THE HITCHHIKERS GUIDE TO RUBRIC QUALITY UNDERSTANDING AND ENRICHMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Rubrics distill notions of expert quality and measure agent performance. However, the quality of rubrics themselves have not been systematically measured and are often left to downstream performance.We import apparatuses from measurement theory built for exactly this: quantitative signals based on the rubric's content, and introduce the **RubrIc-Failure Taxonomy (RIFT)**, of nine possible ways a rubric fails, organized under reliability and content validity. Every mode leaves a distinct signature. To show the signals track failure causally, we seed 720 corruptions, injecting each RIFT mode into clean rubrics at known severity levels. A linear probe over the signals identifies which mode was injected at 75.0% accuracy, beating 56.7% for a frontier model asked to name the failure directly. Surprisingly across GDPval and Terminal-Bench, 10 of 48 expert-authored rubrics weight their criteria backwards, putting more of the score on requirements an expert panel judged less essential. This means a response can fail what matters most and still be graded well. This paper serves as a comprehensive guide on how to understand failure modes in rubrics and create better versions using quality signals, causal experiments, and provides a taxonomy with its rules and examples.

1 INTRODUCTION

Rubrics have become central to the evaluation and training of large language models (LLMs) (Hashemi et al., 2024; Liu et al., 2026; Ding, 2026). Prior work evaluates rubrics from downstream outcomes, including agreement rates between automated judges (Hashemi et al., 2024; Shen et al., 2026) and improvements in models trained using rubric derived rewards (Gunjal et al., 2025; Liu et al., 2026). These methods conflate the rubric with the judge, and both are insensitive to the failure that matters most which is a rubric that is internally consistent, interpretable, but measuring the wrong thing. Recent work has begun to study rubric quality alongside downstream outcomes (Arabzadeh et al., 2024; Siro et al., 2026; Hong et al., 2026; Roy et al., 2026).

Educational measurement and the social sciences have spent decades formalizing the validity of an assessment instrument. Most famously, Lawshe's content validity ratio for expert agreement on item essentiality (Lawshe, 1975), and the decomposition of instrument validity into criterion, content, and construct validity (Straub et al., 2004; Taherdoost, 2016). Importing these into rubric evaluation yields quantifiable signals. However, signals alone tell half a story. Numeric signals can tell a practitioner that something is wrong; it does not tell them what is wrong, and therefore does not tell them what to fix. This necessitates an explanation that connects quality signals to specific failure modes. **This paper supplies both, the quality signals and the taxonomy for explanations.**

To provide intuitive and measurable signals we propose four families of rubric quality signals. The first uses a panel of simulated experts to judge whether each criterion is essential to the task. The second groups criteria by their relevance, from core requirements to peripheral concerns. The third evaluates how well criteria align with the task's domain relevant objectives. The fourth examines whether criterion weights reflect their importance. Together, these four families provide computable signals for assessing rubric quality.

We also introduce **RIFT**, the **R**ubr**I**c **F**ailure mode **T**axonomy, a vocabulary to label rubric quality issues with **nine** rubric failure modes: **subjective language, ungrounded verification, non-atomicity, missing criteria, misalignment, rigidity, redundancy, and hackability, and miscalibrated weighting**. We map quality signals to the taxonomy empirically and causally with known

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

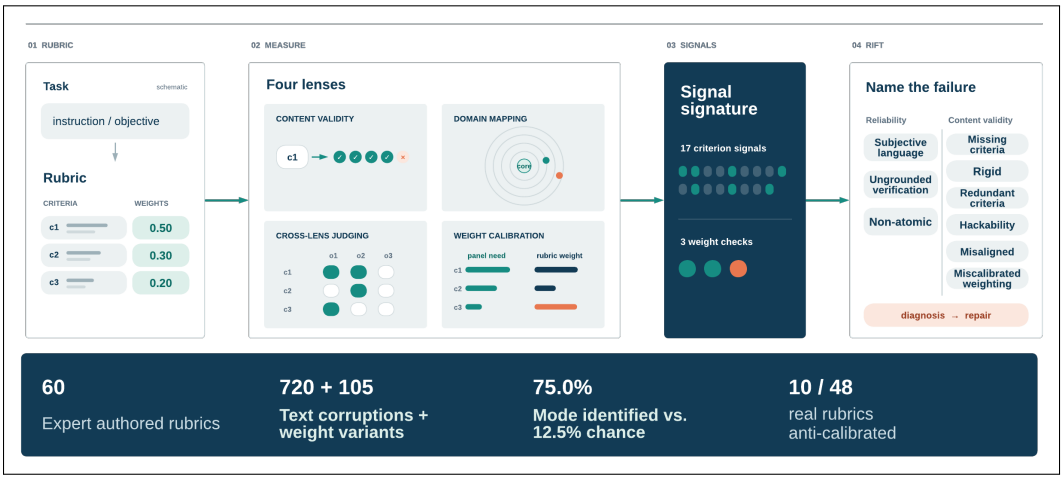

Figure 1: This paper grounds rubric design in well-established quality signals. These help us understand and view causal effects on rubric design. We also provide a taxonomy **(RIFT)** to help designers name specific failure modes.

corruptions: **720** injected rubric variants across **60** expert-human-authored, weighted rubrics drawn from two agentic benchmark families GDPVal (Patwardhan et al., 2025) and Terminal-Bench (Merrill et al., 2026), spanning three corruption levels. The injected mode is known by construction in every case so we can ask directly whether the psychometric signature of a rubric identifies the way it is broken. Every mode produces a statistically distinguishable signature: **46 of 136** mode-by-signal cells are significant under BH-FDR correction, with strong effect sizes up to $|d| = 2.83$. Together these help practitioners identify, quantify, and label rubric failures early in the development process, avoiding expensive failures stemming from leaving problem discovery solely in downstream tasks.

This paper makes three contributions. We define a set of quantifiable quality signals computed directly from a rubric's content (§3), derive a taxonomy of rubric failure modes from measurement theory (§4), and establish through causal experiments that the signals identify which failure mode is present rather than only that something is wrong (§5). Together these give a rubric designer two things they did not have: a way to measure a rubric before it is used, and a taxonomy, grounded in decades of assessment research, for naming what the measurements say is wrong.

2 Related Work

2.1 Rubrics in the era of agents

Rubrics are used to evaluate and guide the training of AI agents. Recent research investigates rubric development methodologies (Liu et al., 2026; Sharma et al., 2026; Shen et al., 2026; Ding, 2026), applies them as the evaluation mechanism itself (Hashemi et al., 2024; Wei et al., 2025), and, most recently, asks whether a rubric is itself *good* (Arabzadeh et al., 2024; Pan et al., 2026; Siro et al., 2026; Hong et al., 2026; Roy et al., 2026; Dhole & Agichtein, 2026). For instance, that third line of research finds machine-generated rubrics diverging from human-authored ones and warns that agreement between judges is not validity (Dhole & Agichtein, 2026). Together this line of work demonstrates that rubrics can yield reliable downstream measurements but lack validity. Rubrics come from educational measurement, a field has spent decades formalizing measurement, labels, and computable statistics. In this paper we bridge these fields, providing practitioners a systematic account of the ways rubrics fail how to holistically quantify their quality before observing downstream effects.

2.2 The history and design of rubrics: lessons outside CS

Rubrics were studied as measurement instruments long before they were used to evaluate LLMs. Educational assessment distinguishes analytic rubrics, which score criteria one at a time and support

formative feedback, from holistic rubrics, which collapse everything into a single judgment that is faster to apply but less informative (Brookhart, 2018), and examines what makes descriptive rubrics effective in practice (Brookhart & Chen, 2015).

A rubric is a measurement instrument. So the broader validity literature applies to it directly. Validity concerns how well an instrument covers the domain under investigation (Ghauri & Grønhaug, 2005) and whether it measures what it is intended to measure (Field, 2005). Taherdoost (2016) decomposes it into criterion, face, content, and construct validity. Of these, content validity, the degree to which items reflect the content universe the instrument generalizes to (Straub et al., 2004), is the most directly computable: Lawshe (1975) estimates it numerically from panel agreement on whether each item is essential, and consensus methods extend the approach to full questionnaires (Montes et al., 2023).

3 Quality Signals

We define four families of quality signals, each importing an established validation measure and answering a different question. *Content validity* asks whether each criterion belongs in the rubric. *Domain mapping* asks whether the criterion set covers the domain in the right proportions. *Cross-lens judging* asks whether criteria align with the task's objectives. *Weight calibration* asks whether scoring mass sits on the criteria that matter. Together, these four families provide a holistic view of rubric quality.

A rubric is a collection of M *criteria* $\{(c_i, w_i)\}_{i=1}^{M}$, each pairing a natural-language statement c_i with a non-zero numerical *weight* w_i expressing its importance. Weights may be negative where a rubric penalizes undesired behavior. We write $s_i = |w_i| / \sum_k |w_k|$ for the *scoring mass* a criterion commands, so that $\sum_i s_i = 1$.

3.1 Content validity ratio

CVR is Lawshe's *content validity ratio* (CVR) (Lawshe, 1975), which is still active recently (Almanasreh et al., 2019). A panel of N experts labels each criterion essential or not, and the content validity ratio for criterion i is

$$\mathrm{CVR}_i \;=\; \frac{n_i^e - N/2}{N/2} \;\in\; [-1, 1], \tag{1}$$

where n_i^e is the number of essential votes for criterion i. The statistic reaches $+1$ under unanimous endorsement, sits at 0 at a bare majority, and turns negative when most of the panel judges the criterion inessential. We report a rubric's mean CVR, following the adjacent content-validity-index tradition's caution to state the aggregation explicitly (Polit & Beck, 2006).

A panel is instantiated with $N{=}5$ task-adapted personas, generated by a GPT-5.5 model per task from the task instruction alone. Each persona is a one-line expert profile covering a distinct specialism relevant to that task. An example for a prehospital cardiac-care task, the generated set comprises an EMS medical director, a cardiac QA paramedic, an interventional cardiologist, a protocol compliance officer, and a clinical documentation auditor. Examples can be found in Appendix G

3.2 Domain mapping

Content validity evaluates criteria one at a time. Domain mapping evaluates the set as a distribution. Instrument-development protocols require an explicit content plan before criteria are judged (Fives & DiDonato-Barnes, 2013; Kalkbrenner, 2021), and validity theory supplies the two questions such a plan must answer: content *relevance* §3.1, and content *representativeness*, whether the set covers the domain in the right proportions (Messick, 1989; Haynes et al., 1995; Sireci, 1998). Alignment methodology turns these into gradeable criteria with explicit thresholds (Webb, 1997; 2007; Porter, 2002), while a fourth tradition places items in a domain *geometry* via expert ratings and multidimensional scaling (Sireci & Geisinger, 1992; 1995), where radial distance from the center is itself measurement-bearing structure (Guttman, 1954; Marshalek et al., 1983). To our knowledge no existing rubric-validation framework combines these elements; our domain mapping automates all four.

For each task a GPT-5.5 model derives from the instruction a *domain map*: five ordered relevance levels, CORE, ADJACENT, BACKGROUND, GENERAL, and IRRELEVANT, each populated with task-specific terms. Let $d_i \in \mathcal{D}$ denote the relevance level assigned to criterion c_i, where $\mathcal{D} = \{\text{CORE}, \text{ADJACENT}, \text{BACKGROUND}, \text{GENERAL}, \text{IRRELEVANT}\}$. The rubric's domain alignment is the fraction of criteria assigned to each level:

$$p_r = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}[d_i = r], \qquad r \in \mathcal{D}, \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. For e.g. in the prehospital cardiac audit, CORE covers ECG timing compliance while GENERAL covers report structure and prose. We record per-level threshold indicators and flag any criterion landing at IRRELEVANT.

3.3 CROSS-LENS EXPERT JUDGING

Following structured elicitation (Dalkey & Helmer, 1963), $E{=}5$ expert personas independently propose objectives from the instruction. These consolidate into a list of objectives carrying normalized weights W_j and one of four centrality tiers (Raymond, 2001; Lawshe, 1975), with an LLM judge (Zheng et al., 2023) scoring panel agreement and objective verifiability. Each expert rates every criterion against every objective on a 0–3 scale, the multidimensional form of item–objective congruence (Rovinelli & Hambleton, 1976; Turner & Carlson, 2003), giving $\bar{R}_{ij} \in [0, 1]$.

Classical construct threats (Messick, 1989) then become computable. *Item relevance* $\frac{1}{M}\sum_i \max_j \bar{R}_{ij}$ and *weighted relevance* $\frac{1}{M}\sum_i \sum_j W_j \bar{R}_{ij}$ catch criteria tied to no objective, or only to peripheral ones. *Representativeness* $1 - \text{JSD}(W \| A)$, with realized coverage $A_j \propto \sum_i s_i \bar{R}_{ij}$ and base-2 Jensen–Shannon divergence (Lin, 1991), measures construct underrepresentation; *contamination*, the s_i-weighted mass on background and irrelevant tiers, construct-irrelevant variance (Haladyna & Downing, 2004); *redundancy*, the mean cosine among criterion pairs at similarity ≥ 0.8 (zero if none), duplicated content (Boyle, 1991; Yen, 1993). Higher is better for the first three, worse for the last two.

Discriminant controls. We build a nomological net (Cronbach & Meehl, 1955) and close it discriminantly (Campbell & Fiske, 1959). Three items from each of four classes are scored identically to the rubric's own criteria: expert positives, adjacent-but-wrong-domain near negatives, generic quality statements, and far negatives. We report the margins separating candidate criteria from the generic and near-negative classes. A valid rubric must separate its criteria from plausible impostors, not merely score them highly, a check aimed at known LLM-judge biases (Zheng et al., 2023; Salewski et al., 2023).

3.4 WEIGHT CALIBRATION

A rubric whose every criterion is sound can still misrepresent the construct if w allocates scoring mass to the wrong criteria. Because practice-analysis methods derive numeric blueprint weights directly from expert judgment (Raymond, 2001; Lawshe, 1975), a rubric's weights are themselves a testable claim about relative importance rather than a scoring convenience. *Ordering calibration* asks whether $|w_i|$ ranks criteria the way the panel does, via the Spearman rank correlation

$$\rho_{\text{ord}} = \text{Spearman}\big(\{|w_i|\}_{i=1}^{M}, \{\text{CVR}_i\}_{i=1}^{M}\big) \in [-1, 1], \tag{3}$$

undefined by convention when either vector is constant. Values near 1 indicate weights that track panel necessity, while negative values indicate *anti-calibration*: the rubric awards more credit for satisfying the less essential criterion of a pair.

Let $\mathcal{C} = \{i : \text{CVR}_i > 0\}$ denote the panel-endorsed *core* criteria. *Failure score retention* (FSR) is the worst-case score fraction a response retains after failing exactly one core criterion while satisfying every other:

$$\text{FSR} = \max_{i \in \mathcal{C}} \left(1 - \frac{|w_i|}{\sum_{k:w_k>0} w_k - \sum_{k:w_k<0} w_k}\right), \tag{4}$$

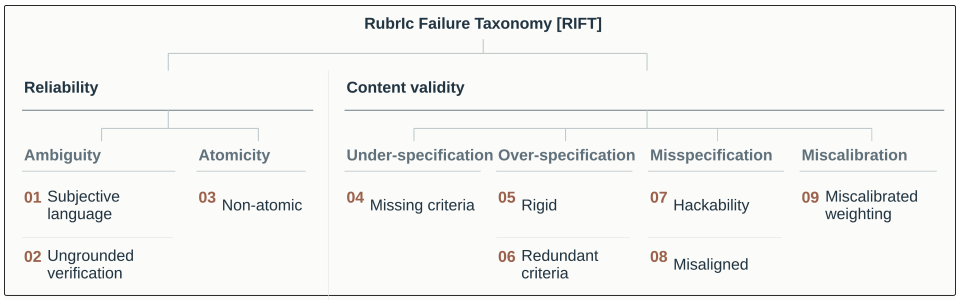

Figure 2: The RubrIc Failure Taxonomy (RIFT) organizes nine modes of rubric failure. See table 2 for complete definitions.

which reduces to $1 - \min_{i \in \mathcal{C}} s_i$ under all-positive weights. High FSR flags rubrics where failing a core criterion is cheaply compensated by credit elsewhere. Relatedly, *core weight share* $\sum_{i \in \mathcal{C}} s_i$ reports what fraction of total scoring mass the panel-endorsed criteria command at all. Higher ordering calibration and core weight share are better; higher FSR is worse.

4 RIFT: The RubrIc Failure Mode Taxonomy

We introduce the **RubrIc Failure mode Taxonomy (RIFT)**, a general framework for identifying and classifying common rubric failures, and the vocabulary we use to explain *how* those failures move the quality signals of section 3. RIFT is hierarchical: quality *dimensions* organize *sub-dimensions*, which group individual *failure modes*. Each mode carries a complete definition, decision rules specifying when it applies, and pass and fail examples that fix its boundaries. Figure 2 previews the structure and Appendix B gives the full taxonomy.

4.1 Grounding RIFT in assessment research

A rubric is a measurement instrument, and measurement theory distinguishes two properties of one. *Reliability* is the reproducibility of a rubric's evaluations. *Content Validity* is the extent to which its content supports the intended interpretation of those evaluations. Reliable evaluations are necessary but not sufficient for valid interpretations (Messick, 1989; Cook & Beckman, 2006).

Research on descriptive rubrics emphasizes clear, focused, substantive, and purpose-appropriate criteria (Brookhart & Chen, 2015; Brookhart, 2018), which RIFT captures through the Reliability sub-dimensions of *Ambiguity* and *Atomicity*. Content-validity research defines quality by whether an instrument's items are relevant to and representative of its intended domain (Lawshe, 1975; Haynes et al., 1995; Sireci, 1998), distinguishing omissions that under-represent the intended content from extraneous requirements and proxies that introduce construct-irrelevant variance (Messick, 1989; Haladyna & Downing, 2004), and identifying rephrased near-duplicates as a source of redundancy (Boyle, 1991). RIFT extends these into the Content Validity sub-dimensions of *Under-Specification*, *Over-Specification*, *Misspecification*, and *Miscalibration*. Appendix A.1 details the corpus, annotation rounds, LLM-assisted critique aggregation, and expert refinement.

4.2 Intervention operators

To support our causal analysis in section 5, each failure mode is realized by an *operator* that applies it to a clean rubric at a requested corruption rate ρ, so the clean rubric serves as its own control and the treatment effect is the difference between the original and the corrupted rubric. Criterion-level operators sample a ρ fraction of criteria to delete, duplicate, merge, or rewrite. The Miscalibrated Weighting operator is deterministic, moving a ρ fraction of scoring mass from panel-endorsed core criteria to non-core criteria. Full operator definitions and prompts are provided in Appendix C.2.

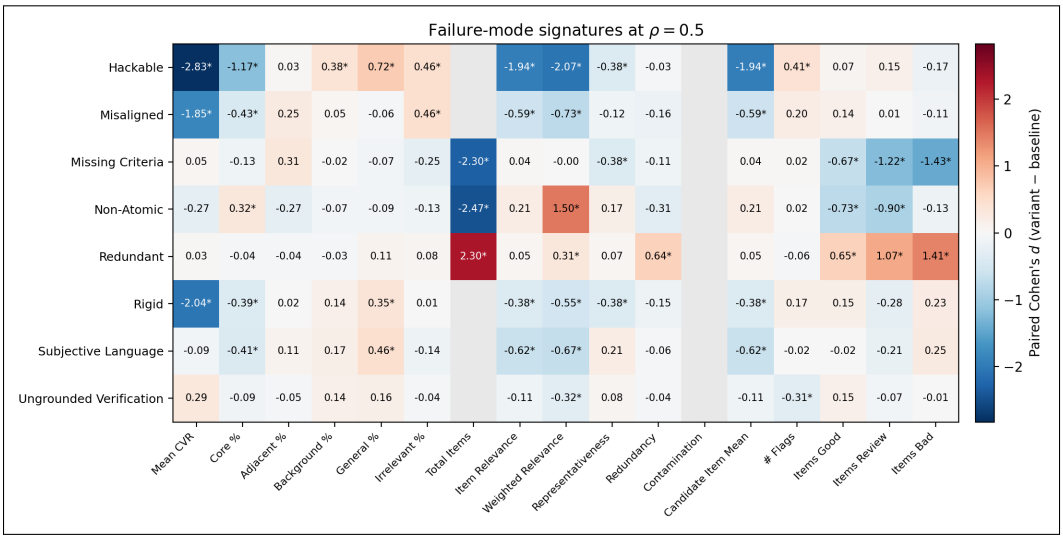

Figure 3: Failure-mode signatures at ρ=0.5: paired Cohen's d (variant − baseline) for each of 8 injected modes across 17 signals, n=60 tasks per mode. Asterisks mark cells significant under Benjamini–Hochberg FDR at $q < .05$ applied across the full 8×17 grid (46 of 136). Blank cells are signals whose delta was identically zero on every task.

5 Experiments and Results

We establish a causal connection between RIFT and quality signals with four main RQs:

- RQ1) *does each mode leave a distinct signature* (§5.1)?,
- RQ2) *do quality signals degrade with corruption level* (§5.2)?,
- RQ3) *do the findings transfer across rubric families and the signals carry independent information* (§5.3)? and
- RQ4) *what do criterion-reading signals not see* (§5.4)?

All four questions need rubrics with known defects. We start from clean expert-authored rubrics and use the intervention operators of section 4.2 to inject RIFT failure modes at controlled severity. We score each clean and corrupted rubric on the quality signals so any movement in the signals is attributable to the mode we injected.

Corpus and injection. We use T=60 human-authored rubrics, 30 from GDPval (Patwardhan et al., 2025) and 30 from Terminal-Bench (Merrill et al., 2026). We apply each of RIFT's eight criterion-level operators at ρ=0.5 across all 60 tasks (480 variants) and at $\rho \in \{0.25, 0.75\}$ on a seeded 15-task subset (240 variants) for corruption-level analysis. The ninth mode, Miscalibrated Weighting, uses the deterministic weight-level operator and is evaluated separately in §5.4. Feature scoring uses GPT-5.5 for the baseline rubric and every variant alike, while corruption is authored by a sonnet-4-6 model, so scorer and corruptor are independent. Task-level instruments (personas, domain map, objective blueprint, discriminant controls) are fixed per task across arms, so within-task deltas isolate the injected mode. Appendix D replicates the main results under additional scoring models to show that our results are not model-specific artifacts.

5.1 Injected failure modes produce mode-specific signatures

Each mode damages a rubric in its own recognizable way. Across the 8 mode $\times$ 17 quality signal grid, **46 of 136 cells** are FDR-significant ($q < .05$), and every mode moves at least one signal at ρ=0.5 (n=60/mode; Figure 3, per-mode effect sizes in Appendix E).

The misspecification modes degrade content validity. *Hackable* most severely of all at d=−2.83 on mean CVR (AUC=0.13). The modes affecting structure (redundant, non-atomic, missing criteria)

move counts by 4 to 7 criteria while content validity holds. One entry runs against type: non-atomic *raises* weighted relevance (d=+1.50), because merging criteria makes the survivors look more relevant.

The signals are selective, not one severity dial. Content validity fires for the three misspecification modes while item-count signals fire for the three structural modes. Redundancy score fires for redundant alone, which is a tight signal-to-construct correspondence. Only weighted relevance behaves as a general severity dial, moving for 7 of 8 modes.

The signature identifies the mode and the damage. A multinomial logistic classifier over the 17-signal delta vector recovers which mode was injected with **75.0% accuracy** (macro-F1 0.746), under leave-one-task-out cross-validation with all 8 arms of a held-out task removed together. Only two confusion pairs are large: subjective-language with ungrounded-verification under Ambiguity, and misaligned with rigid. Collapsing just these pairs raises accuracy to 89.4%. The classifier was never told the taxonomy has sub-dimensions, yet it recovered them from the signals alone. Adding Miscalibrated Weighting as a ninth label costs little (**72.0%**, macro-F1 0.712, chance 11.1%), with mode-9 recall 0.68. Our validation compares each corrupted rubric against its own clean parent. The signals themselves are absolute, as the anti-calibration finding demonstrates, so diagnosis outside this setting requires population-level norms in place of a per-rubric reference.

A judge cannot do this by inspection. We showed the same 540 units (480 corrupted, 60 clean) to an LLM judge given the instruction and the rubric, and asked it to name one of the eight modes or "none." The judge, `claude-opus-4-8`, scores **56.7%** (macro-F1 0.530), **18.3 points** below the feature-based classifier. Its failures are specific: it over-predicts misaligned (recall 1.00, precision 0.33), recognizes healthy rubrics only 18% of the time, and gets zero of 60 ungrounded-verification units right.

5.2 Signals grade severity, not just presence

Knowing a rubric is damaged is less useful than knowing how badly. On the 15-task subset we regress each signal on the realized corruption fraction per unit. The modes that damage content validity grade cleanly: mean CVR falls at -1.55 per unit realized ρ for hackable (p=1.6×10^{-12}), -1.35 for rigid, and -1.10 for misaligned, each replicating independently in both datasets. Non-atomicity grades just as reliably in the opposite direction ($+0.159$, r=+.67), confirming that relevance inflation scales with merging rather than being an artifact of one corruption level. This matters beyond our setting: any quality measure built on mean per-criterion relevance can be improved by consolidating criteria. Content validity therefore acts as a *dosimeter* for misspecification while the domain and cross-lens signals largely act as *presence detectors*, and a practitioner should read a depressed CVR as an estimate of severity but a fired detector only as a flag for inspection. See (Appendix F) for detailed visual results.

5.3 Effects transfer across rubric families, and the signals are independent

Rubric family moderates six of the eight criterion-level modes but never reverses their sign. Deliverable-style GDPval rubrics degrade more sharply than Terminal-Bench's technical checklists under every content-validity mode (Δmean_CVR -0.87 vs. -0.42 for rigid, -0.80 vs. -0.61 for hackable, -0.65 vs. -0.45 for misaligned; all $q < .02$, BH-corrected within the eight moderation tests), plausibly because prose criteria admit more room for the corruption to work. The one reversal is mechanical: Terminal-Bench rubrics are longer, so deleting a fixed fraction removes more criteria.

If the signals were three views of one latent quality factor, the damage one registers would be explained by another. A bootstrap mediation analysis on hackable (2000 resamples) finds the CVR collapse routed neither through the domain shift (0% mediated via Δcore_pct; CI $[-.094, +.066]$) nor through relevance loss (2% via Δitem_relevance, CI straddling zero), against a total effect of -0.700. The signals register the same damage through independent channels, showing diagnostic power lives in the full vector and not in one scalar.

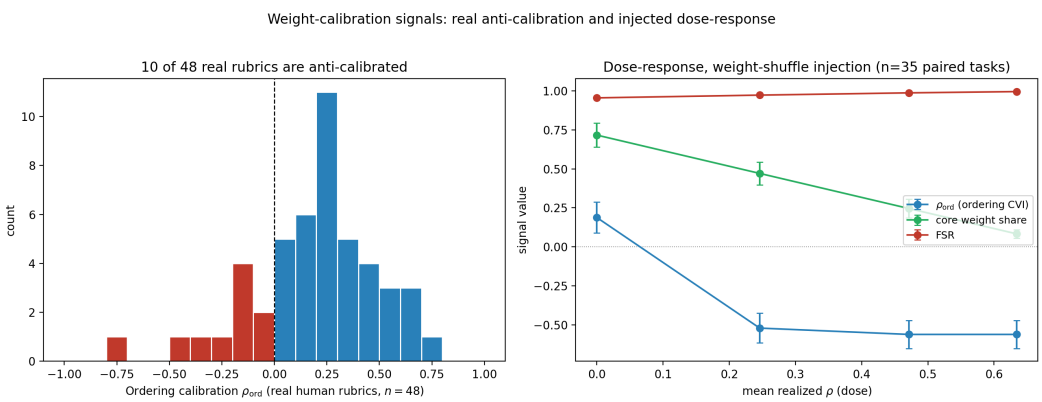

Figure 4: Left: distribution of ρ_{ord} over the 60 human-authored rubrics (n=48 scoreable); red bars are anti-calibrated. Right: the three weight-calibration signals under weight-shuffle injection, mean $\pm 95\%$ CI over 35 paired tasks against mean realized ρ.

Table 1: Weight-calibration signals on human rubrics (n=60, panel-scored) and under weight-shuffle injection (n=35 paired). Higher ρ_{ord} and core weight share are better; higher FSR is worse.

	Human	**Clean**	ρ=.25	ρ=.50	ρ=.75
ρ_{ord}	+0.195	+0.189	−0.522	−0.562	−0.562
Core weight share	0.835	0.717	0.471	0.246	0.083
FSR	0.950	0.956	0.974	0.988	0.996

5.4 The blind spot: weight miscalibration

So far we have dealt with criterion *text*, and every criterion-level operator preserves total rubric weight, so a rubric can pass all of them while allocating scoring mass to the wrong criteria. We close this gap using weight-calibration statistics of §3.4.

Anti-calibration is present in human-authored rubrics. Scored against the real N=5 panel on all 60 rubrics, expert weights modestly track panel-rated necessity on average (mean ρ_{ord}=+0.195, t=4.44, p=5.5 $\times$ 10^{-5}, n=48 scoreable). There still remains a substantial minority: **10 of 48 scoreable rubrics are negative**, ranking criteria against panel-rated necessity, so the rubric awards more credit for the less essential criterion of a pair. This is the Miscalibrated Weighting phenotype in expert-authored rubrics rather than injected. The remaining 12 rubrics have a constant CVR vector, admitting no correlate-able ranking, so we do not score them to preserve the statistic's meaning.

Injection confirms all three statistics respond (Table 1). On 105 corrupted rubrics across 35 paired tasks, ρ_{ord} flips sign at the first corruption level (+0.19 $\rightarrow$ −0.52, d=−1.98, p=1.8 $\times$ 10^{-13}), core weight share falls monotonically (0.72 $\rightarrow$ 0.08, all p<10^{-21}), and FSR rises monotonically (0.956 $\rightarrow$ 0.996, all p<10^{-8}). This is correct since moving mass off core criteria makes a core failure cheaper to absorb. The three differ in *how* they grade. Within the corrupted arms alone, core weight share and FSR remain graded (r=−0.43 and r=0.45, both $p < 10^{-5}$), tracking how much mass has moved, while ρ_{ord} saturates (r=0.005) and reports only whether the move changed the ranking. The operator's positivity constraint caps movement at high ρ, so realized rather than requested ρ is the correct axis throughout.

The criterion-reading signals fail exactly as predicted. Running the same 105 units through the 17-signal grid, 16 of 17 stay silent and the one that fires, `representativeness_score`, is the only signal whose formula consumes weights. Its response is continuous and dose-graded (slope −0.078, p=6.2 $\times$ 10^{-4}, n=105), monotone across bins. This validates our mechanism. The grid responds exactly where the measurement theory says weights should enter, which is evidence the signals measure what they claim. It also shows why the dedicated statistics are needed: representativeness can tell you the weighting is wrong, but not which criteria are at fault.

6 DISCUSSION

In this section we discuss each major takeaway and its interpretation.

From measurement to diagnosis. The central claim of this work is that psychometric signals identify *how* a rubric is damaged. A linear probe over the 17-signal vector recovers the injected failure mode at 75.0% accuracy against a 12.5% chance baseline, and does so under leave-one-task-out validation, so the mapping is not task-specific. It exceeds a strong LLM asked directly to name the failure mode by 18.3 points (56.7%), indicating that the measurement apparatus contributes information that judge intuition does not already supply.

The instrument measures what it claims to measure. Two results speak to construct validity from opposite directions. The redundancy signal fires for injected redundancy and for none of the other modes, giving a signal-to-construct correspondence that is rare in evaluation metrics. More strikingly, the miscalibrated-weighting mode leaves every item's text byte-identical and perturbs only the weight vector; 16 of the 17 signals remain silent, and the one signal that fires is the only one whose computation consumes weights at all. Its response is dose-graded on the realized corruption fraction (slope -0.078, $p = 6.2 \times 10^{-4}$, $n = 105$), with a monotone gradient across dose bins.

A division of labor across families. The quality signals are not interchangeable, and their differences are functional. CVR behaves as a dosimeter for the misspecification modes, falling at -1.55 per unit corruption fraction for hackability ($p = 1.6 \times 10^{-12}$) The domain-mapping lens behaves as a presence detector The cross-lens panel is what catches the modes that invert the logic of the other two. A practitioner reading these signals should therefore read them for different purposes: some answer whether, some answer how much, and some answer only which.

Relevance metrics can be inflated by consolidation. The non-atomicity result deserves emphasis because it generalizes beyond our setting. Merging several criteria into one compound item makes the rubric worse by every substantive standard, yet mean weighted relevance *rises* ($d = +1.50$), and rises further as more items are merged (dose slope $+0.159$, $r = +0.67$). The mechanism is not subtle: a consolidated item touches more of the task, so per-item relevance scores it higher. Any rubric-quality metric built on averaged per-item relevance is therefore optimizable by consolidation, and any rubric-generation procedure that treats such a metric as a reward will drift toward compound, unatomic criteria. This is a concrete instance of the general hazard the taxonomy is meant to surface, and it recommends that mean-relevance measures always be reported alongside item counts.

Implications of RIFT. Readers are now able to name deficiencies in grounded theory. Previously, confronting a low content validity ratio and a depressed core-relevance share pointed to a faulty rubric. The mapping established here licenses a stronger statement: that the rubric is probably hackable, and that the indicated repair is to bind its items to task-specific evidence. Rubric family moderates degradation severity, with deliverable-style rubrics degrading more sharply than technical checklists across six of eight headline comparisons, though direction never reverses. So thresholds should be calibrated per family while the qualitative signature transfers unchanged.

7 CONCLUSION

Rubric quality has been gauged through downstream model performance conflating the rubric with the model applying it. We give practitioners two things instead. First, a set of quantified signals that read a rubric directly, drawn from decades of assessment and measurement research. The second is a taxonomy for naming what those signals find. We validated both by seeding known defects into expert-authored rubrics and recovering which defect was injected from the signals alone. Our results are causal and we show in appendix D that these are not the artifact of a single model or family.

Two limitations follow from our methods. We isolated one failure mode at a time, while rubrics in practice fail in subtler and overlapping ways. We measured absolute weight magnitude, not the penalties, gates, and non-additive scoring rules that some rubrics also use.

The experiment this most immediately motivates is to derive reward signals from clean and corrupted rubrics and compare what the resulting policies learn. While our research points to the MECE (mutual exclusive and collectively exhaustive) nature of our taxonomy, we also encourage readers to always upgrade the evaluations to which we hold other evaluation instruments.

AI DISCLOSURE SECTION

In this work, we used generative AI tools for coding assistance and help with grammatical editing and formatting of the manuscript. We have not used generative AI tools for literature review, writing of the content of the paper, formulation of the research problems and question, or selection of the quality signals. The use of AI in the creation of the RIFT taxonomy is disclosed in Appendix B. The rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Experiments, results, code, citations, and structure, were verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Enas Almanasreh, Rebekah Moles, and Timothy F. Chen. Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, 15(2):214–221, 2019. doi: 10.1016/j.sapharm.2018.03.066.

Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qingyun Wu, Chi Wang, Ahmed Hassan Awadallah, Charles L. A. Clarke, and Julia Kiseleva. Assessing and verifying task utility in LLM-powered applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21868–21888, 2024. doi: 10.18653/v1/2024.emnlp-main.1219.

Gregory J. Boyle. Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, 12(3):291–294, 1991. doi: 10.1016/0191-8869(91)90115-R.

Susan M. Brookhart. Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3:22, 2018. doi: 10.3389/feduc.2018.00022.

Susan M. Brookhart and Fei Chen. The quality and effectiveness of descriptive rubrics. *Educational Review*, 67(3):343–368, 2015. doi: 10.1080/00131911.2014.929565.

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

David A. Cook and Thomas J. Beckman. Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119(2):166.e7–166.e16, 2006. doi: 10.1016/j.amjmed.2005.10.036.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Norman Dalkey and Olaf Helmer. An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3):458–467, 1963. doi: 10.1287/mnsc.9.3.458.

Kaustubh D. Dhole and Eugene Agichtein. Rubricrag: Towards interpretable and reliable llm evaluation via domain knowledge retrieval for rubric generation, 2026. URL https://arxiv.org/abs/2603.20882.

Liang Ding. Adarubric: Task-adaptive rubrics for reliable llm agent evaluation and reward learning, 2026. URL https://arxiv.org/abs/2603.21362.

Andy P. Field. *Discovering Statistics Using SPSS*. Sage Publications Inc., 2005.

Helenrose Fives and Nicole DiDonato-Barnes. Classroom test construction: The power of a table of specifications. *Practical Assessment, Research, and Evaluation*, 18(3), 2013. doi: 10.7275/cztt-7109.

Pervez Ghauri and Kjell Grønhaug. *Research Methods in Business Studies*. FT/Prentice Hall, Harlow, 2005.

Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Publishing Company, Chicago, 1967.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains, 2025.

Louis Guttman. A new approach to factor analysis: The radex. In Paul F. Lazarsfeld (ed.), *Mathematical Thinking in the Social Sciences*. Free Press, Glencoe, IL, 1954.

Thomas M. Haladyna and Steven M. Downing. Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1):17–27, 2004. doi: 10.1111/j.1745-3992.2004.tb00149.x.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745. URL https://aclanthology.org/2024.acl-long.745/.

Stephen N. Haynes, David C. S. Richard, and Edward S. Kubany. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7(3):238–247, 1995. doi: 10.1037/1040-3590.7.3.238.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G. Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Awadalla, and Manaal Faruqui. Advancedif: Rubric-based benchmarking and reinforcement learning for advancing llm instruction following, 2025. URL https://arxiv.org/abs/2511.10507.

Hanhua Hong, Yizhi Li, Jiaoyan Chen, Luu Gia Huy, Sophia Ananiadou, Jung-jae Kim, and Chenghua Lin. Can LLMs write reliable rubrics? A meta-evaluation for experiment reproduction. *arXiv preprint arXiv:2607.12835*, 2026.

Michael T. Kalkbrenner. A practical guide to instrument development and score validation in the social sciences: The MEASURE approach. *Practical Assessment, Research, and Evaluation*, 26 (1), 2021. doi: 10.7275/svg4-e671. Article 1.

C. H. Lawshe. A quantitative approach to content validity. *Personnel Psychology*, 28(4):563–575, 1975. doi: 10.1111/j.1744-6570.1975.tb01393.x.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. OpenRubrics: Towards scalable synthetic rubric generation for reward modeling and LLM alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.791. URL https://aclanthology.org/2026.acl-long.791/.

B. Marshalek, David F. Lohman, and Richard E. Snow. The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, 7:107–127, 1983. doi: 10.1016/0160-2896(83)90023-5.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026.

Samuel Messick. Validity. In Robert L. Linn (ed.), *Educational Measurement*, pp. 13–103. American Council on Education and Macmillan, New York, NY, 3 edition, 1989.

Rosana Montes, Cristina Zuheros, Jeovani M. Morales, Noe Zermeño, Jerónimo Duran, and Francisco Herrera. Design and consensus content validity of the questionnaire for B-learning education: A 2-tuple fuzzy linguistic Delphi based decision support tool. *Applied Soft Computing*, 147:110755, 2023. doi: 10.1016/j.asoc.2023.110755.

Tianjun Pan, Xuan Lin, Wenyan Yang, Qianyu He, Shisong Chen, Licai Qi, Wanqing Xu, Hongwei Feng, Bo Xu, and Yanghua Xiao. RubricEval: A rubric-level meta-evaluation benchmark for LLM judges in instruction following. *arXiv preprint arXiv:2603.25133*, 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks, 2025.

Denise F. Polit and Cheryl Tatano Beck. The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, 29(5):489–497, 2006. doi: 10.1002/nur.20147.

Andrew C. Porter. Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31(7):3–14, 2002. doi: 10.3102/0013189X031007003.

Mark R. Raymond. Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education*, 14(4):369–415, 2001. doi: 10.1207/s15324818ame1404_4.

Richard J. Rovinelli and Ronald K. Hambleton. On the use of content specialists in the assessment of criterion-referenced test item validity. Technical Report ED121845, ERIC Document Reproduction Service, 1976. Also published in *Tijdschrift voor Onderwijsresearch*, 2(2), 49–60, 1977.

Swastik Roy, Rajkumar Pujari, Tharindu Kumarage, Charith Peris, Rahul Gupta, Anna Rumshisky, Pradeep Natarajan, and Venkatesh Saligrama. PReMISE: Policy rubrics as measurement specifications for LLM judges. *arXiv preprint arXiv:2605.30803*, 2026.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean Hendryx, Bradley Kenstler, and Bing Liu. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 90447–90472, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/92808d4b229345ae40fce84b4bf038d0-Paper-Conference.pdf.

William F. Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking rubric generation for improving llm judge and reward modeling for open-ended tasks, 2026. URL https://arxiv.org/abs/2602.05125.

Stephen G. Sireci. The construct of content validity. *Social Indicators Research*, 45:83–117, 1998. doi: 10.1023/A:1006985528729.

Stephen G. Sireci and Kurt F. Geisinger. Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement*, 16(1):17–31, 1992. doi: 10.1177/014662169201600102.

Stephen G. Sireci and Kurt F. Geisinger. Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement*, 19(3):241–255, 1995. doi: 10.1177/014662169501900303.

Clemencia Siro, Pourya Aliannejadi, and Mohammad Aliannejadi. Learning to judge: LLMs designing and applying evaluation rubrics. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6371–6389, 2026.

Detmar Straub, Marie-Claude Boudreau, and David Gefen. Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, 13:380–427, 2004.

Hamed Taherdoost. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM)*, 5, 2016. hal-02546799.

Ronna C. Turner and Laurie Carlson. Indexes of item-objective congruence for multidimensional items. *International Journal of Testing*, 3(2):163–171, 2003. doi: 10.1207/S15327574IJT0302_5.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models, 2025. URL https://arxiv.org/abs/2507.18624.

Norman L. Webb. Criteria for alignment of expectations and assessments in mathematics and science education. Technical Report Research Monograph No. 6, Council of Chief State School Officers and National Institute for Science Education, University of Wisconsin–Madison, Madison, WI, 1997. ERIC ED414305.

Norman L. Webb. Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1):7–25, 2007. doi: 10.1080/08957340709336728.

Yuchen Wei, Dennis Pearl, Matthew Beckman, and Rebecca J. Passonneau. Concept-based rubrics improve llm formative assessment and data synthesis, 2025. URL https://arxiv.org/abs/2504.03877.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. Auto-rubric: Learning to extract generalizable criteria for reward modeling, 2025. URL https://arxiv.org/abs/2510.17314v1.

Wendy M. Yen. Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3):187–213, 1993. doi: 10.1111/j.1745-3984.1993.tb00423.x.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

A Appendix

A.1 RIFT Taxonomy Development Process

RIFT was developed using a grounded-theory methodology and constant-comparative refinement (Glaser & Strauss, 1967). Rather than beginning from a fixed rubric-quality schema, we iteratively derived failure modes from expert critiques of real rubrics. The development corpus contained 85 rubrics and 255 expert annotations drawn from five sources spanning human-authored and synthetic rubric-generation settings. The human-authored sources were AdvancedIF (He et al., 2025), which provides expert-authored criteria for complex instruction following, and ResearchRubrics (Sharma et al., 2026), which focuses on factual grounding and multi-step analytical reasoning for deep research agents. The synthetic sources were WildChecklists (Viswanathan et al., 2025), OpenRubrics (Liu et al., 2026), and AutoRubrics (Xie et al., 2025), which cover checklist-style requirements, contrastive rubric generation, and reward-oriented rubric induction.

The 85 rubrics were divided into four subsets and processed in four sequential refinement rounds. In the first round, three experts independently wrote open-ended *rubric critiques* for each rubric without access to a predefined taxonomy. These critiques were randomly shuffled, batched in groups of five, and provided to GPT-5.2 to propose an initial set of candidate failure modes. The LLM was used only as an aggregation and proposal mechanism: all proposed modes, names, definitions, and

Table 2: RIFT: RubrIc Failure Mode Taxonomy. Nine failure modes under two quality dimensions. Full definitions, decision rules, and pass/fail examples in Appendix B.

Dimension	Sub-dimension	Failure mode
Reliability	Ambiguity	**Subjective Language**: evaluative terms left unanchored to objective expectations **Ungrounded Verification**: checkable facts with no reference, source, or procedure
	Atomicity	**Non-Atomic**: bundled requirements prevent consistent partial credit
Content Validity	Under-Specification	**Missing Criteria**: a prompt requirement that no criterion evaluates
	Over-Specification	**Rigid**: stricter or narrower requirements than the prompt asks for **Redundant Criteria**: one behavior scored by two or more criteria
	Misspecification	**Hackability**: top score reachable by inflating proxies **Misaligned**: grades the wrong objective or embeds incorrect assumptions
	Miscalibration	**Miscalibrated Weighting**: secondary mass offsets failing the core requirement

boundaries were manually reviewed and revised by the expert panel before being used in subsequent rounds.

In later rounds, experts labeled the next subset of rubrics using the current taxonomy and assigned all applicable failure-mode labels. They also wrote additional rubric critiques for failures not covered by the current taxonomy and *taxonomy critiques* for problems in the taxonomy itself, such as unclear definitions, overlapping labels, missing distinctions, or structural inconsistencies. After each round, the panel reviewed the collected labels and critiques, refined the taxonomy, and then applied the updated version to the next subset. The process terminated when the panel agreed that additional rubrics were no longer producing new failure modes or substantive taxonomy revisions.

The resulting taxonomy, summarized in Tab. 2, organizes rubric failures into two dimensions. *Reliability* captures failures that make rubric judgments inconsistent or difficult to reproduce across graders. *Content Validity* captures failures in the relationship between rubric criteria and the intended evaluation target, including missing, overly narrow, misaligned, gameable, redundant, or misweighted criteria.

B RIFT: RUBRIC FAILURE MODE TAXONOMY

This section provides the complete descriptions of each failure mode in the RIFT taxonomy (Tab. 2), including detailed decision rules for annotation. Each failure mode includes: (1) when to apply the label, (2) how to determine whether a rubric exhibits the failure mode, and (3) boundary conditions specifying when *not* to apply the label and which alternative label to consider instead. For each failure mode, we also provide illustrative pass examples (rubrics that do *not* exhibit the failure mode) and fail examples (rubrics that *do* exhibit the failure mode). Note: in the "Do NOT apply" cross-references below, CSV-internal label identifiers have been replaced with the display names used in Tab. 2.

B.1 RELIABILITY FAILURES

Reliability failures lead to inconsistent grading across annotators or evaluation runs, reducing the reproducibility and trustworthiness of rubric-based evaluation.

Subjective. Apply when the rubric uses inherently subjective evaluative terms (e.g., "clear," "appropriate," "credible," "comprehensive," "professional," "engaging," "well-written," "good sources") and does NOT sufficiently anchor them with objective expectations.

How to determine:

- Identify criteria dominated by inherently subjective terms.
- Check whether the rubric provides ANY anchoring attempt such as:
  - concrete checklists ("includes X/Y/Z"),
  - measurable thresholds (word count, required sections, required elements),
  - examples/anti-examples of what qualifies vs does not qualify, or
  - explicit decision rules ("count as clear if it defines the term and gives one example").
- If the rubric relies primarily on grader judgment and provides no meaningful anchors, apply.

Important clarification:

- Do NOT apply if the rubric gives examples or non-trivial decision rules that explain what the subjective term means (even if the term is still somewhat subjective).

Do NOT apply if:

- The core issue is missing expected answers/tolerances or a bounded verification procedure for a groundable requirement (use *Ungrounded*).
- The requirement is entirely absent (use *Missing Criteria*).

Illustrative examples:

	Input Context	**Rubric**
PASS	Write a professional email declining a meeting.	*2 pts: Includes a decline + proposes an alternative time. 2 pts: Uses a greeting and sign-off. 1 pt: No negative or insulting language.*
FAIL	Summarize the study.	*10 pts: The summary is clear and sufficiently detailed.*

Non-Atomic. Apply when the rubric does not provide a parseable, consistently scorable structure OR uses bundled (non-atomic) criteria that prevent consistent partial credit.

Triggers (any sufficient):

Non-atomic (bundled) criteria

- One scored item bundles multiple independently scorable requirements with no partial-credit rule or separable sub-scores (e.g., "clear, comprehensive, accurate, and well-cited" as a single 10-pt item).

Do NOT apply when:

- Subparts are separately scored or the rubric provides explicit level anchors (e.g., "1 point each for A/B/C" or a 0–2 scale per dimension with definitions).
- The rubric is scorable but uses subjective language (use *Subjective*) or is missing requirements (use *Missing Criteria*).

Illustrative examples:

	Input Context	**Rubric**
PASS	Write a short answer with two supporting reasons.	*2 pts: Answers the question. 1 pt: Reason #1 supports the answer. 1 pt: Reason #2 supports the answer. 1 pt: Total length ¡=150 words.*
FAIL	Summarize the article.	*10 pts: Summary is clear, accurate, comprehensive, concise, and engaging.*

Ungrounded. Apply when the rubric requires verification that is plausibly groundable/boundable, but the rubric does not provide the necessary grounding (answer keys/acceptable variants/tolerances/decision rules) OR does not bound the verification procedure (what to check, how much to check, and how to judge conflicts).

How to determine (any sufficient):

(A) *Groundable determinate tasks lack grading anchors.* The task has a knowable target output given fixed inputs (e.g., extraction, classification, translation, math, SQL result, code output), but the rubric provides no expected answers, acceptable variants, label mappings, tolerances, or decision rules. The criterion may be clearly worded (e.g., "totals are correct"), yet graders still lack what they need to check correctness.

(B) *Open-world requirements lack bounded audit procedure.* The rubric demands broad verification (e.g., "all facts are true," "restaurants are open right now," "fully original/no plagiarism," "links work") without bounding: what to check (scope/sample size), which sources/tools are allowed, how to resolve conflicting evidence, and the pass/fail threshold.

(C) *Measurement standard is unspecified but could be made checkable.* The rubric requires a measurement that depends on an unspecified standard (e.g., "exactly 20 pages") without defining the rendering/formatting standard or offering a workable proxy.

Do NOT apply if:

- The requirement is simply missing from the rubric (use *Missing Criteria*).
- The main issue is subjective wording without anchors (use *Subjective*). The rubric provides a representative list of examples to demonstrate expected content.

Illustrative examples:

	Input Context	**Rubric**
PASS	Extract all email addresses from the text.	*1 pt per correct email address; accepted forms include plus-addressing. Gold list of emails: a@x.com, b.y@z.org, ... Deduct 1 pt per missing email.*
FAIL	Compute the correct totals for these 30 invoices.	*10 pts: Totals are correct.*

B.2 CONTENT VALIDITY FAILURES

Content validity failures arise when rubric criteria are misaligned with the intended evaluation target, either by grading the wrong objective or by failing to cover essential requirements.

Misaligned or Rigid. Apply when the rubric (a) grades the wrong objective for the prompt or embeds incorrect assumptions, OR (b) imposes unnecessarily strict/narrow requirements not asked for by the prompt or reasonably inferred from the prompt, predictably penalizing prompt-faithful high-quality answers.

How to determine (any applies):

- Wrong task / shifted objective: makes non-requested deliverables mandatory for points.
- Incorrect embedded assumptions: assumes a context not in the prompt (jurisdiction, audience, tools, constraints) and scores accordingly.
- Penalizes good practice: scores down reasonable caveats/uncertainty/safety practices when the prompt does not forbid them.
- Arbitrary brittleness/over-constraint: mandates a specific tool/library/method/structure/formatting or false precision when multiple reasonable alternatives would satisfy the prompt.

Do NOT apply when:

- The prompt itself imposes the strictness at any point (e.g., exact JSON keys, or a direct instruction from the user earlier in a chat conversation).
- The requirement is missing entirely (use *Missing Criteria*).
- The main problem is internal contradiction (use *Self-Contradictory*).
- The main problem is rubric-level proxy gaming (use *Hackable*).

Illustrative examples:

	Input Context	Rubric
PASS	Write Python code to parse CSV.	*5 pts: Correct parsing. 3 pts: Handles quoted commas. 2 pts: Includes brief usage example. (Does not mandate pandas vs csv module.)*
FAIL	Write a haiku about winter.	*5 pts: Includes at least 5 academic citations. 5 pts: Uses APA format reference list.*

Missing Criteria. Apply when the prompt implies at least one checkable must-have requirement, but the rubric provides no criterion that allows a grader to evaluate that requirement at all.

How to determine:

1. List the prompt's core requirements:
   - required deliverables/components,
   - must/must-not constraints,
   - required format/ordering/sections,
   - and genre-critical qualities the prompt clearly expects (e.g., functional correctness for code; "two sentences"; "valid JSON"; "include 10 items"; "chronological order").
2. For each requirement, check whether ANY rubric criterion covers it.
3. If one or more requirements have no corresponding criterion, apply.

Do NOT apply if:

- The rubric mentions the requirement but is vague or subjective (use *Subjective*).
- The rubric mentions the requirement but it cannot be graded consistently due to missing keys/tolerances/bounded audit steps (use *Ungrounded*).
- The rubric grades a different task or adds arbitrary constraints (use *Misaligned or Rigid*).

Illustrative examples:

	Input Context	Rubric
PASS	Return ONLY valid JSON with keys: name (string) and age (integer).	*3 pts: Output parses as JSON. 2 pts: Contains exactly keys name and age. 2 pts: name is a string; age is an integer. 1 pt: No surrounding commentary.*
FAIL	Write a 200-word email and include a subject line.	*10 pts: Tone is professional. 5 pts: Grammar and spelling are correct.*

Miscalibrated Weighting. Apply when the rubric's point distribution does not reflect the relative importance of requirements in the prompt. A response can earn an implausibly high score despite failing a core task requirement, or lose disproportionate credit for a secondary requirement.

How to determine:

1. **Core-failure retention:** Identify a requirement essential to completing the prompt, then calculate how much score a response can retain if it completely fails that requirement while satisfying everything else.

2. **Compensating credit:** Check whether points from secondary, formatting, or supporting criteria can offset failure on the primary task objective.
3. **Relative importance:** Compare each criterion's weight to its importance in the prompt. Core requirements should materially affect the achievable score; secondary requirements should not dominate it.
4. **Whole-rubric sanity check:** Compare a substantively successful answer with one that misses a major requirement. The rubric should produce a meaningful score difference between them.

Do NOT apply if:

- The criterion evaluates the wrong objective or assumes context not present in the prompt (use *Misaligned*).
- The objective is appropriate, but the criterion demands one unnecessarily narrow method or answer path (use *Rigid*).
- A necessary requirement is not evaluated anywhere in the rubric (use *Missing Criteria*).
- The problem is that one criterion combines multiple independently scorable requirements (use *Non-Atomic*).
- The rubric can be gamed through proxy signals or superficial compliance, regardless of its numerical weights (use *Hackability*).
- A low-weight requirement is genuinely secondary to the prompt, even if an external reference or gold answer treats it as important.

Illustrative examples:

	Input Context	**Rubric**
PASS	Write a decision memo on whether to freeze Palisade for external review.	*5 pts: Recommends holding the freeze because the merge-time guardrail failed. 3 pts: Reports the 128 KiB median error. 1 pt: Correct filename.*
FAIL	Write a decision memo on whether to freeze Palisade for external review.	*5 pts: Correct filename. 5 pts: Reports the 128 KiB median error. 5 pts: Clear headings. 1 pt: Recommends holding the freeze.*

B.3 CONSEQUENTIAL VALIDITY FAILURES

Consequential validity failures reduce the downstream usefulness and discriminative power of rubric-based evaluation, even when individual criteria may be well-defined.

Hackable. Apply when the rubric is gameable at the rubric level: a responder could easily achieve a top score by inflating proxy metrics (length, number of bullets/sections/items/citations/examples/brands, repeated keywords) without materially improving correctness, relevance, or fulfillment of the prompt—and the rubric lacks strong quality gates that tie points to substantive, prompt-aligned success.

Core question (required):

- Could I easily achieve full marks on this rubric while still not satisfying the prompt requirements or producing a low-quality response?

How to determine (any sufficient):

- Most points come from "more" ($\geq$N tips/citations/examples/pros/cons/sections) while relevance, non-duplication, correctness, and prompt-specific success conditions are weakly specified or absent.
- Rewards merely asserting attributes ("quiet," "fast Wi-Fi," "no fees") without requiring evidence, linkage to the task, or checks against duplication.

- Counting proxies dominate while key prompt requirements have only weak gates (e.g., no requirement that citations support specific claims; no requirement that items be distinct and on-topic).

Do NOT apply when:

- Quantity minimums are paired with robust quality controls that make padding ineffective (e.g., each item must be non-duplicative, tied to a specific claim or user need, and verifiably grounded/bounded).
- The main issue is that the rubric is generic and doesn't discriminate at all (use *Low Signal*).
- The main issue is a specific criterion that shifts the task or overconstrains acceptable answers (use *Misaligned or Rigid*).

Illustrative examples:

	Input Context	**Rubric**
PASS	Provide 5 study tips.	*1 pt each for 5 tips that are (a) non-duplicative and (b) each includes a concrete example of how to apply it.*
FAIL	Provide a recommendation.	*5 pts: At least 10 pros. 5 pts: At least 10 cons. (No check for relevance or duplication.)*

C PROMPTS

This section provides the prompt templates used in the RIFT pipeline. Template variables are denoted with double curly braces (e.g., {{`variable`}}).

C.1 TAXONOMY REFINEMENT PROMPT

The following prompt template drives the iterative taxonomy refinement process described in Section 4 (Tab. 2). During each iteration of the grounded theory pipeline, expert annotator feedback—comprising open-ended rubric critiques and taxonomy critiques—is batched and provided to GPT-5.2 alongside the current taxonomy state. The model proposes refinements (merges, additions, clarifications, splits, removals, or renames), which are then reviewed and finalized by the expert panel.

```
You are an expert at analyzing rubric quality feedback and refining
failure mode taxonomies.  Your task is to output a complete refined
failure mode taxonomy.

## Original Failure Mode Taxonomy

This is the original taxonomy before any refinements in this session:
[For each failure mode: label and description. If none defined: "No failure modes have been defined
yet."]

## Current Running Refinement

This is the taxonomy as refined so far in this session (may be
identical to original if this is the first batch):
[For each failure mode: label, description, rationale, and counts of pass/fail examples. If none refined:
"No refinements have been made yet."]

## Annotator Feedback to Analyze

Below are annotations with two types of critiques:
- Rubric Critique:  Issues the annotator observed in the rubric that
were NOT captured by the original taxonomy labels (may suggest new
```

```
failure modes)
- Taxonomy Critique:  Critique of the ORIGINAL taxonomy (unclear
definitions, overlapping categories, missing categories, etc.).  Note:
these critiques were written against the original taxonomy, not the
running refinement.

[For each annotation in the batch:]

Annotation {{loop.index}}
Input Context:  {{item.input_context}}
Rubric:  {{item.rubric}}
Rubric Critique:  (issues not captured by original taxonomy)
{{item.rubric_critique or "None provided"}}
Taxonomy Critique:  (critique of the original taxonomy)
{{item.failure_mode_critique or "None provided"}}

[End of annotation loop]
```

The prompt further specifies a detailed **taxonomy philosophy** and **refinement guidelines** that constrain the model's proposed changes:

```
## Taxonomy Philosophy

CRITICAL: This taxonomy will be used by human annotators.  The primary
goal is to create a taxonomy that is:

     • Compact: Aim for 7–10 total failure modes. Fewer distinct categories is ALWAYS better than
       many granular ones.
     • Easily distinguishable: A human should be able to distinguish between any two fail-
       ure modes in under 30 seconds. If two categories require careful reading to tell apart, they
       should be merged or their distinction should be clarified by refining the label names and or the
       description.
     • Actionable: Each category must be clearly applicable without ambiguity.

Consolidation over proliferation:  When in doubt, MERGE rather than
add.  Two failure modes that are 80% similar should become one
category, not two.  The cost of a slightly imperfect merge is far
lower than the cost of a bloated, hard-to-use taxonomy.

## Guidelines

     • Clear descriptions: Each failure mode description must be clear, specific, and action-
       able. The description should explicitly specify HOW to determine if a rubric exhibits this failure
       mode. An annotator should be able to read the description and confidently apply it to any rubric.
     • No overlapping failure modes: The taxonomy should not contain failure modes
       with overlapping meanings. If two labels capture the same concept, merge them or refine them
       to make them distinct. Do NOT add a new failure mode if its meaning already exists under a
       different label.
     • Self-contained rationales: Each rationale must be a self-contained justification that
       will be used for manual review. It should explain WHY this failure mode exists, what evidence
       from critiques supports it, and how it differs from other failure modes.  A reviewer should
       understand the rationale without needing to see the original critiques.
     • Cumulative applicability: The refined taxonomy must be applicable to ALL critiques
       that have been seen in this session (including previous batches), not just the current batch.
       Do not remove or change failure modes in ways that would make them inapplicable to earlier
       critiques that supported them.


## Task
```

```
Analyze BOTH the rubric critiques and taxonomy critiques above.
Before adding any new failure modes, first consider whether existing
categories should be merged.

FIRST: Consider merging existing failure modes when:

     • Two or more categories have similar descriptions or capture
       closely related issues
     • Categories are difficult to distinguish without careful reading
     • A broader category could capture multiple narrower ones without
       losing important distinctions
     • The taxonomy has grown beyond 12 failure modes

PREFERRED action - merge:  Combine overlapping, redundant, or closely
related labels into one.  This is the most important refinement
action.  If you're unsure whether two categories are distinct enough,
merge them.

Add new failure modes ONLY when ALL of the following are true:

     • The issue is clearly NOT capturable by ANY existing failure
       mode (even with minor rewording)
     • The issue appears in MULTIPLE critiques (not just one
       annotation)
     • The new category is easily distinguishable from ALL existing
       categories
     • Adding it would NOT push the taxonomy beyond 12 failure modes

Other refinement actions:

     • clarify:  Make a label's description clearer, more specific,
       or more actionable (especially clarifying HOW to identify the
       failure mode)
     • split:  Divide an overly broad label into more specific ones
       (use sparingly---only when a category is genuinely too broad to
       apply consistently)
     • remove:  Eliminate labels that are not useful, are duplicates,
       or are too similar to other categories
     • rename:  Change a label name to be more descriptive

Output:
1.  failure_modes:  The complete list of failure modes after applying
changes.  Each failure mode should have:

     • label:  concise identifier (e.g., contradictory_criteria,
       missing_edge_cases)
     • description:  clear, specific, and actionable description that
       explains HOW to determine if a rubric has this failure mode
       (what to look for, what conditions must be met)
     • rationale:  a self-contained justification for this failure
       mode that can be understood without seeing the original
       critiques.  Explain why it exists, what patterns it captures,
       and how it differs from related failure modes.  If this is a
       NEW category, explicitly explain why it cannot be captured by
       any existing category.
     • examples:  REQUIRED: 3-5 pass_examples AND 3-5 fail_examples
       for each failure mode.  Multiple diverse examples are essential
       for annotator training.  You may use real examples from the
       annotations or synthesize clear illustrative examples.  Each
       example should illustrate a distinct scenario or nuance.

2.  changes_summary:  A list of strings describing what changes
you made (e.g., "Added 'contradictory_criteria' based on rubric
```

```
critiques", "Clarified description of 'ambiguous_criterion'", "Merged
'x' and 'y' into 'z'")

If no changes are needed based on these critiques, return the current
running refinement unchanged with an empty changes_summary.
```

C.2 Controlled Intervention Prompts

The following templates specify the model-backed operators. The task, full indexed rubric, and target indices are supplied in the user message:

```
The task the rubric evaluates:
---
{{instruction}}
---

The rubric ({{num_criteria}} criteria):
{{index}}: {{criterion}}
[Repeated for every criterion.]

Target indices: {{target_indices}}
[For Non-Atomic: Target pairs to merge: {{target_pairs}}.]
```

Each system prompt identifies one RIFT failure mode, includes its taxonomy definition, and instructs the model to act as a research tool that injects that known defect so its downstream effects can be measured. All prompts prohibit changes outside the supplied targets and prohibit mentioning the rubric, transformation, or instructions in criterion text. The operator-specific instructions are:

```
Rewrite each target criterion so that it evaluates roughly the same
aspect of the response but through inherently subjective, unanchored
language: replace objective expectations, named quantities, exact
values, and verifiable conditions with subjective evaluative terms
(``clear'', ``appropriate'', ``well-written'', ``high quality'',
``sufficiently detailed'', ``reasonable''). The rewritten criterion
must remain a single true/false statement about the response, but one
whose verdict now depends on the grader's taste. Keep each rewrite
on the same topic as the original---degrade its precision, not its
subject.
```

Subjective Language.

```
Rewrite each target criterion by stripping its grounding details while
keeping the demand for factual verification: remove reference values,
expected answers, named sources, formulas, tolerances, time scopes,
and verification procedures, so the criterion still asks the grader to
check correctness but no longer says against what. Keep each rewrite
a single true/false statement demanding factual verification---do not
turn it into a subjective quality judgment.
```

Ungrounded Verification.

```
Merge each target pair into ONE bundled criterion that demands both
original conditions at once, entangled so they cannot be scored
separately: join the requirements with conjunctions or clause chains
```

```
(``and'', ``while also'', ``as well as'') into a single pass/fail
statement where partial fulfillment has no consistent verdict.
Preserve the substance of both originals---the defect is bundling,
not lost content.
```

Non-Atomic.

```
Replace each target criterion with an unjustifiably rigid version
of the same underlying requirement:  demand one exact phrasing,
format, structure, ordering, length, or method where the task permits
many valid ones.  A response could fully satisfy the task and still
fail the criterion on the arbitrary constraint.  Keep the original
requirement recognizable and keep each replacement a single, judgeable
true/false statement.
```

Rigid.

```
For each target criterion, write a paraphrased duplicate that
evaluates exactly the same underlying requirement with the same
pass/fail boundary, but is worded differently enough that the
duplication is not obvious from a casual read.  The duplicate will
be inserted alongside the original, so the same behavior is scored
twice.  Any response passing one must pass the other.
```

Redundant Criteria.

```
Replace each target criterion with a misaligned version of its
requirement:  grade an objective the task did not ask for, or embed
a factually or contextually incorrect assumption about the task.  The
replacement should stay superficially on-topic and professional.  The
misalignment must be substantive:  a grader following it would reward
or penalize the wrong thing for this task.  Keep each replacement a
single, judgeable true/false statement.
```

Misaligned.

```
Rewrite each target criterion so its decision rule keys on an easily
inflated surface proxy instead of the underlying requirement:
presence of keywords or section headers, response length or count
thresholds, use of hedging or citation-like formatting, or mentioning
a term rather than being correct about it.  A responder who games
the proxy should pass without actually satisfying the original
requirement.  Keep the proxy superficially related to the original
and express it as a single, judgeable true/false statement.
```

Hackability. The five rewrite operators and Redundant Criteria return:

```
{"rewrites":  [{"index":  <target index>, "criterion":  "<rewritten
criterion>"}, ...]}
```

Non-Atomic returns:

```
{"merges": [{"indices": [<index a>, <index b>], "criterion":
"<merged criterion>"}, ...]}
```

The response must contain exactly one entry per target and no prose or code fences. Missing Criteria and Miscalibrated Weights are deterministic and therefore have no model prompt.

D Cross-Judge Robustness Check

§5.1's loudest signatures hackable, rigid, and misaligned, all driven by a mean CVR collapse and are scored throughout by a single judge, GPT-5.5. To rule out the effect being an artifact of that particular scorer, we rescored the clean baseline and each of these three modes at ρ=0.5 (240 rubrics total) with an independent 3-judge panel a `claude-opus-4-8`, a `claude-haiku-4-5`, and a `gemini-3.1-pro`, sharing no model or provider with GPT-5.5. We scored 716 of 720 planned units (99.4%; the remainder failed on a reasoning-budget limit or a malformed tool call and were not retried further).

Table 3 compares the paired Cohen's d (variant − clean) obtained under this panel against the GPT-5.5 values already reported in Table 4. All three effects replicate in direction and in comparable magnitude, each significant at $p<10^{-20}$: the mean CVR collapse under hackable, rigid, and misaligned is not an artifact of the scoring judge.

Table 3: Paired Cohen's d for the mean CVR collapse (variant − clean, n=60/mode) under GPT-5.5 versus an independent 3-judge panel (`claude-opus-4-8`, `claude-haiku-4-5`, `gemini-3.1-pro`).

Mode	GPT-5.5	3-judge panel	p
Hackable	−2.83	−2.51	2.6×10^{-27}
Rigid	−2.04	−1.89	2.9×10^{-21}
Misaligned	−1.85	−2.24	8.5×10^{-25}

E Full Per-Mode Signatures

Table 4 reports every FDR-significant signal for each of the eight criterion-level failure modes at ρ=0.5, expanding the heatmap of Figure 3 into exact effect sizes. Modes are grouped by the family the data produces rather than by a grouping we imposed: the misspecification modes collapse content validity while leaving criterion counts intact, the structural modes move counts while content validity holds, and the ambiguity modes produce only a faint relevance dip.

Two signals, `contamination_score` and `adjacent_pct`, do not reach significance for any mode and are reported as honest nulls. Ungrounded verification is the sole mode without a strong signature: its largest effect is $|d|$=0.32, and the classifier of §5.1 recovers it at 0.78 recall only from the joint pattern of a mild relevance dip occurring *without* the accompanying content-validity collapse.

Table 4: Per-mode signatures at ρ=0.5 (n=60 per mode). Paired Cohen's d (variant − baseline); all entries shown are FDR-significant at $q < .05$ over the full 8×17 grid. k is the number of significant signals for that mode. AUC is the unpaired Mann–Whitney P(variant $>$ baseline), where 0.5 indicates a signal blind to the corruption.

Mode	k	Signal	d	Note
Misspecification: content validity collapses, criterion counts unchanged				
Hackable	10	mean_cvr	−2.83	AUC .13, p=5 × 10^{-30}; loudest effect in the corpus
		weighted_relevance	−2.07	
		item_relevance	−1.94	
		candidate_item_mean	−1.94	
		core_pct	−1.17	AUC .29
		general_pct	+0.72	Mass drifts outward from the task domain
		irrelevant_pct	+0.46	
		num_flags	+0.41	
		background_pct	+0.38	
		representativeness	−0.38	
Rigid	7	mean_cvr	−2.04	AUC .15
		weighted_relevance	−0.55	
		core_pct	−0.39	
		item_relevance	−0.38	
		candidate_item_mean	−0.38	
		representativeness	−0.38	
		general_pct	+0.35	
Misaligned	6	mean_cvr	−1.85	AUC .18
		weighted_relevance	−0.73	
		item_relevance	−0.59	
		candidate_item_mean	−0.59	
		irrelevant_pct	+0.46	Distinguishes misaligned from rigid
		core_pct	−0.43	
Structural: criterion counts move, content validity intact				
Redundant	6	total_items	+2.30	Mean +7.3 criteria
		items_bad	+1.41	
		items_review	+1.07	
		items_good	+0.65	
		redundancy_score	+0.64	Fires for this mode and no other
		weighted_relevance	+0.31	
Non-atomic	5	total_items	−2.47	Mean −3.9 criteria
		weighted_relevance	+1.50	Inversion: merged criteria appear *more* relevant
		items_review	−0.90	
		items_good	−0.73	
		core_pct	+0.32	
Missing criteria	5	total_items	−2.30	Mean −7.3 criteria
		items_bad	−1.43	
		items_review	−1.22	
		items_good	−0.67	
		representativeness	−0.38	Surviving criteria stay individually healthy; only composition shifts
Ambiguity: neither pattern, only a faint relevance dip				
Subjective language	5	weighted_relevance	−0.67	
		item_relevance	−0.62	
		candidate_item_mean	−0.62	
		general_pct	+0.46	
		core_pct	−0.41	mean_cvr is *not* significant here (−0.09)
Ungrounded verification	2	weighted_relevance	−0.32	$\Delta = -0.006$
		num_flags	−0.31	$\Delta = -0.22$; all other signals silent

F CORRUPTION-LEVEL RESPONSE

Figure 5 and Table 5 report how each signature scales with the amount of corruption applied. Every unit is regressed on its *realized* corruption fraction rather than the requested level, since the operators do not always achieve the requested fraction exactly.

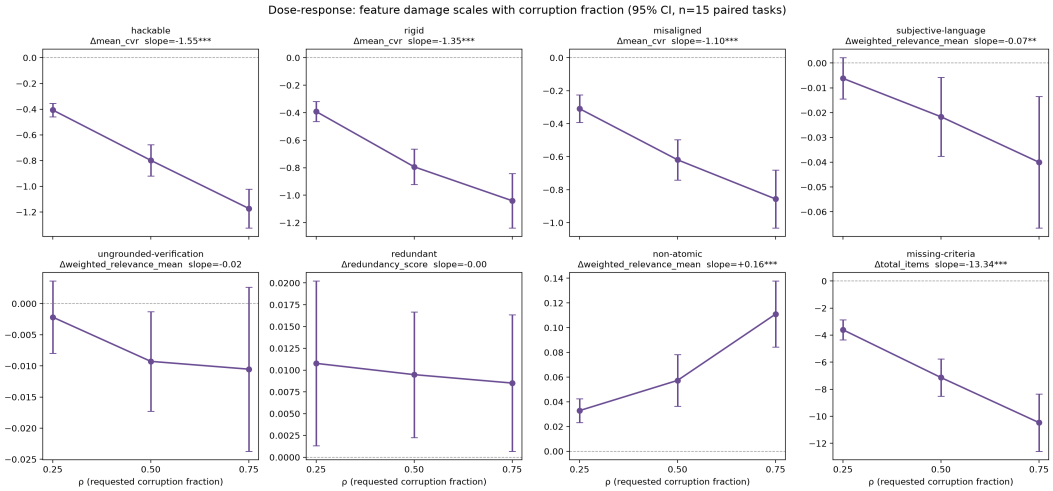

Figure 5: Signal response against realized corruption fraction ρ (95% CI, n=15 paired tasks per level). The CVR-based signals and `total_items` descend steadily with corruption, while `redundancy_score` and ungrounded verification's weighted relevance stay flat across the full range.

Table 5: OLS slope of Δsignal on realized ρ over the 15-task subset (3 corruption levels, 45 paired points per mode). Per-family slopes give the same regression fit within the GDPval and Terminal-Bench subsets independently. Realized corruption fractions average 0.267, 0.528, and 0.766 at the three requested levels, with non-overlapping ranges.

Mode	**Signal**	**Slope**	r	p	**GDPval / TB**
Graded: slope estimates severity					
Hackable	Δmean_cvr	-1.55	$-.83$	1.6×10^{-12}	-1.83 / -1.21
Rigid	Δmean_cvr	-1.35	$-.73$	1.6×10^{-8}	-1.62 / -1.01
Misaligned	Δmean_cvr	-1.10		5×10^{-7}	
Missing criteria	Δtotal_items	-13.3		$< .001$	Mechanical
Non-atomic	Δweighted_relevance	$+0.159$	$+.67$	$< .001$	Inverted
Subjective language	Δweighted_relevance	-0.073		$< .01$	
Weakly graded: detects presence, estimates severity poorly					
Hackable	Δcore_pct	-0.23	$-.35$	$< .05$	
Flat: presence only					
Redundant	Δredundancy_score	-0.005		n.s.	Saturates
Ungrounded verif.	Δweighted_relevance	-0.018		n.s.	Stealthy at every level

The division is functional rather than following the boundaries between signal families. Content validity behaves as a *dosimeter* for the misspecification modes, so the magnitude of the drop estimates how much of the rubric has been compromised. The domain-mapping signals behave as *presence detectors*: `core_pct` registers that hackability has been introduced but grades its extent only weakly. The cross-lens signals catch the modes that invert the logic of the other two, most visibly non-atomicity, whose weighted relevance climbs as more criteria are merged. That inversion has a consequence beyond this corpus: any rubric-quality measure built on mean per-criterion relevance can be improved by consolidating criteria rather than by improving them.

G TASK-ADAPTED PERSONA PANELS: EXAMPLES

Each task's panel is generated once from the task instruction, never from the rubric, yielding five one-line expert profiles that are persisted and reused unchanged across the baseline and every injected condition, and across both the CVR panel and the cross-lens expert panel (the latter substitutes a neutral base prompt for the CVR-specific one). Every persona shares the task's base prompt and differs only in its profile line, so the five raters vary in *vantage point* while receiving an identical task and response format.

Table 6 reproduces the generated panels for two GDPval and two Terminal-Bench tasks. Two regularities recur across the corpus. First, panels span complementary roles rather than five near-synonyms: practitioner, specialist, compliance, audit, and stakeholder perspectives appear consistently, adapted to the domain. Second, the generator tracks task genre. GDPval deliverables elicit professional-stakeholder personas, whereas Terminal-Bench tasks elicit verification-oriented roles keyed to the task's contract, each phrased around what a rubric must check.

Table 6: Generated persona panels for four tasks. Each panel is produced once per task from the instruction alone and reused across all experimental conditions. Profile lines are verbatim; task descriptions are our one-line summaries.

GDPval: EMS cardiac quality-assurance review. *A paramedic reconstructs county EMS STEMI cases for a medical-direction QA report.*

EMS medical director	Focused on whether STEMI recognition, alert activation, and protocol adherence are clinically justified by the available field data.
Cardiac QA paramedic	A senior QA reviewer assessing reconstructed EMS timelines, ECG timing, documentation consistency, and field decision-making.
Interventional cardiologist	Evaluates whether prehospital STEMI suspicion aligns with ECG findings, ED assessment, and cath lab outcomes.
Protocol compliance officer	Determines whether ECG acquisition timing and STEMI alert decisions meet county EMS standards.
Documentation auditor	Checks that conclusions are supported by reconciled evidence from PCRs, monitor exports, hospital notes, and protocols.

GDPval: B2B SaaS release retrospective. *A QA lead analyses defect leakage after Release 4.7 and recommends test-plan changes.*

QA retrospective lead	Experienced in post-release defect leakage retrospectives for B2B SaaS products.
Test coverage analyst	Focused on mapping requirements, executed test cases, automation inventory, and production defects.
Customer impact reviewer	Evaluates whether defect severity and customer impact are accurately represented.
Automation strategy manager	Assesses whether proposed automation candidates are justified by escaped defects and implementation effort.
VP of quality governance	Judges whether the retrospective is direct, actionable, and credible enough for executive review.

Terminal-Bench: `rollsnap` Go CLI. *Implement a standard-library-only CLI producing a deterministic, dependency-first deployment snapshot.*

CLI contract reviewer	Whether criteria are necessary to verify the CLI follows the specified command-line behavior and documented contract.
Deterministic rollout validator	Whether criteria are essential for confirming dependency-first deterministic snapshot generation.
Input resilience auditor	Whether criteria are necessary to judge tolerance of recoverable corrupt rollout mirror inputs and issue reporting.
Standard-library enforcer	Whether criteria are essential for enforcing standard-library-only implementation and prohibited imports.
Output artifact checker	Whether criteria are essential for validating parent-directory creation and exact JSON/TSV output artifacts.

Terminal-Bench: signed-receipt parser differential (CTF). *Probe an HTTP service under a request budget and forge a signed receipt that verifies as authorized.*

HTTP probe recon analyst	Whether criteria are necessary for correctly probing the service, using `/verify` as ground truth, and respecting request-count constraints.
Parser disagreement specialist	Whether criteria are necessary for identifying the canonical parser disagreement channel and exploiting divergent claim interpretation.
Receipt forgery reviewer	Whether criteria are necessary for producing a forged signed receipt that verifies as authorized while parsers disagree.
Output schema validator	Whether criteria are necessary for writing stage JSON files with exact key order, digest formats, and required paths.
Grader compliance auditor	Whether criteria are necessary for satisfying grader-visible constraints such as action-log limits, required probes, and persistence across stages.