A Formal Statistical Measurement Instrument for the Safety & Harmful Content Dimension in AI Red-Teaming

Part I of II — Definition and Mathematical Foundations

Juan F. Cobo *Author*[1] and Leonel Leenen *Reviewer*[1]

[1]EthiCompass

Manuscript prepared for external review and eventual submission
September 4, 2026

Abstract

We present the formal definition of a statistical instrument for converting the continuous output of a harmful-content judge (the "Safety & Harmful Content" dimension of an AI red-teaming engine) into an auditable violation rate, complete with a confidence interval and a sensitivity/specificity correction. The instrument combines five harm categories (hate, violence, self-harm, sexual content, dangerous instructions) through a per-trace maximum collapse over turns, an OR triggering rule across categories, per-category thresholds selected via the Youden index, and a Rogan–Gladen prevalence correction with delta-method variance, together with a Wilson-type confidence interval for the observed rate. We give the complete mathematical definitions, the minimal worked numerical examples needed to follow each formula by hand, and the justification for two central design decisions: why the trace — not the turn — is the correct sampling unit, and why the instrument is, by construction, agnostic to the nature (lexical, semantic, LLM-as-judge, or composite) of the judge that produces the input score. This document reports no calibration of free parameters (τ_c, sensitivity, specificity) and no empirical result: both are treated in full in the companion paper of this series.

Keywords: AI red-teaming, metric definition, Youden index, Rogan–Gladen correction, Wilson interval, content safety.

1 Introduction and Motivation

AI red-teaming systems need to convert an LLM's judgment of whether a conversational turn contains harm ("hate," "violence," etc.) into an auditable figure: a violation rate, with its uncertainty, that a client or a regulator can interpret and reason about longitudinally (did it improve or worsen between two versions of the evaluated system?). The problem is non-trivial for two reasons.

First, the judge that produces the continuous per-turn score is imperfect: it has its own sensitivity and specificity, which depend on the instrument used and on the population of text to which it is applied. Treating its output as literal truth, without correction, biases any reported rate. Second, a dimension such as Safety is not a single question but five simultaneous harm categories, evaluated turn by turn within a conversational trace — an explicit rule is needed to collapse that matrix

into a single per-trace decision, and that rule has non-trivial statistical consequences (discussed in detail, with data, in the companion paper).

This document consolidates the *definitional* work for such an instrument for the Safety dimension: what it measures, how it aggregates, what form its confidence interval and prevalence correction take, and why the design decisions taken were taken. It does not address whether the calibrated instrument performs well or poorly in practice — that is an empirical question, treated in the companion paper ("Calibration and Empirical Validation...").

2 Related Work

The architecture combines classical results from diagnostic statistics with the state of the evidence on the calibration of harmful-content classifiers:

- **Youden index (Youden, 1950).** Standard in clinical diagnostics for choosing the cut-point of a continuous test that maximizes sensitivity + specificity − 1 over an ROC curve. It does not require the score to be a calibrated probability, only that it rank reasonably well.
- **Rogan–Gladen correction (Rogan & Gladen, 1978).** A classical epidemiological formula for correcting an observed prevalence by the known sensitivity and specificity of the test used, when the test is imperfect. This is exactly the problem of an imperfect judge measuring a violation rate.
- **Wilson-type interval (Wilson, 1927).** A confidence interval for a binomial proportion with better behavior than the naive normal interval at the extremes (low rates), which is the typical regime of a violation rate.
- **HateCheck (Röttger et al., 2021).** A functional test suite for hate-speech classifiers, deliberately built to expose a classifier's blind spots (negation, counter-speech, slur reclamation, orthographic obfuscation). It exemplifies the "functional test suite, designed to calibrate" paradigm that informs this instrument's design; it is used as the real-data validation source in the companion paper.
- **Evidence that no universal threshold exists.** Our own literature review found that neither the academic literature (Jigsaw/Google, on the Perspective API) nor industrial practice (the OpenAI Moderation API, which does not publish its thresholds; MLCommons AILuminate, which grades by relative comparison against reference models rather than an absolute threshold) supports the claim that a harmful-content detection threshold can be derived from the literature without one's own empirical data. This finding is the defining reason why τ_c is defined in this document as a **free parameter, to be determined empirically on a golden set** (§3.6), rather than as a constant declared by design or by fiat.

3 Formal Definition of the Metric

3.1 The Score Matrix

Let $i \in \{1, \ldots, N\}$ index the trace (a complete conversation), $t \in \{1, \ldots, T\}$ index the turn within the trace, and $c \in \{1, \ldots, K\}$ index the harm category, with $K = 5$: {hate, violence, self-harm, sexual, dangerous-instructions}. The judge produces, for every combination, a continuous score

$$s_{i,t,c} \in [0, 1],$$

interpreted as the degree to which turn t of trace i violates category c.

Layer independence. Everything that follows in this document — the maximum collapse, the OR rule, the Youden sweep, the Wilson interval, the Rogan–Gladen correction — operates exclusively on $s_{i,t,c}$ as a number in $[0, 1]$. This instrument does not know, and cannot know, whether that number came from a lexical classifier, a semantic model, an LLM-as-judge, or a composite grader — that decision is resolved in an earlier layer (the judge that generates the score) and reaches it already resolved. This is not a simplification adopted for convenience: it reflects a deliberate architectural separation between a *measurement* layer — which emits only a continuous score per turn and variable, together with its rationale, and never a threshold, a rate, or a severity — and an *aggregation-and-interpretation* layer — which consumes those numbers without needing visibility into how they were computed — under the general principle that a number that measures belongs to the former, and a meaning that rests on many numbers belongs to the latter. This is why, when the companion paper evaluates a specific lexical instrument, the label "lexical" is a property of *that particular* judge, relevant for diagnosing why its J is low — not a condition that the calibration method requires or distinguishes. The same mathematics, without a single formula changed, would apply equally if $s_{i,t,c}$ came from a semantic judge or a composite grader.

3.2 Turn-to-Trace Collapse

Turns are collapsed by taking the maximum — a trace violates category c as much as its most violating turn, not an average (a single severely harmful turn should not be diluted among clean turns):

$$m_{i,c} = \max_{t=1,\ldots,T} s_{i,t,c}.$$

3.3 Per-Category Binarization and the Composite Rule

Each category has its own threshold τ_c (a free parameter; see §3.6 for the selection criterion, and the companion paper for its empirical calibration), calibrated separately:

$$b_{i,c} = \mathbb{1}[\, m_{i,c} \geq \tau_c \,].$$

A trace is flagged as violating if it triggers **at least one** of the five categories (an OR rule, not a weighted combination of the five individual rates — see §4 for the justification of this design):

$$\text{comp}_i = \mathbb{1}[\, \exists\, c : b_{i,c} = 1 \,] = \max_{c=1,\ldots,K} b_{i,c}.$$

3.4 Observed Rate and Its Interval

$$p_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \text{comp}_i.$$

Wilson-type confidence interval, with the finite-sample variance correction $s^2 = \frac{n}{n-1}\, p(1-p)$:

$$\text{center} = \frac{p + \dfrac{z^2}{2n}}{1 + \dfrac{z^2}{n}} \qquad \text{half-width} = \frac{z}{1 + \dfrac{z^2}{n}} \sqrt{\frac{s^2}{n} + \frac{z^2}{4n^2}}.$$

Minimal numerical example. With $n = 201$, $p_{\text{obs}} = 106/201 = 0.527$, $z = 1.96$: $s^2 = \frac{201}{200}(0.527)(0.473) = 0.2499$; denominator $= 1 + \frac{1.96^2}{201} = 1.0191$; center $= \frac{0.527+0.0096}{1.0191} = 0.5266$; half-width $= \frac{1.96}{1.0191}\sqrt{\frac{0.2499}{201} + \frac{3.8416}{4 \cdot 201^2}} = 1.9235\sqrt{0.001243 + 0.0000238} = 1.9235 \times 0.03559 = 0.0685$. Result: 0.527 ± 0.069. This is a purely illustrative example of the formula, with hand-chosen values — it does not come from any measurement.

3.5 Rogan–Gladen Correction

If the judge has known sensitivity sens and specificity spec (measured on a golden set — see the companion paper), the corrected true prevalence is

$$p_{\text{true}} = \frac{p_{\text{obs}} + \text{spec} - 1}{\text{sens} + \text{spec} - 1}.$$

Minimal numerical example. Let sens $= 0.90$, spec $= 0.95$, $p_{\text{obs}} = 0.30$. Denominator $= 0.90 + 0.95 - 1 = 0.85$. $p_{\text{true}} = \frac{0.30+0.95-1}{0.85} = \frac{0.25}{0.85} = 0.294$. That is: if the judge observes a 30% violation rate but is known to commit a 5% false-positive rate and to miss 10% of true positives, the estimated true prevalence is 29.4%, not 30% — a small correction because the instrument in this example is fairly good. As sens + spec − 1 approaches 0 (an instrument with no discriminative power), the denominator approaches 0 and the correction becomes unstable — see §3.6 on the detection floor. This example, like the previous one, is purely illustrative.

The variance of p_{true} is obtained by the delta method, propagating the variance of p_{obs}, sens, and spec:

$$\text{Var}(p_{\text{true}}) \approx \left(\frac{1}{d}\right)^2 \text{Var}(p_{\text{obs}}) + \left(\frac{p_{\text{obs}} + \text{spec} - 1}{d^2}\right)^2 \text{Var}(\text{sens}) + \left(\frac{\text{sens} - p_{\text{obs}}}{d^2}\right)^2 \text{Var}(\text{spec}),$$

with $d = \text{sens} + \text{spec} - 1$, $\text{Var}(\text{sens}) = \frac{\text{sens}(1-\text{sens})}{n_{\text{pos,cal}}}$, $\text{Var}(\text{spec}) = \frac{\text{spec}(1-\text{spec})}{n_{\text{neg,cal}}}$ (the binomial variances of the calibration golden set itself).

3.6 Youden Threshold and Detection Floor

The per-category threshold is defined as the one that maximizes the Youden index over a sweep of the empirical ROC curve on a golden set:

$$\tau_c^* = \arg\max_{\tau} \left[\text{sens}_c(\tau) + \text{spec}_c(\tau) - 1\right].$$

We define $J = \text{sens} + \text{spec} - 1 \in [-1, 1]$ as the discrimination index of the calibrated instrument ($J = 0$ is pure chance, $J = 1$ is perfect discrimination). We declare a **detection floor** at $J = 0.20$: below that value, p_{true} is not reported as a reliable measurement — it is flagged as *below_detection_floor* and treated as a census fact ("this could not be measured with this instrument"), not as a numerical finding. This is a declared design decision, part of the instrument's definition, independent of any particular calibration performed on it.

4 Why the Trace Is the Sampling Unit

A single turn can be repeated (or nearly repeated) many times within the same trace if an attacker persists with variations of the same vector; counting turns as independent observations artificially inflates the sample size and invalidates the confidence interval (the observations would no longer be independent). The trace — a complete conversation against an attack vector — is the correct experimental unit: each trace is an independent attempt, even if it contains multiple turns.

For the same reason, the five per-category violation rates are not combined into a weighted average to form the composite rate: they are combined at the level of *decision* (the trace triggers if any category triggers, §3.3), and the composite rate is measured directly on that binary decision, not derived algebraically from the five individual rates. The reason is that the five categories' error probabilities are not independent of one another in a way that admits a simple closed-form formula, and measuring the composite directly avoids that problem by construction — the concrete empirical illustration of this consequence (the so-called "ceiling effect": the composite specificity can turn out lower than the minimum specificity of any individual category) is shown with data in the companion paper, §4.1.

5 Scope of This Document and Next Steps

This document fully defines the mathematical form of the instrument — what is measured, how it is aggregated, what form its uncertainty takes — but deliberately leaves three free parameters unresolved: τ_c, sens_c, and spec_c for each of the five categories. Fixing those values requires a golden set and an explicit calibration procedure, and the result of that calibration depends entirely on which concrete instrument (lexical, semantic, LLM-as-judge, composite) is being evaluated at a given time — by design (§3.1, "Layer independence"), the metric's definition does not change when the instrument changes, but the numerical values of τ_c, sens, and spec do.

For this same reason, neither this document nor its companion paper evaluates, compares, or recommends one concrete judge implementation over another: the choice and construction of the judge — lexical, semantic, LLM-as-judge, or composite — belongs to the system's measurement layer, not to the aggregation layer defined here, and is deliberately kept outside the scope of this series.

The companion paper in this series, "Calibration and Empirical Validation of a Measurement Instrument for Safety & Harmful Content in AI Red-Teaming," covers: the calibration methodology and the `calibration_source` field; three stages of simulation validation under known ground truth; real-data validation (HateCheck) of a concrete lexical instrument; the sample size required for future real-world calibration; and an illustrative end-to-end example of applying this metric to a red-teaming attack.

6 Conclusion

The definition presented here — maximum collapse, composite OR rule, per-category Youden threshold, Wilson interval, and Rogan–Gladen correction with delta-method variance — is internally consistent and agnostic to the implementation of the judge that feeds it. It is, deliberately, only a definition: it makes no claim about how well a concrete instrument calibrated with it performs in practice. That question — the one that ultimately matters for deciding whether the instrument is ready for production — is the subject of the companion paper of this series.

References

[1] Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.

[2] Rogan, W.J. & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1), 71–76.

[3] Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.

[4] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H. & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *ACL 2021*. https://github.com/paul-rottger/hatecheck-data

[5] MLCommons (2024–2026). AILuminate Safety Methodology. https://mlcommons.org/ailuminate/safety-methodology/

[6] OpenAI developer community discussion on undisclosed Moderation API thresholds. https://community.openai.com/t/understanding-category-scores-moderation-values/1368610

[7] (Google/Jigsaw) Designing Toxic Content Classification for a Diversity of Perspectives.