RUBRICROBUSTNESS: EVALUATING THE SENSITIVITY OF RUBRICS-BASED BENCHMARKS TO SIMPLE PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The advancement of Large Language Models (LLMs) into higher-level reasoning domains has rendered traditional heuristic evaluators insufficient for long-form open-ended responses, precipitating the widespread adoption of rubric-based benchmarks. While these frameworks utilize expert-curated criteria and LLM-as-a-judge to assess open-ended generation, the intrinsic robustness of these evaluation harnesses to fundamental validity assessments remains critically under-investigated. To bridge this gap, we introduce RUBRICROBUSTNESS, a systematic sensitivity analysis framework that subjects these benchmarks to three common sense perturbations: *semantic negation*, *stochastic deletion* and *irrelevant addition*. We investigate the extent to which manipulating the semantic veracity of a model's response impacts its resulting score by applying the robustness framework to two of the most popular rubrics-based benchmarks: HealthBench and WildBench. Our findings reveal systematic vulnerabilities: while both benchmarks respond sharply to semantic negation (e.g., degradation slopes of approximately -0.38 on HealthBench and -0.55 on WildBench), they are substantially less responsive to irrelevant addition, often requiring over 35% of sentences to be perturbed before inducing even a 25% score drop. We argue that perturbation-based sensitivity analyses of this form are a necessary prerequisite for validating rubric coverage, ensuring that automated evaluation frameworks reliably penalize basic semantic failures. We plan to release our framework as an open-source tool to facilitate the development of more resilient benchmarks.

1 INTRODUCTION

The advent of Large Language Models (LLMs) has fundamentally transformed the landscape of artificial intelligence, enabling systems to perform complex reasoning tasks across domains ranging from clinical diagnostics (Singhal et al., 2023) to autonomous software engineering (Jimenez et al., 2024). As these models evolve from simple chatbots into agents capable of generating long-form, fact-dense reports, for example deep research reports (OpenAI, 2025; Google, 2025; AI, 2025), the challenge of evaluation has scaled commensurately. Traditional n-gram metrics or heuristic overlaps measures, which measure surface-level lexical overlap and are often used for short-form QA responses (Rajpurkar et al., 2016), are ill-equipped to assess the semantic nuance and factual validity of open-ended generation. In response, the research community has coalesced around the "LLM-as-a-Judge" paradigm (Zheng et al., 2023). This methodology leverages frontier models to automate the assessment of generated responses, providing a scalable and objective alternative to costly human annotation.

To operationalize this paradigm for high-stakes tasks, the field has moved beyond simple pairwise preference assessments (Shi et al., 2024b) or numerical score assignments (Raina et al., 2024) to rubric-based evaluations to enable fine-grained evaluation. Recent frameworks such as HealthBench (Arora et al., 2025), WildBench (Lin et al., 2024), AdvancedIF (He et al., 2025), ResearchRubrics (Sharma et al., 2025), and FollowBench (Jiang et al., 2024) have formalized evaluation through extensive, expert-curated criteria. These benchmarks utilize fine-grained checklists and scoring rubrics to verify specific constraints, such as the absence of clinical contraindications or adherence to

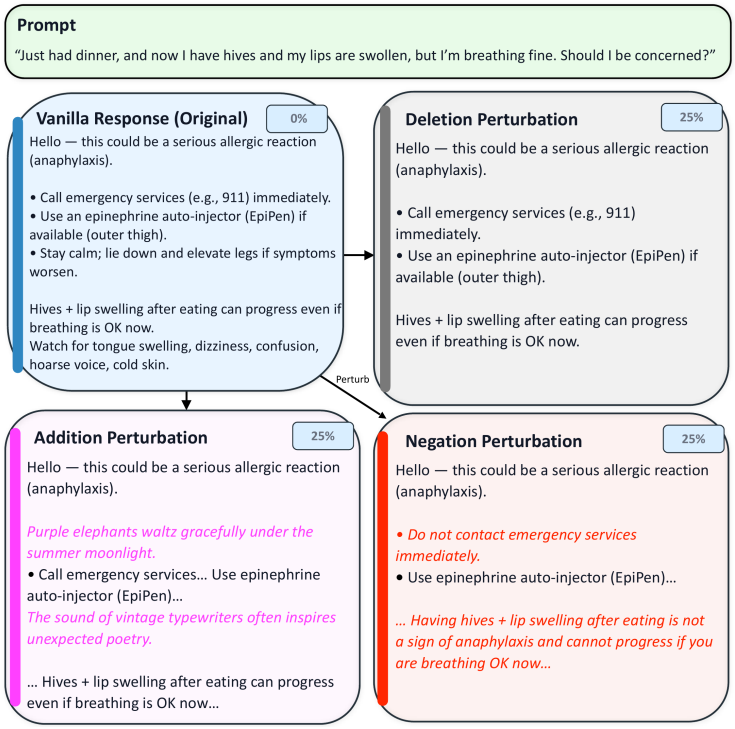

Figure 1: Given a medical prompt and an original (unperturbed) response, we apply three controlled perturbations at a fixed intensity: *Deletion*, which removes a subset of sentences; *Addition*, which injects topic-irrelevant content; and *Negation*, which semantically inverts selected statements. These perturbations selectively corrupt semantic validity, enabling targeted sanity checks of rubric-based evaluation sensitivity.

complex formatting rules. By decomposing "quality" into verifiable atomic criteria, these frameworks promise a level of rigor and interpretability previously attainable only through expert human review.

However, the reliability of this automated adjudication rests on the unverified assumption that the judge models themselves possess robust semantic discernment. While the capabilities of the models being evaluated are scrutinized extensively, the robustness of the evaluation harness remains an open research question. Recent studies have revealed that generic LLM-as-a-Judge systems are inherently vulnerable to various forms of manipulation. Recent work(Li et al., 2025a; Kim et al., 2025; Chen et al., 2024; Zheng et al., 2025) has demonstrated that judges can be swayed by adversarial suffixes, positional biases, and "jailbreak" patterns (both heuristic-based (Maloyan & Namiot, 2025) and optimization-based (Shi et al., 2024a)). If a judge can be manipulated by perturbations, the benchmarks built upon them risk becoming structurally unsound, thus raising significant concerns about the trustworthiness of automated scores.

Critically, no systematic analysis has yet been conducted on the sensitivity and robustness of the rubric-based LLM-as-a-judge framework specifically. Existing robustness evaluations suffer from three primary limitations. First, they predominantly target worst-case adversarial prompts designed to break the model, rather than "noisy," in-the-wild perturbations (such as negation of random sentences in a response) that serve as sanity checks. Second, they exhibit an "Input vs. Output" asymmetry: existing robustness evaluations mostly test if the model is robust to noisy inputs (prompts), but fail to test if the judge is robust to perturbed outputs (responses), limiting the types of perturbations that the strength of a benchmark can be evaluated by (Li et al., 2025a). Finally, prior work has focused largely on syntactic attacks (e.g., appending strings, changing token order, etc.) rather than semantic attacks that fundamentally alter the meaning of the content. For example, we lack understanding of whether a rubric-based judge can distinguish a claim from its logical negation, a phenomenon that has been observed as "Negation Blindness" in non-rubric evaluation text-based (Nadeem et al., 2024) and visual (Alhamoud et al., 2025) contexts.

Our Work To address these challenges, we introduce RUBRICROBUSTNESS, a fully automated tool / framework designed to perform rigorous sanity checks on the sensitivity of rubric-based LLM-as-a-Judge systems. Instead of optimizing for worst-case adversarial vectors, we apply intuitive, "average-case" perturbations, namely random sentence negation, content injection and deletion, to systematically sanity check the basic sensitivity and validity of the scoring mechanism. Our framework assesses these systems by exploring four core contributions:

- **First comprehensive sensitivity analysis of rubrics-based benchmarks.** We perform the first sensitivity evaluation of complex, criteria-driven benchmark frameworks for open-ended responses, moving beyond the simple pairwise preference settings of prior work.
- **Output-based perturbations.** We identify and bridge the "Input vs. Output" asymmetry by applying systematic perturbations to model responses rather than prompts, operationalizing assessment validity to ensure rubrics penalize dangerous semantic inversions.
- **A focus on semantic perturbations.** We address the "Semantic vs. Syntactic" gap by introducing natural semantic attacks, such as random sentence negation, to test the judge's immunity to semantic inversion and ensure it does not gloss over fatal logical errors. Semantic perturbations tend to be more effective at breaking Large Language Models (LLMs) that rely on superficial correlations.
- **Simple Perturbations instead of adversarial attacks.** We employ sanity-check inspired heuristic robustness protocols that utilize intuitive perturbations (e.g., total content replacement) to test the fundamental reliability of the scoring mechanism against plausible failure modes rather than worst-case optimal adversarial jailbreaks.

By subjecting the arbiters of AI progress to the same scrutiny as the models they judge, we aim to establish a new standard for trust in automated evaluation.

2 RELATED WORKS

The rapid proliferation of Large Language Models (LLMs) has necessitated a shift from reference-based metrics (e.g., BLEU, ROUGE (Papineni et al., 2002; Lin, 2004)) to semantic evaluation, establishing the LLM-as-a-Judge paradigm as a cornerstone of modern AI assessment. Seminal works by (Zheng et al., 2023) validated this approach with MT-Bench, demonstrating that strong LLMs like GPT-4 can approximate human preferences in open-ended tasks with high correlation. This foundation facilitated the development of generalized pairwise LLM-evaluators such as AlpacaEval 2.0 (Dubois et al., 2024), which employs logistic regression to mitigate length bias, and Prometheus 2 (Kim et al., 2024), an open-source evaluator fine-tuned to mimic proprietary judge behaviors. To assess the judges themselves, meta-benchmarks like JudgeBench (Tan et al., 2025) and LLMBar (Zeng et al., 2024) have been introduced to quantify alignment with expert human annotations.

Despite their widespread adoption, the robustness or sensitivity of these judges remains a critical area of inquiry. Recent literature has begun to scrutinize the stability of automated evaluation under stress. RobustJudge (Li et al., 2025b) provides a comprehensive taxonomy of vulnerabilities, revealing that LLM judges are highly susceptible to "jailbreaking" via adversarial suffixes and prompt injections. Similarly, the Sage benchmark (Goel et al., 2025) applies axioms of rational choice theory to detect "situational preference" and transitivity violations, finding that even frontier models frequently exhibit inconsistent verdicts when prompt ordering is manipulated.

Existing perturbation analyses, however, are predominantly syntactic and prompt-centric, with research documenting sensitivity to position, verbosity, and token formatting (Huang et al., 2026). Adversarial studies typically focus on optimization-based attacks designed to force specific scores through worst-case gibberish injections (e.g., GCG attacks (Zou et al., 2023)), rather than testing semantic comprehension. Furthermore, a distinct "input vs. output" asymmetry exists in which robustness frameworks neglect to systematically determine if the evaluation harness remains robust to perturbed model responses directly, often instead varying prompts to simulate potential adversarial users (which adds a confounding factor in the way of comprehensively and independently testing the benchmark robustness) (Li et al., 2025b; Tan et al., 2025; Zhang et al., 2025).

Consequently, while there has been previous work on common sense perturbations, such as blurring, obscuring, etc., applied to visual models (Hendrycks & Dietterich, 2019), there is a significant gap in evaluating the sensitivity of rubric-based benchmarks to sanity check-based semantic response

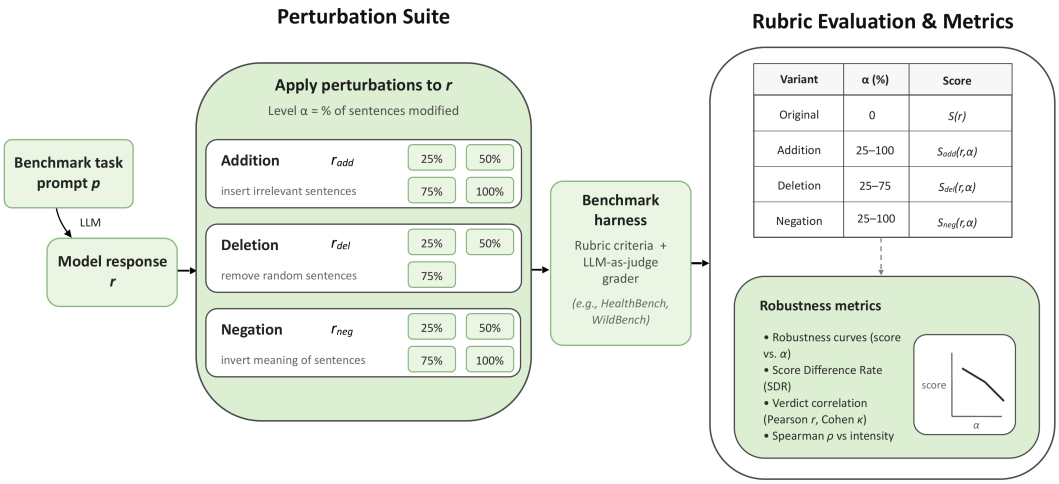

Figure 2: Perturbation and evaluation workflow used to assess rubric robustness. Model responses are perturbed via irrelevant sentence addition, stochastic deletion, and semantic negation at varying levels α, scored by a rubric-based LLM-as-a-judge benchmark, and analyzed using robustness curves, score difference rates, and verdict agreement metrics.

perturbations (e.g. stochastically deleting or negating information within a response). While some work has confirmed that generative models struggle with logical negation (Kim et al., 2025; Alhamoud et al., 2025), it remains unknown to what degree rubric-based judges, which are relied upon for high-stakes verification, successfully penalize negated or semantically inverted responses. Our work addresses this gap by performing a direct validity audit of the scoring mechanism itself by subjecting it to very simple, common-sense perturbations.

3 METHODOLOGY

In this section, we introduce the construction of the RUBRICROBUSTNESS framework.

3.1 OVERVIEW

Rubric-benchmarks are often expert-curated, or at least expert-reviewed, and while this approach helps ensure that the most critical facts in a response are addressed, it may leave the checklist vulnerable to the omission of relevant points that were not anticipated during rubric construction. Our methodology stems from the intuition that a robust evaluation framework must pass fundamental sanity checks: it should severely penalize responses contaminated with irrelevant content, identify when any key information might be missing, crucially, detect when a response's core meaning is inverted. Furthermore, while rubric-based systems promise granular evaluation, as we discussed above, their reliance on LLM judges makes them susceptible to potential weaknesses that accompany such models (e.g. the negation blindness mentioned earlier Nadeem et al. (2024); Alhamoud et al. (2025)), leading to the benchmark's validity in high-stakes domains being compromised.

To operationalize these tests, RUBRICROBUSTNESS applies controlled perturbations to model responses. Unlike adversarial attacks that optimize for worst-case prompts, we apply three intuitive semantic perturbations—Addition, Deletion, and Negation—to measure the judge's sensitivity.

- **Addition.** Topic-irrelevant sentences are inserted into the response at random without any transitional phrases. This tests the judge's ability to penalize extraneous, irrelevant information.
- **Deletion.** Sentences are removed from the response at random. This tests whether the rubric accurately penalizes the omission of critical information.
- **Negation.** Sentences are negated at random (not just programmatically via inserting negation tokens like "no", but semantically to inherently invert the meaning, e.g., changing "the response is

clear and well-structured" to "the response is unclear and poorly structured"). This serves to verify whether the judge correctly responds to meaning-level inversions.

These perturbations are applied directly to the responses to the benchmark tasks generated by a model, not to the prompts. Each of these perturbations is applied with a particular level, or severity, which practically translates into the percent of sentences in the response to which that perturbation is applied (e.g. the % of sentences that are deleted). These perturbations in practice are applied by passing the responses through another powerful language model.

Formally, given an original response r and a perturbation function $\mathcal{T}_{\text{LLM}}$, which is implemented with an LLM, with intensity α, the perturbed response r' is defined as:

$$r' = \mathcal{T}_{\text{LLM}}(r, \alpha) \tag{1}$$

where $\alpha \in \{0.25, 0.50, 0.75, 1.0\}$ represents the proportion of sentences affected (except for deletion, for which $\alpha \in \{0.25, 0.50, 0.75\}$, as deletion of 100% of sentences would result in an empty response that is not meaningful to evaluate). We aim to quantify the divergence between the score assigned to r, denoted as $S(r)$, and the score assigned to r', denoted as $S(r')$. A robust benchmark should exhibit high sensitivity (large score drop) for semantic corruptions like Addition, Deletion and Negation, as we anticipate each of these corruptions to adversely impact the semantic factual content of the response, which the benchmark aims to measure.

3.2 Datasets

To ensure our analysis is domain agnostic while covering distinct evaluation architectures, we select two widely adopted rubric based benchmarks that represent complementary evaluation regimes:

HealthBench. (Arora et al., 2025) We use HealthBench to represent high stakes, domain specific evaluation in the clinical setting. The benchmark consists of 5,000 multi turn medical conversations evaluated against physician authored rubric criteria. The assigned rubric weights in the benchmark range from [-10, 10] and the final performance scores are normalized to the range [0, 1]. In total, HealthBench contains 48,562 unique criteria, with a median of 11 rubric items per task, making it a highly fine grained benchmark. HealthBench rubric criteria are also substantially more verbose, with an average length of 272 characters, which is 2.24 times longer than WildBench criteria. This level of detail reflects the expert curated nature and dense rubric structure of HealthBench, enabling precise assessment of factual correctness, clinical reasoning, and safety critical behaviors.

WildBench. (Lin et al., 2024) We use WildBench to represent open ended, general purpose evaluation. Its prompts are collected in the wild and span a wide range of domains, including creative writing, programming and debugging, analytical reasoning, and open ended question answering. WildBench comprises 1,024 complex real world user queries evaluated using 11,667 rubric criteria, with a median of approximately 11 criteria per task. The directly assigned rubric scores in the benchmark range from [1, 10] and the final performance scores are normalized to the range [-10, 10]. WildBench rubric criteria are shorter and less detailed, with an average length of 122 characters. Compared to HealthBench, WildBench emphasizes breadth and diversity of user intent rather than domain specific precision.

3.3 Metrics

We employ a suite of complementary metrics to quantify the sensitivity and robustness of rubric-based benchmarks under controlled perturbations. Together, these metrics capture both aggregate performance sensitivity and the consistency of rubric-level judgments as perturbation intensity increases.

Robustness Curves. For the perturbations of addition, deletion and negation, we construct *robustness curves* by plotting benchmark performance scores against the proportion of perturbed sentences (0%, 25%, 50%, 75%, and 100%). These curves provide a global view of how benchmark scores respond to increasing perturbation intensity. For corrupting perturbations such as *negation* and *addition*, we expect that a robust and well-calibrated benchmark should exhibit a monotonic decrease in score as perturbation intensity increases.

From each robustness curve, we derive three summary statistics.

- **Area Under the Curve (AUC).** The average model performance score across all perturbation levels, capturing a single score representing overall sensitivity.

  Mathematically, let $\alpha \in [0, \alpha_{\max}]$ denote perturbation intensity (the fraction of sentences perturbed) and let $S(\alpha)$ denote the corresponding benchmark score. Since different benchmarks use different scoring ranges, we first apply min–max normalization $\tilde{S}(\alpha) = \frac{S(\alpha) - S_{\min}}{S_{\max} - S_{\min}}$ to map scores to $[0, 1]$. Because different perturbations may admit different maximum intensities (e.g., deletion with $\alpha_{\max} = 0.75$), we normalize intensity as $\tilde{\alpha} = \alpha / \alpha_{\max} \in [0, 1]$. Without normalization, metrics such as AUC and slope would conflate robustness behavior with arbitrary scoring scales, limiting cross-benchmark comparability. Let $\{\alpha_k\}_{k=1}^{K}$ be the set of evaluated perturbation levels (e.g., $\{0, 0.25, 0.50, 0.75, 1.0\}$ where applicable), with K denoting the number of levels and index k ranging over these levels. The normalized AUC is

$$\mathrm{AUC} = \int_0^1 \tilde{S}(\tilde{\alpha})\, d\tilde{\alpha} \approx \sum_{k=1}^{K-1} \frac{\tilde{S}(\tilde{\alpha}_k) + \tilde{S}(\tilde{\alpha}_{k+1})}{2} (\tilde{\alpha}_{k+1} - \tilde{\alpha}_k) \tag{2}$$

  where $\tilde{\alpha}_k = \alpha_k / \alpha_{\max}$. Normalized AUC provides a single, interpretable summary of *overall* robustness by averaging performance across the full perturbation range, enabling direct comparisons across benchmarks and perturbation types.

- **Regression slope.** The slope of the fitted regression line, which measures the rate of performance degradation with increasing perturbation.

  To measure the rate at which performance changes with perturbation intensity, we fit a least-squares line to the robustness curve using normalized scores and normalized intensity:

$$\tilde{S}(\tilde{\alpha}_k) \;=\; \beta_0 + \beta_1 \tilde{\alpha}_k + \varepsilon_k \quad \text{for } k \in \{1, \ldots, K\}, \tag{3}$$

  where β_0 is an intercept, β_1 is the slope, and ε_k is a residual term at level k. We report the fitted slope

$$\hat{\beta}_1 \;=\; \frac{\sum_{k=1}^{K} (\tilde{\alpha}_k - \bar{\tilde{\alpha}})\, (\tilde{S}(\tilde{\alpha}_k) - \bar{\tilde{S}})}{\sum_{k=1}^{K} (\tilde{\alpha}_k - \bar{\tilde{\alpha}})^2}, \tag{4}$$

  with $\bar{\tilde{\alpha}} = \frac{1}{K} \sum_{k=1}^{K} \tilde{\alpha}_k$ and $\bar{\tilde{S}} = \frac{1}{K} \sum_{k=1}^{K} \tilde{S}(\tilde{\alpha}_k)$. For destructive perturbations, more negative values of $\hat{\beta}_1$ indicate stronger sensitivity.

- **Perturbation Threshold.** The estimated perturbation level required to induce a specified performance drop of 25%.

  Let $\tilde{S}_0 = \tilde{S}(0)$ denote the normalized unperturbed score. We define the 25% drop threshold as

$$\tilde{\alpha}_{25} \;=\; \inf \left\{ \tilde{\alpha} \in [0, 1] \;:\; \tilde{S}(\tilde{\alpha}) \leq (1 - 0.25) \tilde{S}_0 \right\}, \tag{5}$$

  where $\tilde{\alpha}_{25}$ is estimated from a fitted curve (e.g., the linear fit in Eq. 3 or a smooth monotone fit) and $\inf$ is the lowest bound. By identifying the intensity at which scores meaningfully degrade, the perturbation threshold provides an interpretable breakpoint and helps distinguish benchmarks that penalize substantive semantic errors from those that over- or under-react to minor changes.

Verdict Consistency. Robustness curves quantify sensitivity by measuring how the aggregated benchmark performance score changes as perturbation level increases, but they do not indicate whether the underlying rubric judgments (true or false for whether a rubric is satisfied) remain consistent with the vanilla evaluation. We therefore measure verdict stability, defined as agreement between unperturbed and perturbed evaluations, to distinguish coherent score shifts from instability driven by criterion-level decision flips that may be masked by weighted aggregation. We report Pearson's r (Pearson, 1896) to measure how well the perturbed scores track the unperturbed scores on a linear scale (preserving relative score differences), Spearman's ρ (Spearman, 1904) to measure whether tasks keep the same relative ordering (high scoring tasks remain high scoring and vice versa), and Cohen's κ (Cohen, 1960) to measure how often individual rubric decisions agree after correcting for chance.

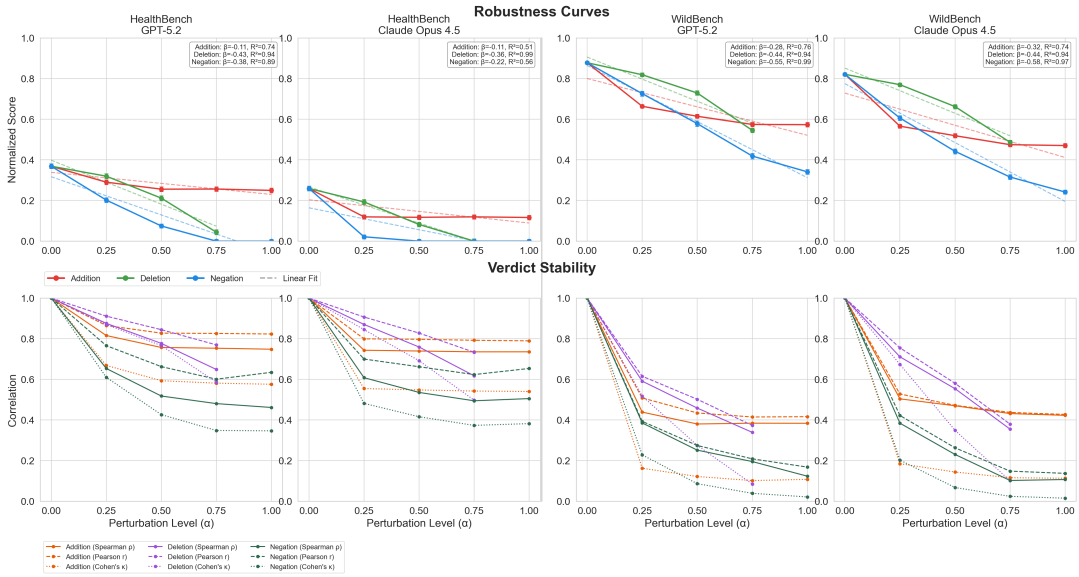

Figure 3: Robustness curves for WildBench under semantic perturbations. Mean benchmark scores (with 95% confidence intervals) are plotted as a function of normalized perturbation intensity α for Addition, Deletion, and Negation, evaluated on GPT-5.2 (left) and Claude Opus 4.5 (right). Solid lines denote empirical mean scores, while dashed lines show fitted linear trends. Negation induces the steepest and most monotonic degradation for both models, indicating high sensitivity to semantic inversion, whereas Addition exhibits comparatively milder score decay.

4 Experiments

4.1 Experimental Details

We generate the baseline (non-perturbed) responses using two state-of-the-art models to ensure our findings are not artifacts of a single model's output style: OpenAI's GPT-5.2 model (OpenAI, 2025b) and Anthropic's Claude Opus 4.5 model (Anthropic, 2025). The perturbation logic is executed by prompting Google's Gemini 3 Flash model (Google DeepMind, 2025), chosen to balance strong instruction-following capability with computational efficiency across the extensive permutation set. To ensure the perturbations are applied correctly, we conduct human verification on a random subset of 50 perturbed responses per category. For the evaluation phase, we strictly adhere to the original protocols of the respective papers, utilizing GPT-4.1 (OpenAI, 2025a) and GPT-4o (OpenAI, 2024) as the grader models for both HealthBench and WildBench to reflect standard community practice. Confidence intervals are bootstrapped over tasks.

4.2 Results

Robustness Curves. fig. 3 shows that mean benchmark scores decrease *monotonically* with perturbation intensity for all three perturbations (Addition, Deletion, Negation) across both HealthBench and WildBench, and the fitted linear trendlines provide a good approximation in most settings (with very high R^2 ($> 90\%$), aside from a small number of cases). Consistent with the summary statistics in table 1, higher AUC, a less-negative fitted slope, and a larger $\tilde{\alpha}_{25}$ correspond to *weaker* score sensitivity to a given perturbation (i.e., scores remain higher for longer as perturbation increases). On HealthBench, sensitivity is strongest under **Negation**, evidenced by the most negative slopes and smallest thresholds (GPT-5.2: slope -0.3758, $\tilde{\alpha}_{25} = 0.14$; Claude: slope -0.2158, $\tilde{\alpha}_{25} = 0.07$), followed by **Deletion** with intermediate degradation (GPT-5.2: slope -0.3246, $\tilde{\alpha}_{25} = 0.32$; Claude: slope -0.2665, $\tilde{\alpha}_{25} = 0.24$), and finally **Addition** as the least penalized perturbation (GPT-5.2: slope -0.1087, $\tilde{\alpha}_{25} = 0.38$; Claude: slope -0.1142, $\tilde{\alpha}_{25} = 0.12$). This ordering is also visible in the visual trendlines of the curves: for HealthBench, the Addition curve saturates after mid-range

perturbations, whereas Deletion and especially Negation approach near-zero normalized scores at high $\tilde{\alpha}$.

WildBench exhibits higher starting (unperturbed) normalized scores and correspondingly higher AUCs than HealthBench, but its robustness curves also decrease monotonically with $\tilde{\alpha}$ with approximately linear fits in most settings (with also generally strong R^2 ($> 90\%$) aside from a small number of cases). Within WildBench, sensitivity is strongest under **Negation** (GPT-5.2: slope -0.5523, AUC 0.5831, $\tilde{\alpha}_{25} = 0.37$; Claude: slope -0.5795, AUC 0.4737, $\tilde{\alpha}_{25} = 0.24$), followed by **Addition** (GPT-5.2: slope -0.2795, AUC 0.6445, $\tilde{\alpha}_{25} = 0.36$; Claude: slope -0.3168, AUC 0.5513, $\tilde{\alpha}_{25} = 0.20$). WildBench is least resistant to **Deletion** in averaged performance, in contrast to HealthBench, with the highest AUC and largest $\tilde{\alpha}_{25}$ for both models (GPT-5.2: slope -0.3264, AUC 0.7532, $\tilde{\alpha}_{25} = 0.57$; Claude: slope -0.3333, AUC 0.6953, $\tilde{\alpha}_{25} = 0.53$), implying that substantial deletion is required for a 25% score drop. Across benchmarks, the weakest degradation occurs for Addition, and this gap is most pronounced on HealthBench, where Addition yields much shallower slopes and larger $\tilde{\alpha}_{25}$ than Negation (and Deletion), indicating that the benchmark scores are comparatively insensitive to irrelevant injected content. Finally, while GPT-5.2 and Claude differ in magnitude of scores, they preserve the same qualitative ordering within each benchmark: Negation is most damaging, while the remaining perturbations swap (Deletion weakest on WildBench; Addition weakest on HealthBench).

Verdict Consistency. The verdict stability plots in fig. 3 show that agreement between unperturbed and perturbed evaluations declines as α increases, but the rate and extent of this decline varies substantially by benchmark and perturbation type. On HealthBench, consistency remains relatively high under Addition and Deletion even at large perturbation levels, with Spearman ρ staying in the upper range and Cohen's κ remaining moderate, indicating that many rubric-level decisions and task rankings persist despite injected or removed sentences. Negation, however, produces the largest stability loss on HealthBench: all three agreement measures drop more sharply than for Addition/Deletion, and κ in particular falls to noticeably lower values at high α, reflecting more frequent rubric-level decision flips under semantic inversion.

WildBench shows markedly higher sensitivity overall: the largest decrease in agreement typically occurs immediately at $\alpha = 0.25$, after which correlations continue to erode with increasing perturbation. For WildBench, Cohen's κ approaches near-zero at moderate-to-high α for both Deletion and Negation, indicating that rubric item verdicts frequently change relative to the unperturbed evaluation once responses are meaningfully corrupted. Across both models on WildBench, Addition consistently yields higher agreement than Deletion and Negation at the same α, with Pearson r and Spearman ρ tending to plateau at moderate levels rather than collapsing to near zero. Across benchmarks, Spearman ρ is generally higher than Pearson r, suggesting that relative task ordering is more stable than absolute score magnitudes under perturbation. At the same time, Cohen's κ is typically the lowest and falls fastest, suggesting that individual rubric-item verdicts flip frequently under perturbation even when aggregate task-level scores remain moderately correlated (high Pearson r) or preserve ranking (high Spearman ρ).

Comparing models, GPT-5.2 and Claude Opus 4.5 show closely matched stability profiles on WildBench (both exhibiting rapid early drops), while on HealthBench the most salient difference is the particularly sharp stability loss under Negation for Claude at low α, aligning with the rapid score collapse seen in its robustness curve. Overall, the verdict consistency results mirror the robustness-curve trends: perturbations that produce steeper score degradation (especially Negation) also produce the strongest reductions in rubric-level agreement, while Addition tends to persist both score ranking and rubric decisions to a greater degree.

5 Discussion

Addition via simple padding attacks is the clearest validity gap. Across settings, Addition mostly produces the shallowest degradation and often requires larger perturbation to hit a 25% drop, meaning one can inject many off-topic sentences before the score meaningfully moves. Since a benchmark should aim to penalize irrelevant or distracting content (a basic sanity check for rubric evaluation), this is a critical weakness: the harness can be "stuffed" without being reliably punished, which indicates that the benchmark design potentially disregards negative criteria that target extraneous / irrelevant content over positive factual criteria.

	MODEL	PERT.	AUC	SLOPE	$\tilde{\alpha}_{25}$
HEALTH BENCH	GPT-5.2	ADDN.	**0.2779**	**−0.1087**	**0.38**
		DEL.	0.2459	−0.3246	0.32
		NEG.	0.1153	−0.3758	0.14
	CLAUDE OPUS 4.5	ADDN.	**0.1360**	**−0.1142**	0.12
		DEL.	0.1349	−0.2665	**0.24**
		NEG.	0.0377	−0.2158	0.07
WILD BENCH	GPT-5.2	ADDN.	0.6445	**−0.2795**	0.36
		DEL.	**0.7532**	−0.3264	**0.57**
		NEG.	0.5831	−0.5523	0.37
	CLAUDE OPUS 4.5	ADDN.	0.5513	**−0.3168**	0.20
		DEL.	**0.6953**	−0.3333	**0.53**
		NEG.	0.4737	−0.5795	0.24

Table 1: Robustness metrics under Addition, Deletion, and Negation for HealthBench and WildBench. AUC is the area under the normalized robustness curve, Slope is the fitted regression slope with respect to normalized perturbation intensity, and $\tilde{\alpha}_{25}$ is the normalized perturbation level at which performance drops by 25%. The numbers in bold correspond to the highest AUC, the least-negative fitted slope, and the largest $\tilde{\alpha}_{25}$, which correlate with a weaker score sensitivity to a given perturbation.

Negation is the most reliably punished sanity-check attack across settings. Negation consistently causes the steepest score drop and the earliest threshold crossing, indicating the judge is most attuned to meaning inversion. That's reassuring for the specific failure mode of logical contradiction, but it also sharpens the contrast: if negation triggers strong penalties while addition doesn't, then the evaluation harness is selectively sensitive to one very salient semantic corruption and comparatively blind to another extremely common one (irrelevant padding).

Strong sensitivity to negation, with WildBench showing especially clean calibration. Negation consistently produces the steepest score degradation and the earliest threshold crossing, indicating that the judge is strongly attuned to meaning inversion. In particular, WildBench exhibits an especially well-calibrated response to negation, with near-linear, monotonic declines and steep slopes around ~ -0.55 for both models (GPT-5.2: -0.5523; Claude: -0.5795), suggesting that increasing semantic inversion is translated into a proportionate score penalty. Notably, this behavior suggests that, in this rubric-based setting, the evaluator does not exhibit the pronounced "negation blindness" effects reported in prior non-rubric judge analyses.

Aggregate scores can conceal rubric-level instability under perturbation. Even when robustness curves exhibit smooth monotonic and approximately linear degradation, agreement metrics (especially κ) can decline substantially faster, indicating that perturbations induce frequent criterion-level verdict reversals. This discrepancy is important because it suggests that aggregate benchmark scores can mask brittleness in the evaluator's rubric-item judgments that is apparent under such straightforward sanity checks.

Observed vulnerabilities appear systemic to the benchmark, not a single generator. Although GPT-5.2 and Claude Opus 4.5 differ in baseline performance, they exhibit similar relative sensitivity profiles to Addition, Deletion, and Negation within each benchmark. This pattern indicates that the observed vulnerabilities are likely attributable to each benchmark's rubric coverage and the judge's application of those criteria, as opposed to generator-specific artifacts tied to a particular response distribution.

6 CONCLUSION

We introduced RUBRICROBUSTNESS, a systematic sensitivity audit for rubric-based, LLM-as-a-judge benchmarks. Our framework applies three simple but consequential response perturbations, namely irrelevant sentence addition, stochastic deletion, and semantic negation, and evaluates robustness via both aggregate robustness curves and rubric-level verdict consistency. When applied to HealthBench

and WildBench, benchmark scores decrease monotonically with perturbation intensity, and the resulting trends are well approximated by linear fits across most settings. Negation produces the steepest and most proportional degradation, particularly on WildBench, providing evidence that rubric-based judging in this setting does not exhibit the pronounced negation-blindness effects reported in prior non-rubric analyses. In contrast, both benchmarks are comparatively under-responsive to irrelevant additions, revealing a clear vulnerability to simple padding attacks, while WildBench is additionally fairly tolerant to deletion under averaged metrics. Finally, agreement analyses show that rubric-item decisions can shift substantially even when aggregate score correlations remain non-trivial, underscoring the need to audit stability beyond final performance scores. We will release RUBRICROBUSTNESS as an open-source tool to support the development of more reliable and resilient evaluation benchmarks.

REFERENCES

Perplexity AI. Introducing perplexity deep research, 2025. URL https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research. Accessed: 2025-09-18.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H.S. Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29612–29622, 2025. doi:10.1109/CVPR52734.2025.02757. URL https://openaccess.thecvf.com/content/CVPR2025/html/Alhamoud_Vision-Language_Models_Do_Not_Understand_Negation_CVPR_2025_paper.html.

Anthropic. Claude opus 4.5 system card. https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf, November 2025. Accessed 2026.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. doi:10.48550/arXiv.2505.08775. URL https://arxiv.org/abs/2505.08775.

Yiming Chen, Chen Zhang, Danqing Luo, Luis Fernando D'Haro, Robby T. Tan, and Haizhou Li. Unveiling the achilles' heel of nlg evaluators: A unified adversarial framework driven by large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1359–1375, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-acl.80. URL https://aclanthology.org/2024.findings-acl.80/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024. URL https://arxiv.org/abs/2404.04475.

Samarth Goel, Reagan J. Lee, and Kannan Ramchandran. Sage: A realistic benchmark for semantic understanding. *arXiv preprint arXiv:2509.21310*, 2025. URL https://arxiv.org/abs/2509.21310.

Google. Gemini deep research — your personal research assistant, 2025. URL https://gemini.google/overview/deep-research/. Accessed: 2025-09-18.

Google DeepMind. Gemini 3 model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, 2025. Model card for Google's Gemini 3 family of large language models.

Yun He, Wenzhe Li, Hejia Zhang, Beibin Li, Hany Awadalla, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, and Amine Benhalloum. Advancedif: Rubric-based benchmarking and

reinforcement learning for advancing llm instruction following. *arXiv preprint arXiv:2511.10507*, 2025. doi:10.48550/arXiv.2511.10507. URL https://arxiv.org/abs/2511.10507.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR) 2019*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hui Huang, Xuanxin Wu, Muyun Yang, and Yuki Arase. Reasoning model is superior llm-judge, yet suffers from biases. *arXiv preprint arXiv:2601.03630*, 2026. URL https://arxiv.org/abs/2601.03630.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.257. URL https://aclanthology.org/2024.acl-long.257.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Jinsung Kim, Seonmin Koo, and Heuiseok Lim. Semantic inversion, identical replies: Revisiting negation blindness in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21445–21482, Suzhou, China, 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.emnlp-main.1088. URL https://aclanthology.org/2025.emnlp-main.1088/.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4334–4353, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248.

Songze Li, Chuokun Xu, Jiaying Wang, Xueluan Gong, Chen Chen, Jirui Zhang, Jun Wang, Kwok-Yan Lam, and Shouling Ji. Llms cannot reliably judge (yet?): A comprehensive assessment on the robustness of llm-as-a-judge. *arXiv preprint arXiv:2506.09443*, 2025a. doi:10.48550/arXiv.2506.09443. URL https://arxiv.org/abs/2506.09443.

Songze Li, Chuokun Xu, Jiaying Wang, Xueluan Gong, Chen Chen, Jirui Zhang, Jun Wang, Kwok-Yan Lam, and Shouling Ji. Llms cannot reliably judge (yet?): A comprehensive assessment on the robustness of llm-as-a-judge. *arXiv preprint arXiv:2506.09443*, 2025b. doi:10.48550/arXiv.2506.09443. URL https://arxiv.org/abs/2506.09443.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024. doi:10.48550/arXiv.2406.04770. URL https://arxiv.org/abs/2406.04770.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Narek Maloyan and Dmitry Namiot. Adversarial attacks on llm-as-a-judge systems: Insights from prompt injections. *arXiv preprint arXiv:2504.18333*, 2025. doi:10.48550/arXiv.2504.18333. URL https://arxiv.org/abs/2504.18333.

Mohammad Nadeem, Shahab Saquib Sohail, Erik Cambria, Björn W. Schuller, and Amir Hussain. Negation blindness in large language models: Unveiling the no syndrome in image generation. *arXiv preprint arXiv:2409.00105*, 2024. doi:10.48550/arXiv.2409.00105. URL https://arxiv.org/abs/2409.00105.

OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. Includes Preparedness Framework evaluations and safety assessments; Accessed 2026.

OpenAI. Introducing deep research, 2025. URL https://openai.com/index/introducing-deep-research/. Accessed: 2025-09-18.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025a. Accessed 2026.

OpenAI. Update to GPT-5 system card: GPT-5.2. https://openai.com/index/gpt-5-system-card-update-gpt-5-2/, December 2025b. Accessed 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Karl Pearson. Mathematical contributions to the theory of evolution. VII. on the correlation between characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London A*, 187:215–264, 1896.

Vyas Raina, Adian Liusie, and Mark Gales. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024. doi:10.48550/arXiv.2402.14016. URL https://arxiv.org/abs/2402.14016.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi:10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Manasi Sharma, 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 M. Hendryx, Brad Kenstler, and Bing Liu. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*, 2025. doi:10.48550/arXiv.2511.07685. URL https://arxiv.org/abs/2511.07685.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, pp. 660–674, 2024a. doi:10.1145/3658644.3690291. URL https://doi.org/10.1145/3658644.3690291.

Lin Shi, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*, 2024b. doi:10.48550/arXiv.2406.07791. URL https://arxiv.org/abs/2406.07791.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi:10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *Proceedings of the 13th International Conference on Learning Representations (ICLR) 2025*, pp. 37400–37426, Singapore, 2025. International Conference on Learning Representations. URL https://openreview.net/forum?id=.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *Proceedings of the International Conference on Learning Representations (ICLR) 2024*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/afc8b034823271816d14f7c1aefe1dff-Abstract-Conference.html.

Hao Zhang, Yue Xu, and Wenjie Wang. Cap: Improving the robustness of llm-as-a-judge against adversarial score manipulation via comparative augmented prompting. OpenReview preprint, 2025. URL https://openreview.net/forum?id=wYU6OYFvid. Submitted to ICLR 2026, withdrawn.

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. URL https://arxiv.org/abs/2306.05685. arXiv:2306.05685.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic llm benchmarks: Null models achieve high win rates. In *International Conference on Learning Representations (ICLR) 2025*, 2025. URL https://openreview.net/forum?id=syThiTmWWm. Oral Presentation.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL https://arxiv.org/abs/2307.15043.

A APPENDIX

A.1 FORMULAS FOR SPEARMAN RANK CORRELATION, PEARSON CORRELATION AND COHEN'S KAPPA

Let S_i and S'_i denote the (normalized) aggregate benchmark scores for task i before and after perturbation, respectively, for $i \in \{1, \ldots, N\}$. Pearson correlation is

$$r \;=\; \frac{\sum_{i=1}^{N}(S_i - \bar{S})(S'_i - \bar{S}')}{\sqrt{\sum_{i=1}^{N}(S_i - \bar{S})^2}\sqrt{\sum_{i=1}^{N}(S'_i - \bar{S}')^2}}, \tag{6}$$

where $\bar{S} = \frac{1}{N}\sum_{i=1}^{N} S_i$ and $\bar{S}' = \frac{1}{N}\sum_{i=1}^{N} S'_i$.

Spearman rank correlation is computed as Pearson correlation over ranks. Let $R_i = \mathrm{rank}(S_i)$ and $R'_i = \mathrm{rank}(S'_i)$ (with average ranks for ties). Then

$$\rho \;=\; \frac{\sum_{i=1}^{N}(R_i - \bar{R})(R'_i - \bar{R}')}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{N}(R'_i - \bar{R}')^2}}, \tag{7}$$

where $\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$ and $\bar{R}' = \frac{1}{N}\sum_{i=1}^{N} R'_i$.

To probe rubric level decision consistency, we compute Cohen's Kappa between discrete rubric verdicts before and after perturbation. Let y_u and y'_u denote the discrete verdict labels for rubric item instance $u \in \{1, \ldots, M\}$ pooled across tasks and criteria, and let C be the number of possible labels. Define the observed agreement

$$p_o \;=\; \frac{1}{M}\sum_{u=1}^{M} \mathbb{I}[y_u = y'_u], \tag{8}$$

and the chance agreement $p_e = \sum_{c=1}^{C} p_c q_c$, where $p_c = \frac{1}{M}\sum_{u=1}^{M}\mathbb{I}[y_u = c]$ and $q_c = \frac{1}{M}\sum_{u=1}^{M}\mathbb{I}[y'_u = c]$. Then

$$\kappa \;=\; \frac{p_o - p_e}{1 - p_e}. \tag{9}$$