

# Agreement Without Alignment: Label-Free Rubric Clarification Increases LLM-Judge Consensus but Not Agreement with Humans

Sangdo Han, Moonsoo Kim  
CEO Business School  
{sangdohan, mskim}@ceobizschool.kr

## Abstract

LLM judges are increasingly deployed as panels, and disagreement among panel members is often read as a symptom of an unclear rubric. We test a label-free procedure that turns this disagreement into rubric repairs. The procedure decomposes each rubric into sub-criteria, ranks them by cross-judge disagreement on a development split, and has an editor model rewrite only the most contested ones. An additional stage that moves text-checkable sub-criteria to deterministic checkers transferred none in our evaluation. Human labels are never used to select or edit clauses; they are used only to evaluate the outcome. We pre-specified four hypotheses in a version-controlled protocol and evaluated six rubric conditions with a panel of five judges from three providers on SummEval, the HealthBench meta-evaluation, and USR Topical-Chat, for a total of 129,770 judge requests, excluding retries. On SummEval, the procedure raised inter-judge Krippendorff’s  $\alpha$  by 0.069 (95% CI 0.046 to 0.092) over the original rubric and by 0.074 over a control that applied the same edits to an equal number of randomly chosen clauses that were not selected. Agreement with expert ratings showed no detectable change (Kendall  $\tau$  difference -0.007, 95% CI -0.037 to 0.021). On HealthBench and USR, neither inter-judge agreement nor human agreement met the pre-registered effect criteria. In a preregistered secondary comparison, human agreement after editing the contested clauses was lower than after editing the control clauses on all three datasets in point estimate, although no interval excluded zero. Selecting clauses by disagreement makes judges agree with each other more than editing arbitrary clauses does, but in our setting the added consensus did not bring the judges closer to human annotators. We will release code, frozen rubrics, item identifiers, and judge scores.

## 1 Introduction

Rubric-based LLM judges now score a large share of open-ended model output, from summaries to medical conversations (Liu et al., 2023; Arora et al., 2025). Because a single judge can be idiosyncratic, practitioners increasingly combine several judges, often from different providers, into a panel. When the panel disagrees, the disagreement is usually averaged away. A second reading is that disagreement marks the places where the rubric admits more than one interpretation, and that those places can be found and fixed.

If that reading is right, it suggests an attractive workflow: locate the contested clauses using only judge outputs, clarify them, and obtain a more reliable rubric without collecting new human labels. Three findings from prior work complicate this picture. First, almost any rubric edit that adds examples or context can raise agreement (Huynh et al., 2026), so an improvement after targeted edits does not show that the targeting mattered. Second, LLM-written elaborations of rubric text can lower agreement with humans (Sharma et al., 2025). Third, agreement among LLM judges is not evidence of agreement with humans, because judges can converge on a shared axis that annotators do not weight (Mukherjee et al., 2026; Song et al., 2026).

We ask two questions. Does selecting clauses by cross-judge disagreement improve inter-judge agreement more than editing other clauses in the same way? And does any gain in inter-judge agreement carry over to agreement with human annotators? To answer them, we specify a three-stage procedure with deterministic selection rules (Section 3) and evaluate it on three benchmarks with existing human labels (Section 4). The procedure never reads human labels. We compare it against the original rubric, a published judging method, a one-shot LLM rewrite, and a matched control that edits the same number of unselected clauses with the same

editor prompt and example rule. Hypotheses, metrics, and tests were fixed in a version-controlled protocol before any evaluation call.

Our contributions are as follows.

- A fully specified, label-free procedure that maps cross-judge disagreement to a bounded set of clause edits, with selection and example rules that can be rerun without human judgment.
- A preregistered evaluation with a matched random-clause control, five judges from three providers, and three human-annotated benchmarks.
- Evidence that targeting contested clauses increases inter-judge agreement beyond untargeted edits on SummEval, that this gain does not transfer to agreement with expert ratings, and null results on HealthBench and USR.
- Practical observations on running multi-provider judge panels, including a case where structured-output constraints were not enforced on one provider’s batch transport.

## 2 Related Work

**Rubric-based LLM judges.** G-Eval prompts an LLM with task criteria and evaluation steps and reported improved correlation with human ratings on SummEval (Liu et al., 2023). HealthBench grades medical conversations against physician-written criteria with an LLM grader (Arora et al., 2025). Autorubric unifies rubric-based evaluation on non-verifiable tasks (Rao and Callison-Burch, 2026b), and Bagaria et al. (2026) show that rubric wording itself can drive judge scores. We take rubrics from existing benchmarks as given and study how to repair them.

**Effects of rubric edits.** Huynh et al. (2026) quantify how rubric modifications affect human-autorater agreement on essay scoring and instruction following, finding that representative examples, added context, and separate calls per criterion tend to help while higher rubric complexity hurts. ResearchRubrics (Sharma et al., 2025) reports gains from concrete examples and losses from LLM-based augmentation of criteria. These studies vary the type of edit. We hold the edit operation fixed and vary which clauses receive it, selecting them without labels.

**Label-free reliability signals.** CalibratedRubric estimates the measurability of each rubric item from inter-judge agreement alone and filters out items with low measurability, raising agreement with human gold labels on JudgmentBench (Chen et al., 2026). Our procedure uses the same signal for a different purpose: it keeps contested clauses and rewrites them rather than discarding them. PReMISE audits and repairs rubrics using pairwise human preference data (Roy et al., 2026); our procedure uses no human data.

**Consensus and alignment.** Mukherjee et al. (2026) show geometrically that LLM judges can agree with one another as much as humans do while agreeing much less with humans. Song et al. (2026) argue that judge consensus can rest on shared surface heuristics. These results motivate our decision to treat human agreement (H2) as a primary outcome separate from inter-judge agreement (H1).

**Reporting agreement.** Rao and Callison-Burch (2026a) show that several commonly reported agreement statistics coincide on binary data and propose a reporting checklist covering scale, tie and abstention handling, and aggregation level. We follow it in Section 4.5. On the human side, Borghain and Mariathas (2026) decompose physician disagreement in HealthBench, which bounds what any judge can achieve on that benchmark.

## 3 Procedure

The procedure takes a rubric for one evaluation dimension, a development split of items, and a panel of J judges, and returns an edited rubric. All model-generated steps use fixed prompts and are run once.

**Decomposition.** An editor model splits the rubric for each dimension into atomic sub-criteria. For HealthBench, each of the 34 official criteria categories is used unchanged as a sub-criterion. Sub-criteria are identified as C1, C2, and so on within each dimension; for HealthBench these identifiers index the official categories in sorted order. Across the three datasets this yields 75 sub-criteria.

**Probing.** Each judge scores every development item with a probe output: a verdict for each sub-criterion (pass, fail, or not applicable), a one- or two-sentence rationale, and an overall score. Probe outputs are used only for selection and examples; they are not part of the evaluation.

**Stage 1: transfer to code.** The editor maps each sub-criterion to one of four deterministic checkers

(n-gram repetition, truncated ending, formatting artifacts, and lexical overlap with the source or knowledge snippet) or to none. For a mapped sub-criterion, a threshold is chosen on the development split to maximize agreement with the judge majority, counting only items where at least three of five judges gave a pass or fail verdict. A sub-criterion is transferred only if this agreement is at least 0.9. Transferred sub-criteria are removed from the judge rubric, and the checker result is combined with the judge score through a declarative rule. In our evaluation the editor mapped none of the 75 sub-criteria to a checker, so no sub-criterion was transferred and the 0.9 threshold was never tested. The Stage 1 condition therefore differs from the original rubric only in presenting the rubric as a list of sub-criteria, and our results do not speak to the effect of transferring criteria to code.

**Stage 2: localization and clarification.** For each remaining sub-criterion, we compute the rate at which the 10 judge pairs give different verdicts, excluding not-applicable and missing verdicts. Let  $n$  be the number of remaining sub-criteria in a dimension. We set  $k = \min(5, \lfloor n/2 \rfloor)$  and select up to  $k$  sub-criteria with a positive disagreement rate, in decreasing order of rate, breaking ties by identifier. The editor rewrites each selected clause once, given the clause, the development items on which judges split, and the judges’ rationales. If the development split contains items on which all five judges passed and items on which all five failed the sub-criterion, up to two pass-fail pairs are attached as examples; otherwise no example is attached. Rubric text outside the selected clauses is not changed.

The bound  $\lfloor n/2 \rfloor$  guarantees that at least as many unselected clauses remain as were selected, so a matched control can always be formed (Section 4.3). It replaced an initial bound of five after a rehearsal showed that with three to six sub-criteria per dimension the initial rule selected every clause in most dimensions. The change was made before any evaluation call (Appendix A).

**Stage 3: residual reporting.** Disagreement that remains after clarification is not edited further. We report it in three layers: variation across repeated runs of the same judge, differences among judges from the same provider, and differences across providers.

## 4 Experimental Setup

### 4.1 Data

We use three benchmarks with released human labels and annotation guidelines (Table 1). Each guideline serves as the original rubric.

Items are split by cluster (article or conversation) into 20% development and 80% evaluation with seed 20261005, so no source text appears in both splits. A further 20% of evaluation clusters form a repeat sample that is scored three times under the original and full conditions. Human labels for the evaluation split were extracted only after the final manifest was sealed; development labels were never extracted.

### 4.2 Judges

The panel consists of five judges (Table 2). Each judge starts at the lowest reasoning setting its provider supports. A judge that fails a pre-flight check (Section 4.4) may be raised by one reasoning step once. All calls use the providers’ batch APIs with JSON Schema-constrained output. The editor model is Claude Opus 5.5.

Two panel decisions were made during pre-flight checks and before any evaluation call. Gemini 3.1 Flash-Lite, the originally planned Google judge, was replaced by Gemini 3.5 Flash-Lite after failing the probe-format check; we later found that these failures were caused by the batch transport not enforcing the response schema (Appendix C). The Opus score-output cap was raised from 64 to 512 tokens after outputs were truncated by reasoning tokens in a cost pilot.

### 4.3 Conditions

Each evaluation item is scored by all five judges under six rubric conditions.

- **Original:** the benchmark’s annotation guideline.
- **External:** a published judging method run with the same judges. For SummEval we use the G-Eval prompts verbatim, without probability weighting. For USR we use the same template with evaluation steps generated once by the editor. For HealthBench we use the official grader template.
- **Rewrite:** the editor rewrites the original rubric once without disagreement information,

| Dataset                                          | Domain                                                    | Sample                    | Dimensions and scale                   | Human labels                      | Dev units | Eval units |
|--------------------------------------------------|-----------------------------------------------------------|---------------------------|----------------------------------------|-----------------------------------|-----------|------------|
| SummEval (Fab-bri et al., 2021)                  | News summaries (Hermann et al., 2015)                     | 400 items; 25 clusters    | 4 dimensions; 1 to 5                   | Mean of 3 experts                 | 320       | 1,280      |
| HealthBench meta-evaluation (Arora et al., 2025) | Medical conversations                                     | 1,000 items; 890 clusters | Criterion met, binary                  | Physician majority; ties excluded | 201       | 799        |
| USR Topical-Chat (Mehri and Eskenazi, 2020)      | Knowledge-grounded dialogue (Gopalakrishnan et al., 2019) | 360 items; 60 clusters    | 6 qualities; binary, 1 to 3, or 1 to 5 | Mean of 3 annotators              | 432       | 1,728      |

Table 1: Datasets. A unit is one item-dimension pair.

| Judge                 | Provider  | Tier   | Reasoning setting                            | Output cap (score / probe) |
|-----------------------|-----------|--------|----------------------------------------------|----------------------------|
| Claude Opus 5.5       | Anthropic | Large  | effort low (adaptive thinking on)            | 512 / 1,024                |
| Claude Sonnet 5.5     | Anthropic | Medium | effort low, no thinking before output        | 64 / 1,024                 |
| Claude Haiku 4.5      | Anthropic | Small  | thinking off, temperature 0                  | 64 / 1,024                 |
| GPT-6 Luna            | OpenAI    | Small  | reasoning effort low (raised once from none) | 512 / 2,048                |
| Gemini 3.5 Flash-Lite | Google    | Small  | thinking level minimal                       | 512 / 2,048                |

Table 2: Judge panel. The HealthBench official grader output is capped at 1,024 tokens for all judges.

using the same editor instructions as clarification, which forbid adding new constructs. HealthBench criteria are rewritten one at a time.

- **Stage 1:** the decomposed rubric after transfer to code. Since no sub-criterion was transferred, this is the original rubric presented as a list of sub-criteria.
- **Full:** Stage 1 plus clarification of the selected clauses.
- **Random clause:** Stage 1 plus the same clarification procedure applied to  $m$  clauses drawn with seed 20261005 from the clauses not selected in that dimension, regardless of their disagreement rate. The editor prompt and example rule are identical to the full condition.

We also compute majority and mean aggregates of the original-condition scores, which require no additional calls. Each aggregate yields one score per item, so inter-judge  $\alpha$  is not defined for them and we report only their agreement with humans. The majority aggregate requires at least three judges to give the same score and is missing otherwise.

#### 4.4 Pre-flight checks

Before development probing and evaluation, every judge scored 30 synthetic items (five clearly good

and five clearly bad per dataset) in the score-only and probe formats, and the HealthBench subset of these items in the official grader format. The items were written for this purpose and contain no benchmark text. A judge passed if, within one retry, all outputs parsed without truncation, the first-attempt parse failure rate was at most 10%, no pair was inverted, at most one pair was tied, and the mean score of good items exceeded that of bad items. The tie allowance was added after one tied pair in the Claude Haiku 4.5 official-format check; it is a post hoc relaxation, and a retrospective check confirmed that no other past failure would have passed under it (Appendix A).

#### 4.5 Metrics and statistical tests

For each condition, dataset, and dimension we compute Krippendorff’s  $\alpha$  over the five judges (Hayes and Krippendorff, 2007), with an ordinal distance for 3- and 5-point scales and a nominal distance for binary scales. Missing scores are handled by  $\alpha$ ’s pairwise coincidence formulation. Agreement with humans is computed for each judge and for a five-judge ensemble, which uses the first run of an item only when all five judges returned a valid score: Kendall  $\tau$  with the expert mean for SummEval, macro F1 against the physician majority for HealthBench, and Spearman  $\rho$  with the annotator mean for USR. For SummEval and USR the ensemble is the unrounded mean of the five scores.

For HealthBench it is the 3-of-5 majority verdict, and macro F1 averages the F1 of the two verdict classes, not of the 34 criteria; Cohen’s  $\kappa$  is reported as a secondary metric. Metrics that are undefined, for example when a class is absent, are reported as missing. Multi-dimension datasets use the equally weighted mean of per-dimension differences.

The preregistered primary comparisons, three per dataset and nine in total, are full minus original in  $\alpha$  (H1), full minus original in human agreement (H2), and full minus random clause in  $\alpha$  (H3). Each comparison uses items scored under both conditions and a paired cluster bootstrap with 10,000 resamples (Efron, 1979), reporting percentile 95% intervals and two-sided p-values adjusted with Holm’s procedure over the nine tests (Holm, 1979). An  $\alpha$  difference counts as an effect only if it is at least 0.05 and its interval excludes zero. The 0.05 threshold is not applied to human agreement metrics. H4, an exploratory comparison of pairwise weighted  $\kappa$  among the three Claude judges against the three small judges from different providers, is not included in the correction.

#### 4.6 Preregistration and execution

The hypotheses, metrics, tests, and selection rules were committed under a version-control tag before any evaluation call or label extraction. This is a self-registration, not an entry in an external registry. All six rubric conditions, the judge settings, the sample, and the submission order were sealed in a manifest before evaluation (SHA-256 prefix 3f49c9c4). Every change to the original plan is listed with its date and the state of the experiment at that time in Appendix A.

## 5 Results

### 5.1 Execution

All 129,770 planned judge requests (25,954 per judge) were collected over 23.95 hours. After at most one retry, 27 outputs (0.021%) remained unparseable; the highest rate in any judge-condition cell was 0.105%. No request ended in a provider-side failure. Evaluation calls cost US\$236.69 at batch prices. No sub-criterion was transferred to a checker in Stage 1. Every dimension had at least one selected clause, so the full and random-clause conditions differ from Stage 1 in all 11 dimensions; 23 clauses were edited in each of the two conditions. Example pairs were attached to 12 of the 23 selected clauses and to 16 of the 23 control clauses;

none of the five selected HealthBench criteria had an eligible example pair.

### 5.2 Primary comparisons

Table 3 reports the nine preregistered comparisons. Absolute agreement for all six conditions is given in Table 4.

**Secondary conditions.** Presenting the rubric as a list of sub-criteria (Stage 1) slightly lowered inter-judge  $\alpha$  on SummEval (0.577 to 0.556) and HealthBench (0.375 to 0.358). Relative to Stage 1, which is the direct parent of both edited conditions, clarification of contested clauses raised SummEval  $\alpha$  by 0.090 and clarification of control clauses by 0.016. The control condition reached the highest ensemble agreement with humans on SummEval ( $\tau$  0.516 against 0.494 for full) and USR ( $\rho$  0.748 against 0.735). In the preregistered secondary comparison on paired items, the full condition was below the control in human agreement on all three datasets: SummEval  $\tau$  -0.022 (95% CI -0.043 to 0.000, unadjusted  $p = 0.049$ ), HealthBench macro F1 -0.007 (-0.027 to 0.013,  $p = 0.497$ ), and USR  $\rho$  -0.012 (-0.025 to 0.001,  $p = 0.060$ ). All three intervals include zero; the SummEval upper bound is +0.0001. The one-shot rewrite showed no detectable change in human agreement on any dataset and gave the highest  $\alpha$  and macro F1 on HealthBench (0.389 and 0.689). The official HealthBench grader produced the lowest macro F1 (0.619), although it asks for an explanation before the verdict; in a paired exploratory comparison it was 0.062 below the original (95% CI -0.088 to -0.035, unadjusted). G-Eval prompts on SummEval gave the lowest  $\alpha$  (0.541) with human agreement similar to the original.

**SummEval.** The full procedure met the effect criterion against both the original rubric (H1) and the random-clause control (H3). The gain over the control (0.074) was as large as the gain over the original (0.069), so editing an equal number of unselected clauses with the same editor and example rule produced little of the increase. Agreement with expert ratings showed no detectable change (H2). Following the preregistered reporting rule, the primary result for SummEval is that consensus increased without moving the judges closer to human annotators.

**HealthBench.** Inter-judge  $\alpha$  showed no detectable change under either comparison. Macro F1 against the physician majority decreased by 0.023; the unadjusted interval excludes zero, but the Holm-

| Dataset     | Comparison                   | Difference | 95% CI lower | 95% CI upper | Holm p  |
|-------------|------------------------------|------------|--------------|--------------|---------|
| SummEval    | Full - Original (alpha)      | 0.069      | 0.046        | 0.092        | < 0.001 |
| SummEval    | Full - Original (human)      | -0.007     | -0.037       | 0.021        | 1.000   |
| SummEval    | Full - Random clause (alpha) | 0.074      | 0.047        | 0.097        | < 0.001 |
| HealthBench | Full - Original (alpha)      | -0.012     | -0.031       | 0.007        | 1.000   |
| HealthBench | Full - Original (human)      | -0.023     | -0.043       | -0.003       | 0.155   |
| HealthBench | Full - Random clause (alpha) | 0.003      | -0.017       | 0.023        | 1.000   |
| USR         | Full - Original (alpha)      | 0.027      | 0.007        | 0.047        | 0.070   |
| USR         | Full - Original (human)      | 0.005      | -0.011       | 0.023        | 1.000   |
| USR         | Full - Random clause (alpha) | 0.007      | -0.009       | 0.023        | 1.000   |

Table 3: Preregistered primary comparisons. Differences are equally weighted means over dimensions. Intervals are unadjusted percentile intervals from a paired cluster bootstrap; p-values are Holm-adjusted over the nine tests. The effect criterion (difference at least 0.05 and interval excluding zero) applies to  $\alpha$  only.

adjusted p-value is 0.155, so we do not report a significant decrease.

**USR.** Inter-judge  $\alpha$  increased by 0.027 with an unadjusted interval excluding zero, below the pre-registered effect size of 0.05 and not significant after adjustment ( $p = 0.070$ ). The other two comparisons showed no difference.

### 5.3 Exploratory: within- and across-provider agreement (H4)

In all three datasets, the three Claude judges agreed with each other more than the three small judges from different providers did. Clarification narrowed this gap only on SummEval. Because the two groups differ in both provider and model size, and Haiku belongs to both, this comparison does not isolate a provider effect.

Run-to-run disagreement ranged from near zero for Claude Haiku 4.5, run at temperature 0, to 0.246 for GPT-6 Luna on SummEval. On SummEval, clarification lowered run-to-run disagreement for the two small judges from other providers and raised it for Claude Opus 5.5 and Claude Sonnet 5.5. The gain in cross-judge agreement was therefore not accompanied by uniformly more stable individual judges.

## 6 Discussion

**Targeting contested clauses matters for consensus.** On SummEval, editing the clauses on which judges disagreed raised inter-judge agreement, while editing an equal number of other clauses with the same editor prompt and example rule did not (examples were attached to 12 selected and 16 control clauses). This argues against the simplest explanation, that any edit with examples raises agreement, at least for this dataset. Disagreement on a development split was a usable signal

for where judges read a rubric differently.

**Consensus did not become alignment.** The same edits produced no detectable change in agreement with expert ratings. This is consistent with reports that LLM judges can converge on an axis that human annotators do not weight (Mukherjee et al., 2026). A label-free procedure has no information about which of several readings of a clause the annotators held. When the editor resolves an ambiguous clause, it may commit the panel to a reading that is internally consistent but not the annotators’ reading. Our results suggest using disagreement to locate ambiguous clauses, and a small set of human labels or the original annotation instructions to decide how to resolve them.

**Why HealthBench and USR showed no effect.** On HealthBench, only 5 of 34 criteria were edited, and the items assessed under those criteria make up 120 of the 799 evaluation units (15%), which dilutes any effect on the pooled  $\alpha$ . None of the five selected criteria had a unanimous pass-fail example pair on the development split, so their edits changed wording only. Physician agreement on this benchmark is itself limited (Borgohain and Mariathas, 2026). On USR, several dimensions already had high development agreement, and the dataset is small. We cannot separate these explanations with the present design.

**Edits that raise consensus are not the edits that help alignment.** The secondary conditions sharpen this point. Editing unselected clauses barely changed inter-judge agreement but gave the highest human agreement of any edited condition on SummEval and USR, while editing contested clauses did the reverse; in paired comparisons the human-agreement gap favored the control on all three datasets, with intervals that include zero. A one-shot rewrite without any disagreement signal

| Dataset     | Condition           | $\alpha$ | Human metric | Human agreement | Defined dims. | Valid units |
|-------------|---------------------|----------|--------------|-----------------|---------------|-------------|
| SummEval    | Original            | 0.577    | $\tau$       | 0.501           | 4             | 1280        |
| SummEval    | Stage 1             | 0.556    | $\tau$       | 0.480           | 4             | 1280        |
| SummEval    | Full                | 0.646    | $\tau$       | 0.494           | 4             | 1280        |
| SummEval    | External            | 0.541    | $\tau$       | 0.506           | 4             | 1280        |
| SummEval    | Rewrite             | 0.570    | $\tau$       | 0.501           | 4             | 1280        |
| SummEval    | Random clause       | 0.572    | $\tau$       | 0.516           | 4             | 1280        |
| SummEval    | Majority (Original) | NA       | $\tau$       | 0.550*          | 4             | 974         |
| SummEval    | Mean (Original)     | NA       | $\tau$       | 0.501           | 4             | 1280        |
| HealthBench | Original            | 0.375    | Macro F1     | 0.681           | 1             | 654         |
| HealthBench | Stage 1             | 0.358    | Macro F1     | 0.681           | 1             | 653         |
| HealthBench | Full                | 0.362    | Macro F1     | 0.659           | 1             | 651         |
| HealthBench | External            | 0.369    | Macro F1     | 0.619           | 1             | 655         |
| HealthBench | Rewrite             | 0.389    | Macro F1     | 0.689           | 1             | 653         |
| HealthBench | Random clause       | 0.359    | Macro F1     | 0.665           | 1             | 654         |
| HealthBench | Majority (Original) | NA       | Macro F1     | 0.681           | 1             | 654         |
| HealthBench | Mean (Original)     | NA       | Macro F1     | 0.681           | 1             | 654         |
| USR         | Original            | 0.659    | $\rho$       | 0.730           | 6             | 1728        |
| USR         | Stage 1             | 0.663    | $\rho$       | 0.730           | 6             | 1728        |
| USR         | Full                | 0.685    | $\rho$       | 0.735           | 6             | 1728        |
| USR         | External            | 0.670    | $\rho$       | 0.728           | 6             | 1728        |
| USR         | Rewrite             | 0.647    | $\rho$       | 0.732           | 6             | 1728        |
| USR         | Random clause       | 0.678    | $\rho$       | 0.748           | 6             | 1728        |
| USR         | Majority (Original) | NA       | $\rho$       | 0.679*          | 6             | 1646        |
| USR         | Mean (Original)     | NA       | $\rho$       | 0.730           | 6             | 1728        |

Table 4: Inter-judge  $\alpha$  and ensemble agreement with humans for all conditions on the evaluation split. Values are equally weighted means over dimensions, each computed on all valid units of that condition, so row differences are descriptive and can differ slightly from the paired differences in Table 3.  $\alpha$  is not defined for single-score aggregates. The mean aggregate equals the original ensemble by definition, and for binary HealthBench so does the majority aggregate. \*Computed only on units where at least three judges gave the same score (974 of 1,280 for SummEval, 1,646 of 1,728 for USR), so not comparable with the other rows.

| Dataset     | Original, Claude | Original, small | Full, Claude | Full, small | Change in gap  |
|-------------|------------------|-----------------|--------------|-------------|----------------|
| SummEval    | 0.707            | 0.578           | 0.757        | 0.654       | 0.026 narrower |
| HealthBench | 0.410            | 0.378           | 0.418        | 0.339       | 0.047 wider    |
| USR         | 0.704            | 0.635           | 0.734        | 0.648       | 0.017 wider    |

Table 5: Mean pairwise weighted  $\kappa$  among the three Claude judges and among the three small judges from different providers (Claude Haiku 4.5, GPT-6 Luna, Gemini 3.5 Flash-Lite). Exploratory; not adjusted.

showed no detectable change in human agreement on any dataset, unlike the degradation reported for LLM-elaborated criteria (Sharma et al., 2025); our rewrite was a single pass over the existing rubric under an instruction not to add new constructs, rather than an expansion with synthetic detail. In paired exploratory comparisons, the rewrite changed human agreement by less than 0.01 on every dataset (Appendix E). Apart from the secondary comparison, none of these contrasts were planned tests, and none of the comparisons in this paragraph is significant; taken together, they argue against using inter-judge agreement alone as the objective when repairing rubrics. We did not run equivalence tests, so the absence of a detectable change should not be read as evidence of no change.

**A negated criterion as a case study.** In pre-

flight checks, two judges handled HealthBench criterion C31, which passes a response only if it does not ask the user follow-up questions, with the same incorrect verdict: they marked a response consisting only of follow-up questions as meeting the criterion. One judge’s explanation correctly identified the failure before emitting criteria\_met = true (Appendix B). The judge’s reasoning was right and its verdict inverted. This is the kind of clause-level ambiguity, here in the mapping from a negated criterion to an output value, that the procedure is meant to find. GPT-6 Luna returned only a score, so we cannot tell whether its error had the same cause.

**Running multi-provider panels.** Structured-output enforcement differed by provider and by transport. For one provider, a response schema

| Dataset     | Condition | Opus  | Sonnet | Haiku | Luna  | Flash-Lite | Claude-3 $\kappa$ | Small-3 $\kappa$ |
|-------------|-----------|-------|--------|-------|-------|------------|-------------------|------------------|
| SummEval    | Original  | 0.086 | 0.062  | 0.022 | 0.246 | 0.102      | 0.707             | 0.578            |
| SummEval    | Full      | 0.113 | 0.092  | 0.013 | 0.214 | 0.051      | 0.757             | 0.654            |
| HealthBench | Original  | 0.033 | 0.021  | 0.000 | 0.086 | 0.025      | 0.410             | 0.378            |
| HealthBench | Full      | 0.029 | 0.016  | 0.000 | 0.099 | 0.025      | 0.418             | 0.339            |
| USR         | Original  | 0.039 | 0.033  | 0.002 | 0.135 | 0.074      | 0.704             | 0.635            |
| USR         | Full      | 0.037 | 0.024  | 0.004 | 0.113 | 0.074      | 0.734             | 0.648            |

Table 6: Residual diagnostics on the repeat sample. Judge columns give the share of the three run pairs per item in which a judge’s scores differed (483 to 1,080 pairs per cell). The last two columns repeat Table 5. The diagnostics are descriptive and do not sum to a total variance.

passed through the batch API was not enforced (0 of 20 outputs conformed in a controlled test), whereas a different documented field was enforced (20 of 20) (Appendix C). Simple score-only checks passed under both, and the failure appeared only with nested probe outputs. Panels that span providers need format checks for every output format they use.

## 7 Limitations

**Panel composition.** Three of the five judges are Claude models, and the other two providers are represented only by small models. Reasoning settings and output caps differ across judges (Table 2), so provider, size, and reasoning effects are confounded. The editor model belongs to the same family as three judges, which may favor edits that those judges read the same way.

**Changes before evaluation.** Several decisions were changed before any evaluation call or label access: the clause-selection bound before the self-registered protocol was committed, and the following after it: the Google judge model, one reasoning-step increase for GPT-6 Luna, the Opus output cap, the HealthBench output instruction, the removal of prompt caching, and a relaxation of the pre-flight ordering criterion after a failure. Each is listed in Appendix A. They were not part of the original plan and should be read as such.

**Procedure scope.** Each clause is edited once with a single editor prompt; we did not study iteration or alternative editors. The procedure’s original formulation included a step for system defects in a deployed pipeline, which public benchmarks cannot exercise; we exclude it. The code-transfer stage was not exercised: the editor mapped none of the 75 sub-criteria to the four available checkers, which target surface defects such as repetition and truncation. Testing that stage would require rubrics that name such defects explicitly or a larger checker

library.

**Data.** The evaluation uses existing human labels, which are themselves noisy, and three English datasets. SummEval is widely used and may appear in judges’ training data; this affects all conditions equally but may raise baseline agreement. HealthBench edits affect a small share of items. The external baseline re-runs published prompts with current judges, and our G-Eval variant omits probability weighting.

**Statistics.** Percentile bootstrap intervals are unadjusted, while p-values are Holm-adjusted; an interval that excludes zero is therefore not by itself a significant result. H4 is exploratory and confounds provider with model size.

## 8 Ethical Considerations

We collected no new human annotations. All datasets are publicly released; for sources without an explicit data license we use them for research evaluation only and do not redistribute source text, requests, or responses. Artifacts to be released contain item identifiers, hashes, judge scores, and aggregates. HealthBench contains medical conversations; our study evaluates judging procedures and makes no clinical claims. We used Claude (Anthropic) as an AI assistant for software development, experiment orchestration, and drafting and editing of this paper. The authors designed the study, made all research decisions, reviewed the generated code and text, and verified the reported numbers against the analysis logs and the references against their sources; the authors take full responsibility for the content.

## References

- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [HealthBench: Evaluating Large Language Models Towards Improved Human Health](#). *Preprint*, arXiv:2505.08775.
- Anshul Bagaria, Sowmya S Sundaram, Gokul S Krishnan, and Balaraman Ravindran. 2026. [Judging LLM-as-a-Judge: Concerning Rubric Artifacts in LLM-based Automated Text Generation Evaluation](#). *Preprint*, arXiv:2609.02942.
- Satya Borgohain and Roy Mariathas. 2026. [Decomposing Physician Disagreement in HealthBench](#). *Preprint*, arXiv:2602.22758.
- Mengting Chen, Yanshu Sun, Wanting Liang, Beidi Luan, Rui Sun, Dezhi Chen, Jing Li, and Zuo Bai. 2026. [CalibratedRubric: Task-Adaptive Rubric Banks for Open-Ended LLM Evaluation](#). *Preprint*, arXiv:2607.29252.
- B. Efron. 1979. [Bootstrap Methods: Another Look at the Jackknife](#). *The Annals of Statistics*, 7(1).
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating Summarization Evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qintiang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. [Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations](#). In *Interspeech 2019*, pages 1891–1895.
- Andrew F. Hayes and Klaus Krippendorff. 2007. [Answering the Call for a Standard Reliability Measure for Coding Data](#). *Communication Methods and Measures*, 1(1):77–89.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching Machines to Read and Comprehend](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Sture Holm. 1979. [A Simple Sequentially Rejective Multiple Test Procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.
- Jessica Huynh, Alfredo Gomez, Athiya Deviyani, Renee Shelby, Jeffrey P. Bigham, and Fernando Diaz. 2026. [Quantifying the Statistical Effect of Rubric Modifications on Human-Autorater Agreement](#). *Preprint*, arXiv:2605.06283.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Shikib Mehri and Maxine Eskenazi. 2020. [USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.
- Sourabrata Mukherjee, Hamna Hamna, Kalika Bali, and Sunayana Sitaram. 2026. [The Geometry of LLM-as-Judge: Why Inter-LLM Consensus Is Not Human Alignment](#). *Preprint*, arXiv:2606.03043.
- Delip Rao and Chris Callison-Burch. 2026a. [Agreement Metrics for LLM-as-Judge Evaluation: What to Report and Why](#). *Preprint*, arXiv:2606.00093.
- Delip Rao and Chris Callison-Burch. 2026b. [Aurubric: A Unifying Framework for Rubric-Based LLM Evaluation on Non-Verifiable Tasks](#). In *Third Conference on Language Modeling*.
- Swastik Roy, Rajkumar Pujari, Tharindu Kumarage, Charith Peris, Rahul Gupta, Anna Rumshisky, Pradeep Natarajan, and Venkatesh Saligrama. 2026. [PREMISE: Policy Rubrics as Measurement Specifications for LLM Judges](#). *Preprint*, arXiv:2605.30803.
- 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 M. Hendryx, Brad Kenstler, and Bing Liu. 2025. [ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents](#). *Preprint*, arXiv:2511.07685.
- Mingyang Song, Mao Zheng, and Chenning Xu. 2026. [Beyond the Illusion of Consensus: From Surface Heuristics to Knowledge-Grounded Evaluation in LLM-as-a-Judge](#). *Preprint*, arXiv:2603.11027.

## A Changes to the Plan

All changes below were made before any evaluation call and before evaluation labels were extracted, unless stated otherwise.

| Date                              | Change                                                                                                                               | Reason                                                                                                                   |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| 2026-10-05                        | Selection bound changed from top 5 to $k = \min(5, \text{floor}(n/2))$                                                               | Rehearsal with three Claude judges found too few unselected clauses to form an equal-size control in 10 of 11 dimensions |
| 2026-10-06                        | HealthBench output instruction made uniform across all 34 criteria; Claude HealthBench probes rerun                                  | Some official criteria instruct a <code>criteria_met</code> field that conflicts with the score-only output schema       |
| 2026-10-06                        | GPT-6 Luna reasoning raised from none to low (one step)                                                                              | Failed the ordering check at the lowest setting                                                                          |
| 2026-10-06                        | Earlier rule “lowest setting with reasoning enabled” withdrawn                                                                       | Other judges also answered without reasoning and passed; rule replaced by one-step escalation on failure                 |
| 2026-10-06                        | Google judge changed from Gemini 3.1 Flash-Lite to Gemini 3.5 Flash-Lite; Google batch transport changed to an enforced schema field | Probe-format outputs failed on both models; a controlled test showed the original batch field was not enforced           |
| 2026-10-07                        | Pre-flight ordering criterion relaxed to allow one tied pair                                                                         | One tied pair for Claude Haiku 4.5 on C31; retrospective check of 67 past cohorts changed no other outcome               |
| 2026-10-07                        | Opus score-output cap raised from 64 to 512 tokens                                                                                   | Reasoning tokens truncated 9 of 90 cost-pilot outputs                                                                    |
| 2026-10-07                        | Budget limits raised; prompt caching added and then removed                                                                          | Conservative cost estimate exceeded limits; cache hits are not guaranteed in batch processing                            |
| 2026-10-08<br>(during evaluation) | Evaluation driver restarted after unexplained termination; one batch of 100 requests resubmitted once                                | Provider records showed the batch had not been received; no settings changed                                             |

Table 7: Changes to the plan (Appendix A).

## B Case Study, Criterion C31

### Synthetic context

User: In this fictional clinic directory, the desk for collecting my printed visit summary is Desk  
↪ Azure. Where does the directory say I should collect it? The directory gives all the information  
↪ needed.

### High target

Collect the printed visit summary at Desk Azure.

### Low target

What desk does your directory name, and what document are you collecting?

### Criterion (verbatim)

Step 1: Does the response seek more context from the user by asking follow up questions that have the  
↪ intent of gaining more information for the model to produce a better next response? If yes, fail.  
Step 2: Does the response seek more context from the user by stating what information would be helpful  
↪ for the user to provide to the model so the model can produce a better response? If yes, fail.  
If both steps pass, mark `criteria_met` = true.  
If any of the steps fail, mark `criteria_met` = false.

### Haiku high response (verbatim)

```
{"criteria_met": true, "explanation": "The response does not seek more context from the user by asking  
↪ follow-up questions (Step 1 passes), nor does it state what information would be helpful for the  
↪ user to provide (Step 2 passes). The assistant directly answers the user's question based on the  
↪ information provided in the fictional clinic directory. Both steps pass, so the criteria is met."}
```

### Haiku low response (verbatim)

```
{"criteria_met": true, "explanation": "The assistant's response asks follow-up questions ('What desk  
→ does your directory name, and what document are you collecting?') with the intent of gaining more  
→ information to produce a better next response. According to Step 1 of the rubric, if the response  
→ seeks more context from the user by asking follow up questions that have the intent of gaining  
→ more information, it should fail. Since the response does ask follow-up questions with this  
→ intent, the criteria is met (meaning it fails the rubric requirement)."}"
```

### Luna high response (verbatim)

```
{"score":1}
```

### Luna low response (verbatim)

```
{"score":1}
```

The synthetic user states the desk name and says the directory gives all needed information. The bad response asks which desk the directory names and which document the user is collecting. C31 fails a response that seeks more context through follow-up questions. Claude Haiku 4.5’s explanation states that the response asks follow-up questions and therefore fails Step 1, then outputs `criteria_met = true`. GPT-6 Luna at its lowest reasoning setting gave the same response a passing score, with no explanation available.

## C Structured-Output Enforcement Test

| Provider and field                           | Synchronous     | Batch           | Used in study          |
|----------------------------------------------|-----------------|-----------------|------------------------|
| Google, responseJsonSchema                   | 20/20 conform   | 0/20 conform    | No (failures recorded) |
| Google, _responseJsonSchema                  | 20/20 conform   | 20/20 conform   | No                     |
| Google, responseSchema                       | Rejected schema | Rejected schema | No                     |
| Google, responseFormat with schema           | 20/20 conform   | 20/20 conform   | Yes                    |
| Anthropic (Haiku), output_config JSON schema | 20/20 conform   | 20/20 conform   | Yes                    |
| OpenAI (Luna), strict JSON schema            | 20/20 conform   | 20/20 conform   | Yes                    |

Table 8: Conformance when the prompt asks for an output that contradicts the schema. Ten requests each for a flat and a nested schema, per mode.

## D Development-Split Agreement

| Dimension               | Units | $\alpha$ (five judges, original probes) |
|-------------------------|-------|-----------------------------------------|
| HealthBench Fulfillment | 201   | 0.404                                   |
| SummEval Coherence      | 80    | 0.612                                   |
| SummEval Consistency    | 80    | 0.755                                   |
| SummEval Fluency        | 80    | 0.341                                   |
| SummEval Relevance      | 80    | 0.472                                   |
| USR Engaging            | 72    | 0.671                                   |
| USR Maintains Context   | 72    | 0.498                                   |
| USR Natural             | 72    | 0.768                                   |
| USR Overall             | 72    | 0.714                                   |
| USR Understandable      | 72    | 0.466                                   |
| USR Uses Knowledge      | 72    | 0.785                                   |

Table 9: Development-split agreement.

| Dataset  | Dimension         | $n$ | $k$ | $m$ | Selected IDs | Control IDs |
|----------|-------------------|-----|-----|-----|--------------|-------------|
| SummEval | Coherence         | 4   | 2   | 2   | C4,C2        | C1,C3       |
| SummEval | Consistency       | 3   | 1   | 1   | C3           | C1          |
| SummEval | Fluency           | 3   | 1   | 1   | C1           | C2          |
| SummEval | Relevance         | 4   | 2   | 2   | C4,C1        | C2,C3       |
| USR      | Engaging          | 5   | 2   | 2   | C3,C5        | C1,C4       |
| USR      | Maintains Context | 5   | 2   | 2   | C3,C4        | C2,C5       |
| USR      | Natural           | 4   | 2   | 2   | C2,C3        | C1,C4       |
| USR      | Overall           | 6   | 3   | 3   | C4,C3,C5     | C1,C2,C6    |
| USR      | Understandable    | 4   | 2   | 2   | C3,C2        | C1,C4       |
| USR      | Uses Knowledge    | 3   | 1   | 1   | C2           | C3          |

Table 10: Clause selection. Selected clauses are listed in decreasing order of development-split disagreement. Sub-criterion IDs refer to the frozen decomposition, which is released with the code.

| Dataset     | Dimension   | $n$ | $k$ | $m$ | Selected IDs        | Control IDs             |
|-------------|-------------|-----|-----|-----|---------------------|-------------------------|
| HealthBench | Fulfillment | 34  | 5   | 5   | C6, C1, C2, C20, C5 | C14, C18, C19, C23, C28 |

Table 11: HealthBench selection (continuation of Table 10).

## E Exploratory Paired Contrasts

| Dataset     | Contrast            | Metric   | Difference | CI lower | CI upper | Uncorrected $p$ | Paired units |
|-------------|---------------------|----------|------------|----------|----------|-----------------|--------------|
| SummEval    | Stage 1 - Original  | $\alpha$ | -0.021     | -0.038   | -0.005   | 0.014           | 1280         |
| SummEval    | Stage 1 - Original  | $\tau$   | -0.021     | -0.047   | 0.003    | 0.103           | 1280         |
| SummEval    | Rewrite - Original  | $\alpha$ | -0.007     | -0.025   | 0.011    | 0.459           | 1280         |
| SummEval    | Rewrite - Original  | $\tau$   | 0.000      | -0.016   | 0.015    | 0.957           | 1280         |
| SummEval    | External - Original | $\alpha$ | -0.035     | -0.063   | -0.011   | 0.008           | 1280         |
| SummEval    | External - Original | $\tau$   | 0.005      | -0.020   | 0.028    | 0.710           | 1280         |
| HealthBench | Stage 1 - Original  | $\alpha$ | -0.017     | -0.035   | 0.002    | 0.079           | 794          |
| HealthBench | Stage 1 - Original  | Macro F1 | 0.000      | -0.019   | 0.018    | 0.961           | 652          |
| HealthBench | Rewrite - Original  | $\alpha$ | 0.015      | -0.009   | 0.039    | 0.242           | 792          |
| HealthBench | Rewrite - Original  | Macro F1 | 0.006      | -0.017   | 0.030    | 0.605           | 653          |
| HealthBench | External - Original | $\alpha$ | -0.005     | -0.037   | 0.026    | 0.736           | 797          |
| HealthBench | External - Original | Macro F1 | -0.062     | -0.088   | -0.035   | < 0.001         | 654          |
| USR         | Stage 1 - Original  | $\alpha$ | 0.004      | -0.016   | 0.024    | 0.676           | 1728         |
| USR         | Stage 1 - Original  | $\rho$   | 0.000      | -0.019   | 0.019    | 0.985           | 1728         |
| USR         | Rewrite - Original  | $\alpha$ | -0.012     | -0.025   | 0.002    | 0.080           | 1728         |
| USR         | Rewrite - Original  | $\rho$   | 0.002      | -0.010   | 0.015    | 0.783           | 1728         |
| USR         | External - Original | $\alpha$ | 0.012      | -0.007   | 0.030    | 0.213           | 1728         |
| USR         | External - Original | $\rho$   | -0.002     | -0.018   | 0.014    | 0.794           | 1728         |

Table 12: Exploratory paired contrasts, computed on the items scored by all five judges under both conditions, with the same cluster bootstrap as the primary comparisons. p-values are not adjusted, and none of these contrasts was preregistered.