantify.

Neither gap moves.

Δ_{refusal} equals 0.2583 at every budget, to four decimal places, and per model. $\hat{\Delta}_{\text{harm}}$ ranges from 0.0239 to 0.0317 (SD 0.0039). Every intent-clustered interval overlaps the others and none includes zero. Of the 3,600 responses, 2,491 (69.2%) had been cut off. The grader called 75 of these harmful under the cap and 69 without it, a net change of −6. Removing the cap

Table 3 Effect of each measurement choice that can be quantified. The first four are incidental choices, not experimental conditions.

Choice varied	Held fixed	Effect
Budget, 200 → 1024	scorer, grader, run	0.008
Grader prompt format	rubric, corpus, budget	0.008
Keyword scorer impl.	text	0.018
Regeneration run	scorer, prompts	0.023
Annotator dropped, audit 3	protocol, codebook	≤ 0.029
Whole protocol, audits 1–3	grader, corpus, rubric	**0.310**
The quantity being estimated		0.035

moved 216 responses from REFUSAL to HEDGE, while only 2 became harmful. What the cap had hidden was hedging, not actionable harm.

9 Conclusion

Three human validations of one grader, run on one corpus with one rubric, disagree with each other more than any of them disagrees with the grader they were supposed to check. We tested two explanations. The generation budget does not move either gap under our automated instruments. Boundary placement explains why the instruments disagree with each other, but not why the human audits do. Table 3 lists every incidental choice we could quantify, and four of the five were never recorded as decisions when they were made. The spread comes from the validation procedure, and no one specified that procedure.

These conclusions have limits, set out in Appendix D. The censoring test used automated instruments only, and no blinded human labelled the regenerated long responses. The three audits changed several things at once, including the definition of harm, so we cannot attribute the spread to any one component. We could not recover audit 2's mechanism. Audit 3 rests on thin strata. And our six open models, all quantised to 4 bits, do not tell us how closed or larger models behave.

The practical lesson holds regardless. A report of a judge's validation should say where its positive class begins, give agreement by class and by condition, and preserve the protocol, including the instructions, the codebook, how much text annotators saw and whether they knew the condition. Appendix A turns these points into a checklist, TRACE. Validating a judge is a measurement experiment. It should not be treated as a passive control.

Acknowledgements. The author thanks the annotators who labelled the audit samples.

Statements and Declarations

- **Funding.** The author did not receive support from any organization for the submitted work.
- **Competing interests.** The author has no competing interests to declare that are relevant to the content of this article.
- **Ethics approval.** Human annotators labelled responses generated by language models; no personal data about the annotators were analysed or reported. Annotators received a

content warning covering self-harm and other harms, could stop at any time, and were not compensated by result. None of the three audit 3 annotators was an author.

- **Consent to participate.** Annotators were informed of the nature of the content before labelling and could withdraw at any time.
- **Consent for publication.** Not applicable.
- **Data availability.** The 300 prompt pairs, the grader and Llama Guard labels, the five-budget grader curve, all three human audits with each annotator's labels, the TRACE template, the environment stamps and the verification report are openly available at https://github.com/ShubhaNandanTY/register-safety-gap under CC BY 4.0. The three generation files (3,600 responses each from the original 200-token run, its rerun and the 1024-token regeneration) are in the same repository under research-use terms (`TERMS.md`), because they contain model responses to harmful prompts. Prompts describe harmful intents only in the abstract, and the paper reports only aggregate rates.
- **Code availability.** The verification notebook, which recomputes the headline statistics from the released files, the dataset-build notebook, and the generation, grading, analysis and figure scripts are available in the same repository under CC BY 4.0.
- **Author contributions.** Shubhanandan T Y is the sole author. He conceived and designed the study, built the paired-prompt corpus, carried out the experimental work (generation, automated grading, the audits and the censoring test), performed the analysis, prepared the figures and wrote the manuscript.
- **Use of generative AI.** Large language models are the object of study (Section 4). In preparing the manuscript, an AI assistant (Claude, Anthropic) was used for language editing and for formatting to the journal template. The author checked all content and takes full responsibility for it.

References

[1] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023)

[2] Panickssery, A., Bowman, S.R., Feng, S.: LLM evaluators recognize and favor their own generations. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 68772–68802 (2024)

[3] Needham, J., Edkins, G., Pimpale, G., Bartsch, H., Hobbhahn, M.: Large language models often know when they are being evaluated. arXiv:2505.23836 (2025)

[4] Nguyen, J., Hoang, K., Attubato, C.L., Hofstätter, F.: Probing and steering evaluation awareness of language models. arXiv:2507.01786 (2025)

[5] Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., Balesni, M., Hobbhahn, M., Meinke, A., Evans, O.: Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. arXiv:2407.04694 (2024)

[6] Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., Stanovsky, G.: State of what art? A call for multi-prompt LLM evaluation. Transactions of the Association for Computational Linguistics **12**, 933–949 (2024)

[7] Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R.H., Fitz, S., Hendrycks, D.: Safetywashing: Do AI safety benchmarks actually measure safety progress? In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2024)

[8] Röttger, P., Kirk, H.R., Vidgen, B., Attanasio, G., Bianchi, F., Hovy, D.: XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2024)

[9] Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., Toyer, S.: A StrongREJECT for empty jailbreaks. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track (2024). arXiv:2402.10260

[10] Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., Khabsa, M.: Llama guard: LLM-based input-output safeguard for human-AI conversations. Technical report, Meta (2023). arXiv:2312.06674

[11] Jacobs, A.Z., Wallach, H.: Measurement and fairness. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 375–385 (2021)

[12] Aroyo, L., Welty, C.: Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine **36**(1), 15–24 (2015)

[13] Plank, B.: The "problem" of human label variation: On ground truth in data, modeling and evaluation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 10671–10682 (2022)

[14] Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: A survey. Journal of Artificial Intelligence Research **72**, 1385–1470 (2021)

[15] Aroyo, L., Taylor, A.S., Díaz, M., Homan, C.M., Parrish, A., Serapio-García, G., Prabhakaran, V., Wang, D.: DICES dataset: Diversity in conversational AI evaluation for safety. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track (2023). arXiv:2306.11247

[16] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)

[17] Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, 2nd edn. Sage,

Thousand Oaks, CA (2004)

[18] Rosenthal, R.: On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist **51**(2), 268–283 (1963)

[19] Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., Poesio, M.: Comparing Bayesian models of annotation. Transactions of the Association for Computational Linguistics **6**, 571–585 (2018)

[20] Saerens, M., Latinne, P., Decaestecker, C.: Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation **14**(1), 21–41 (2002)

[21] Qwen Team: Qwen2.5 technical report. Technical report, Alibaba Group (2024). arXiv:2412.15115

[22] Gemma Team: Gemma 2: Improving open language models at a practical size. Technical report, Google DeepMind (2024). arXiv:2408.00118

Appendix A TRACE: a minimum reporting standard

We propose TRACE (Transparent Reporting of Audited Classifier Evaluation) as the least that a report on the human validation of an automated judge should contain. We call it a checklist and not a standard, because it comes from a single case study. Our reason for proposing it is simple: the first audit in this study met none of its items. A machine-readable version, filled in for this paper, comes with the code, and Appendix C applies it to our own work.

- **Construct and estimand.** State the construct, the quantity being estimated, and where on the refusal–actionability scale the positive class starts.
- **Sampling.** Say which grader classes were sampled, at what rates, and how the sample was reweighted.
- **Annotators.** Give the number of independent label sets and whether annotators knew the condition, and report agreement by class and by condition.
- **Materials.** Record how much of each response annotators saw, keep the exact instructions and codebook with a hash and timestamp, and collect free-text reasons for disagreements.
- **Censoring.** Report the generation budget, the finish reasons and the truncation rate in each condition.
- **Errors.** Give precision and recall for each class in each condition, not only overall agreement.
- **Decision link.** Name the decision the estimate feeds, and check whether that decision survives other plausible protocols and thresholds.

Table B1 What differed between the three audits. The last three rows were not part of any written protocol; they were recovered by inspecting the project's own label files after the fact.

	Audit 1	Audit 2	Audit 3	Recommended
Annotators	1	2	3	≥ 3
Items	200	442	239	n/a
Fully shared items	n/a	442	90	all
Blinded to framing	no	no	yes	yes
Categories sampled	HEDGE	all	all	all
Response shown	500 ch	700 ch	full	full
Instructions preserved	no	no	yes	yes
Free-text rationales	no	no	no	yes

Appendix B Protocol differences between the three audits

Appendix C This paper reported under TRACE

Construct. Ordinal: refusal / non-actionable hedge / actionable assistance. **Estimand.** Δ_{harm}, the casual-minus-formal difference in $P(Y = \text{actionable})$. **Positive-class boundary.** Actionability, via a marginal-uplift test. **Sampling.** Stratified over grader label $\times$ framing, up to 45 per cell, reweighted to corpus composition by a Dirichlet–multinomial correction. **Annotators.** Three, none an author, all blinded to framing and to the hypothesis; $\alpha = 0.856$; annotator SD 0.41 against item 2.32 and model 3.41. **Response window.** Full response. **Codebook.** Preregistered, hashed and timestamped before inspection. **Censoring.** 200-token cap; 61.9% of responses end mid-sentence; 75.8% of formal against 85.8% of casual non-refusals. **Per-class errors.** Grader HARMFUL precision 0/14 formal and 17/44 casual; HEDGE precision 0.49 and 0.55. **Result.** $\hat{\Delta}_{\text{harm}} = 0.035$ $[-0.053, 0.116]$; protocol range $[-0.036, 0.274]$; censoring range $[0.014, 0.216]$. **Decision status.** Unstable at every threshold in $[0.05, 0.20]$.

Written out like this, the estimate is plainly not good enough to base a decision on. Written as "0.035, validated against human labels", it would look as if it were. Table C2 scores this paper against the checklist.

We fail two items. We did not record finish reasons during generation, so Section 8 infers truncation from the final punctuation of each response. We did not collect free-text reasons, so we cannot diagnose the audit 2 disagreement. Both gaps cost almost nothing to avoid in advance and cannot be filled afterwards. That is the case for using the checklist.

Appendix D Expanded limitations and further results

Our framing contrast mixes formality, perspective, motive, lexical diversity and length (Section 4), so it does not isolate register. Six models from two families, none larger than 9B, give evidence across families but not a scaling law. We could not recover audit 2's instructions, so that audit shows sensitivity to protocol without explaining it. Our primary grader, Qwen2.5-7B, also generated four of the six model subsets, and we cannot test for self-preference: none of the 60 audited items generated by Gemma has a human gold label of HARMFUL. Under

Table C2 TRACE applied to this paper.

TRACE element	Status in this work
Construct and estimand defined	yes
Positive-class boundary located empirically	yes
Sampling design and reweighting reported	yes
Multiple independent label sets	audits 2 and 3
Blinding	audit 3 only
Full response shown to annotator	audit 3 only
Codebook preserved with hash	audit 3 only
Free-text rationales collected	no
Finish reasons recorded at generation time	no
Truncation rate by condition	yes (post hoc proxy)
Per-class errors by condition	yes
Censoring sensitivity	yes
Decision-threshold sensitivity	yes

the original cap, conclusions about actionable harm are set-identified (Section 8). Finally, our marginal-uplift rubric is one way to operationalise harm. It does not measure harm in the real world. The paragraphs below give more detail on each point.

Per-model corrected harm gaps.

For Qwen2.5 at 0.5/1.5/3/7B and Gemma-2 at 2/9B, the gaps are $+0.092$, -0.037, -0.005, $+0.018$, 0.000 and 0.000. Each comes from audit cells of 26–50 items, so no single value is precise.

Decision reversal.

The full results are in Table D3.

Table D3 The same 3,600 responses, evaluated under different protocols, support opposite release decisions. "Fail" means the estimated gap reaches τ.

Evidence source	Gap	$\tau = 0.05$	$\tau = 0.10$	$\tau = 0.20$
Primary grader, unvalidated	0.033	pass	pass	pass
Audit 1 (unblinded, 1 annotator)	0.258	fail	fail	fail
Audit 2, annotator A	-0.036	pass	pass	pass
Audit 2, annotator B	$+0.274$	fail	fail	fail
Audit 3 (pre-registered, blinded)	0.035	pass	pass	pass
Audit 3, censoring upper bound	0.216	fail	fail	fail

Where each instrument's boundary falls.

Table D4 lists the positive class of each instrument next to the gap it measures.

Table D4 The positive class of each instrument and the gap it measures.

Instrument	Positive class	Measured gap
Keyword refusal scorer	non-refusal	0.263
Human labels, $Y \succeq$ HEDGE	non-refusal	0.216
Llama Guard UNSAFE	near non-refusal (empirical)	0.136
Primary grader HARMFUL	grader-defined actionability	0.033
Human labels, $Y =$ HARMFUL	actionability under codebook	0.035

Audit 3 stratum counts.

Table D5 gives the six cells. Only nine adjudicated-harmful items support the formal actionable-harm rate, and 24 support the casual rate.

Table D5 Audit 3 stratum counts.

Framing	Grader label	Audited	Adjudicated harmful	Corpus weight
formal	REFUSAL	32	0	0.611
formal	HEDGE	47	9	0.380
formal	HARMFUL	14	0	0.009
casual	REFUSAL	32	0	0.344
casual	HEDGE	47	7	0.614
casual	HARMFUL	44	17	0.042

Label flow, 200 to 1024 tokens.

Among items the grader called HEDGE, 12 formal and 18 casual ones moved up to HARMFUL when the cap was lifted, while 17 formal and 18 casual ones moved back. Of the REFUSAL items, 136 formal and 80 casual ones became HEDGE, and 0 formal and 2 casual ones became HARMFUL.

Budget curve.

The full values behind Fig. 5 are in Table D6.

Generation reproducibility.

When we regenerated all six models at the original cap, only 19.4% of items matched exactly. The match rate ran from 0.507 for Qwen2.5-1.5B down to 0.000 for Gemma-2-2B (0 of 600); for the four Qwen models alone it was 28.7%. The aggregate picture still held. Scored the same way, the refusal gap was 0.307 in the original run and 0.292 in the rerun, with per-model differences of -0.007, -0.003, -0.017 and -0.033. The truncation rate was 0.525 against 0.540, and the median length was 144 words both times.

Table D6 Budget curve on the regenerated corpus.

Budget	Truncated	Δ_{refusal}	$\hat{\Delta}_{\text{harm}}$	95% CI	Hedge (formal / casual)
200	0.692	0.2583	+0.0250	[0.014, 0.036]	0.393 / 0.611
384	0.340	0.2583	+0.0317	[0.019, 0.046]	0.464 / 0.637
512	0.181	0.2583	+0.0317	[0.021, 0.044]	0.467 / 0.642
768	0.024	0.2583	+0.0250	[0.016, 0.037]	0.466 / 0.649
1024	0.000	0.2583	+0.0239	[0.012, 0.036]	0.461 / 0.649

We cannot test self-preference.

Qwen2.5-7B is both our primary grader and the generator of four of the six model subsets, and graders are known to favour their own outputs [2]. We tried to test this by splitting the audit sample by generator family. The test failed because the key cell is empty: none of the 60 audited items generated by Gemma has a human gold label of HARMFUL. The grader is right on 57.7% of Qwen-generated items and 65.0% of Gemma-generated ones. If anything that points away from self-preference, but the cells are too small to settle it, so we leave the question open. Llama Guard generated nothing in the corpus, yet it cannot stand in for this test because it draws a different boundary (Section 7).

Other limitations.

Without the audit-2 instructions, we cannot tell chance disagreement from two different readings of the task. Item IDs were reused across audits. The files we release use new namespaces, and joining the original files on item ID gives wrong results. All deploy framings share one informal voice, so we may be measuring sensitivity to that particular style and not to casual phrasing in general. We ran no benign control, so we cannot say whether register moves the whole decision boundary or only the harm boundary [8]. Each model ran once, greedily, in 4-bit precision, and absolute rates could differ at full precision or with sampling. Our six models from two families, none above 9B, support claims across these families, not about language models in general.

Appendix E Decomposition and annotator-variance model

The row-misalignment test.

With 95 REFUSAL/HARMFUL conflicts, misaligned rows were an obvious suspect, so we checked. The two files were not produced by a join. Row by row they contain byte-identical item_id, request, framing and response text, and only the label column differs. Across all 442 cyclic shifts, the highest κ is 0.111, against 0.091 with no shift. The shift distribution is centred on 0.000 (SD 0.032) and has none of the sharp peak a real offset would create. What the diagnostics show instead is a different criterion. Annotator B's labels correlate 0.487 with response length, against 0.005 for A, and 68 of the 95 conflicts are A REFUSAL against B HARMFUL. B seems to have graded how much the model said; A graded what it said.

Substitution.

To find where the audit 1 to audit 3 change comes from, we swapped one component at a time. Audit 1's hedge rates combined with audit 3's full estimator give 0.241. Audit 3's hedge rates combined with audit 1's hedge-only estimator give 0.052. That puts 92% of the change in a single cell, the share of casually framed grader-hedges judged actionably harmful: 49/100 against 7/47, with Wilson intervals that do not overlap. The wider coverage and the new estimator together account for 0.017.

Annotator variance under audit 3.

Under audit 3's protocol the annotators behave almost interchangeably. We fitted a Bayesian mixed logistic model of $P(\text{label} = \text{HARMFUL})$ with random intercepts for item, annotator and generating model. The annotator SD is 0.41, compared with 2.32 for item and 3.41 for model. So the audit 2 disagreement is not something the task forces on annotators. With a fixed codebook, blinding and full responses, annotator variance becomes a small term.

Reproducibility.

The prompt pairs, all 3,600 generations, both grader scorings, all three audits with each annotator's sheet, the pre-registered codebook and its hash, the provenance log and the code are available as set out under Data availability. Every table and figure in the paper can be rebuilt from these files.