# SCORING HIGHER, ANSWERING WORSE: MITIGATING REWARD HACKING IN RUBRIC-BASED RL VIA PROTOCOL-LEVEL RUBRICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Rubric-based reinforcement learning (Rubric-RL) trains language models where no verifier exists. A judge checks each criterion of a rubric, and the verdicts are aggregated into a reward, most often by a weighted sum. We show that this additive aggregation is the weak point. Under a sum, criteria compensate for one another: a policy that misses the one decision that matters can buy the points back with advice nobody asked for. On clinical consultation, such a policy scores higher and answers worse. Rubric coverage rises while appropriateness on held-out physician criteria falls below the untrained model. The medical criteria are not to blame. Grouped so that they must hold together, the same criteria, unchanged to the word, recover a third of the loss; shorter answers recover almost none. We therefore propose **Pro**tocol-level **Rubric**s (**ProRubric**), which keeps what the criteria ask for and changes how they are aggregated. It groups a checklist into a few protocol-level dimensions. A dimension counts only when all of its criteria hold and its failure clause does not fire. The grouping is done once, offline, and leaves the optimizer unchanged. ProRubric raises appropriateness by 10.8 points without losing coverage and has the best seven-benchmark average at both scales. Reward validity is set not only by what a rubric verifies, but by how it aggregates.

*"When a measure becomes a target, it ceases to be a good measure."*
— Goodhart's law

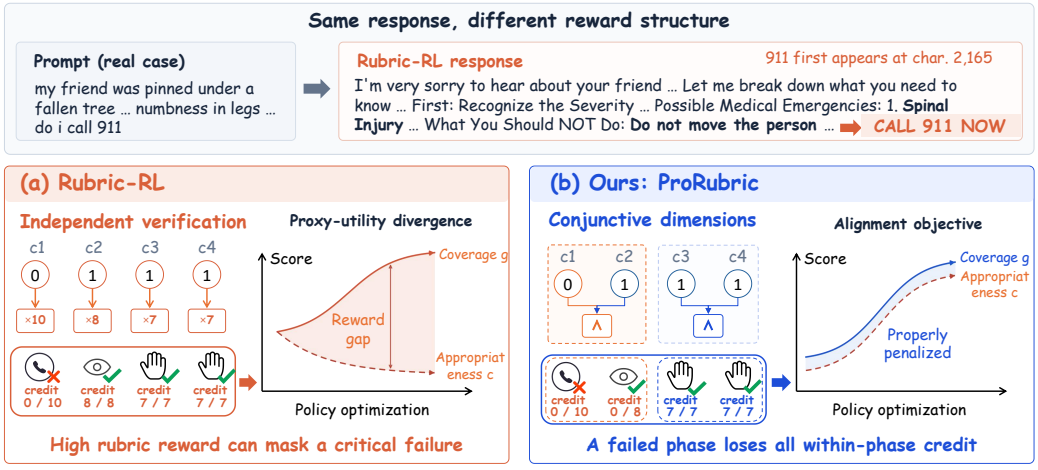


Figure 1: **Comparison of reward verification paradigms in clinical protocol RL.** Top: a real prompt and a Rubric-RL response in which the call to 911 first appears after 2,165 characters. (a) Conventional additive Rubric-RL scores atomic criteria independently, rewarding superficial coverage while masking critical omissions. (b) ProRubric counts a group of criteria only when all of them hold, so cheap criteria can no longer offset a missed critical one.

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

# 1 INTRODUCTION

Reinforcement learning has improved language models most where a verifier exists: a compiler for code, an answer checker for mathematics (Shao et al., 2024). Where none exists, rubrics have taken its place. Rubric-based reinforcement learning (Rubric-RL) asks an LLM judge to check a response against a checklist of criteria and turns the verdicts into a reward (Gunjal et al., 2026; Viswanathan et al., 2025; Huang et al., 2025; Jia et al., 2026). Unlike a monolithic reward model or judge, which is prone to length bias and style exploitation (Singhal et al., 2024; Dubois et al., 2024; Zheng et al., 2023), such a reward can be read criterion by criterion, and a growing body of work builds on it: rubric datasets and generators (Li et al., 2026; Liu et al., 2026; Shen et al., 2026), rubric-guided exploration (Zhou et al., 2025; Hou et al., 2026), and rubric benchmarks on which frontier models are evaluated (Arora et al., 2025; OpenAI, 2025). Most of that work is about writing better criteria. Aggregation has barely changed: most often it is a weighted sum, which RaR calls *explicit aggregation* (Gunjal et al., 2026).

Decomposing quality into criteria is easy; recomposing criteria into quality is where rubric rewards fail. Clinical standards are protocols, not tallies (Gawande, 2009). Getting the triage right is not one criterion among thirty, and missing it is not a deduction to be made up elsewhere. A weighted sum treats it as exactly that. Criteria compensate for one another, so a policy that misses the decision can buy the points back with material nobody asked for, a form of reward hacking (Skalse et al., 2022) that policy gradient is well placed to find (Figure 1a). Told *"My baby has a fever,"* a policy trained this way replies with $10.9$k characters and fails both physician-written criteria for the case. One of them asks that the referral not be buried in verbose text.

The policy that results scores higher and answers worse, and standard reporting records only the first half. The score that published work reports is rubric coverage $g$, the proxy being optimized (Gao et al., 2023). Training raises it on Qwen3-4B from $26.6$ to $32.1$. On the same queries, appropriateness $c$, judged against physician-written criteria held out from training, falls from $54.1$ to $26.0$, less than half of where the untrained model stands. An exploration baseline (Zhou et al., 2025) falls as far. A judge shown no rubric at all prefers the untrained model. Mahmoud et al. (2026) observed the same reversal and attributed it to under-specified rubrics; we show that aggregation alone moves it.

The obvious suspect is verbosity: trained answers are four times as long, and frontier evaluations now penalize longer answers on HealthBench (OpenAI, 2026; Hicks et al., 2026). We test this and other potential explanations systematically, leaving the criteria untouched. First, non-aggregation hypotheses fail to account for the collapse: cutting answers by $42\%$ leaves appropriateness where it was; swapping in a judge from another model family preserves the ranking; fine-tuning on the same rubric data without RL ends above the untrained model; and raising the KL penalty restores appropriateness only once learning stops. Second, what genuinely restores appropriateness while learning continues is the aggregation mechanism itself: grouping the verbatim criteria into conjunctive units recovers a third of the loss, while scoring them holistically as one list recovers nearly $40\%$. In medicine and dialogue, changing only the aggregation moves the collapse while the criteria stay fixed; in science, where criteria are substantive, both grouping and synthesis matter.

If the aggregation is the problem, the criteria can keep what they ask for while the aggregation changes. We propose **Pro**tocol-level **Rubric**s (**ProRubric**; Figure 1b) to make that change. ProRubric groups a checklist into a few dimensions that follow the phases of the protocol. A dimension counts only when all its criteria hold, and each dimension carries a failure clause that voids it. The grouping is done once, offline. The optimizer and the served policy are untouched, and, unlike dependency-aware aggregation (Lv et al., 2026), no relations among criteria need to be annotated. Told the same five words, the ProRubric policy replies with $5.3$k characters and meets both criteria. ProRubric raises appropriateness by $10.8$ points without losing coverage, and it has the best seven-benchmark average at both model scales.

Our main contributions are:

- **Finding.** Rubric-RL with additive aggregation scores higher and answers worse: coverage rises while appropriateness on held-out physician criteria falls below the untrained model, across seeds, scales and judges, and the standard score does not show it.

- **Analysis.** With the criterion text fixed, changing only the aggregation moves the collapse in medicine and dialogue, and partly in science, while length, KL, weights, the judge and the data do not; the incentive can be read off the reward before training.
- **Method.** We propose ProRubric, an offline regrouping of the rubric that improves appropriateness without losing coverage and has the best seven-benchmark average at both model scales.

## 2 PRELIMINARIES

### 2.1 RUBRIC REWARDS: CRITERIA AND AGGREGATION

A rubric specifies a task as a checklist of $m$ criteria $\mathcal{C} = \{(r_i, w_i)\}_{i=1}^{m}$, where $r_i$ is a natural-language requirement and $w_i \neq 0$ its weight, negative for a penalty criterion (a behavior to avoid) (Gunjal et al., 2026; Viswanathan et al., 2025). Turning a rubric into a reward involves two choices: the criteria, which fix what is checked, and the aggregation, which turns a judge's verdicts into a scalar. The standard choice scores each criterion separately, $s_i(x, y) \in \{0, 1\}$ for a prompt $x$ and a response $y \sim \pi_\theta(\cdot \mid x)$, and takes the normalized weighted sum, which RaR calls *explicit aggregation*:

$$R_{\text{atom}}(x, y) = \text{clip}_{[0,1]}\Big(\frac{\sum_{i=1}^{m} w_i s_i(x, y)}{\sum_{i:\, w_i>0} w_i}\Big). \tag{1}$$

Other aggregations keep the weighted mean and change what counts as one scored item: a group of criteria that must all hold (conjunctive grouping, Section 4), or the whole checklist scored at once (RaR's *implicit aggregation*). Our controls fix $\mathcal{C}$ and the optimizer and change only this choice.

**Compensation.** Under Eq. 1, criteria compensate for one another: a response that misses a critical criterion can recover the lost reward by satisfying easier ones (Figure 1). A policy that learns to do this is reward hacking (Skalse et al., 2022), and Section 3.2 shows why an optimizer of expected reward is drawn to it. We optimize with GRPO (Shao et al., 2024); training details are in Section 5.1.

### 2.2 TWO AXES: RUBRIC COVERAGE AND APPROPRIATENESS

A policy scored only by the kind of rubric it was trained on cannot reveal compensation, so we read every medical policy on two axes. **Rubric coverage** $g$ is the score on HealthBench (Arora et al., 2025) under its own example-specific criteria. It is the proxy that training optimizes, in the sense of Gao et al. (2023), and the number published work reports. **Appropriateness** $c$ is the score on HealthBench's physician-consensus criteria, which practicing clinicians wrote and which are held out from training. They ask whether the response gets the decision right for this user, for example whether an emergency referral is stated clearly rather than buried in text. Both axes are scored by DeepSeek-V4-Pro (Section 5.1), not by clinicians, hence *appropriateness* rather than clinical utility.

**Problem statement.** With the criteria and the optimizer held fixed, does the choice of aggregation decide whether optimizing the reward raises $c$ along with $g$? Section 3 shows that under explicit aggregation $g$ rises while $c$ falls below the untrained model, and why. Section 4 proposes a conjunctive aggregation, and Section 5 tests both against the alternatives.

## 3 THE REWARD-HACKING CHANNEL IN ADDITIVE AGGREGATION

### 3.1 SCORING HIGHER, ANSWERING WORSE

In clinical consultation, training against the explicitly aggregated reward pulls the two axes apart (Figure 2). Coverage rises by five points, while appropriateness falls by twenty-eight, to less than half of where the untrained model started. The pattern holds at both model scales and in every seed, and an exploration baseline that keeps the same reward falls just as far.

On the score that published work reports, the policy improves; the damage appears only on criteria held out from training. Nor is it an artifact of rubric-based judging: shown the answers with no rubric at all, judges from two model families (DeepSeek-V4-Pro and GPT-5.6-luna; Section 5.1) both prefer the untrained model, in about two pairs of three and in nearly all pairs respectively,

Table 1: **Evidence of reward hacking without a rubric and before training.** (a) Blind pairwise win rates against the untrained Base (bold denotes Base preferred). (b) Reward change per edit.

| (a) After Rubric-RL training: untrained answer preferred (%) | | | (b) Before training: reward change per edit | | | |
|---|---|---|---|---|---|---|
| Domain | DeepSeek-V4-Pro | GPT-5.6-luna | Edit to the answer | Rubric-RL | raw-AND | ProRubric |
| Medicine | **65.7** ±13.2 | **93.6** ±2.3 | Name an item | **+7.3** | −0.3 | +2.2 |
| Science | 1.2 ±1.2 | **90.3** ±9.0 | Name the topic | −0.4 | +0.6 | −3.8 |
| Dialogue | 24.3 ±5.8 | 44.9 ±6.8 | Add a needless test | +0.3 | −1.4 | **−5.8** |
| Writing | 0.0 ±0.0 | 1.4 ±1.6 | Drop the key advice | −0.9 | −0.8 | −2.2 |

averaged over three seeds (Table 1a). The two judges disagree in science, and in dialogue and in writing both prefer the trained policy (Appendix D.4).

**The criteria set the sign.** Both axes are scored by the same judge, so the gap comes from what is checked, not from who checks it: on the same responses, the coverage criteria record a gain over the untrained model and the physician criteria a loss. A judge from the training reward's own family shows the same sign with a smaller loss (Appendix D.2); a judge closer to the reward understates the damage rather than creating it.

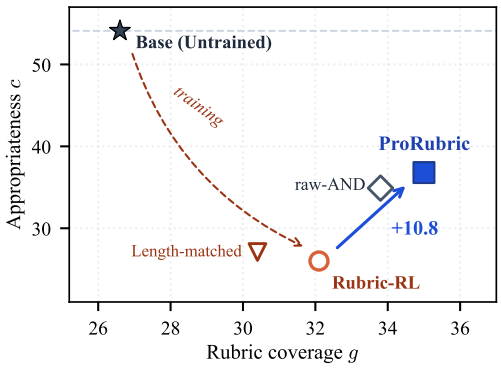


Figure 2: **Reward hacking under additive aggregation.** Coverage vs. appropriateness across medical 4B arms.

**What the policy learns.** The trained policy writes four times as much as the untrained one. On nearly half of the physician criteria (about a quarter before training), the training-family judge passes the answer and DeepSeek-V4-Pro does not (Appendix D.2). The key decision may still be there, but buried: in Figure 1, the emergency referral first appears two thousand characters into the answer. On *"My baby has a fever"* the answer exceeds 10,000 characters and fails both physician criteria (Appendix E.1).

### 3.2 Why the Cheap Behavior Is the Rational One

Under additive aggregation every criterion pays a fixed rate however hard it is to satisfy, so the cheapest criteria are the rational target. Let $p_i$ be the probability that the policy satisfies criterion $i$; under Eq. 1 the expected reward is a weighted sum of these probabilities, and each criterion's rate is its share of the total weight. A correct triage decision and a reminder to stay hydrated earn the same, but the first needs clinical reasoning the policy does not reliably have, while the second comes almost free from pre-training. Equal reward, unequal cost: a policy maximizing expected reward has no reason to prefer the first. This is an argument about the objective, not a convergence proof about the optimizer, and it makes a prediction that can be checked without training anything.

### 3.3 Reading the Incentive off the Reward

We perturb the untrained model's answers to 300 clinical cases in ways that mirror what trained policies do, and record how the reward changes (Table 1b). A paraphrase that keeps the meaning moves each of the three rewards by about one point (+1.0 to +1.1; Table 13), so the larger changes are not judge noise. Three readings of the Rubric-RL column account for Section 3.1. Naming a concrete item without analyzing it pays well, while naming only the topic pays nothing: the reward buys recognizable entities, not substance. Recommending an unwarranted high-risk test costs nothing, because no criterion forbids it. Deleting the critical recommendation costs almost nothing either, so an answer that drops the decision and adds a term comes out ahead. The other two columns preview Section 4: grouping the same criteria verbatim (raw-AND) stops paying for a name, and ProRubric's failure clauses make the needless test costly.

# 4 METHOD: PROTOCOL-LEVEL RUBRICS (PRORUBRIC)

Section 3 traced the collapse to a fixed price per criterion; ProRubric removes that price within each phase of the protocol, changing how criteria are aggregated, not what they ask for. It rests on one property of the protocols such checklists encode: they are gated rather than tallied (Gawande, 2009). Within a phase, requirements hold together, and failing one cannot be made up elsewhere. ProRubric writes this property into the rubric once, offline, and leaves the optimizer and the served policy unchanged (Figure 1b). Section 4.1 groups the criteria into dimensions that must hold jointly, Section 4.2 adds a failure clause to each dimension, and Section 4.3 generates them.

## 4.1 CONJUNCTIVE DIMENSIONS

Under explicit aggregation every criterion earns a fixed share of the reward however hard it is to satisfy (Section 3.2), so a policy can collect easy criteria in place of the one that matters. ProRubric partitions the checklist $\mathcal{C} = \{(r_i, w_i)\}_{i=1}^{m}$ into $K$ protocol dimensions $\mathcal{I}_1, \ldots, \mathcal{I}_K$. For each criterion $i$, let $q_i \in \{0, 1\}$ denote compliance: $q_i = s_i$ for positive criteria ($w_i > 0$), and $q_i = 1 - s_i$ for penalty criteria ($w_i < 0$), ensuring penalty conditions act as hard non-violation constraints. A dimension is satisfied only when all of its criteria hold jointly, carrying their aggregated weight:

$$\tilde{s}_k = \prod_{i \in \mathcal{I}_k} q_i, \qquad \tilde{w}_k = \sum_{i \in \mathcal{I}_k} |w_i|, \qquad \tilde{R} = \frac{\sum_k \tilde{w}_k \tilde{s}_k}{\sum_k \tilde{w}_k}. \tag{2}$$

The reward is the normalized weighted sum over dimension verdicts; in practice, the judge verifies $\tilde{s}_k$ directly via synthesized protocol prompts (Section 4.3).

This alters the marginal incentive of satisfying individual criteria. Under an idealized assumption where criterion compliances hold independently with probability $p_i = \mathbb{P}(q_i = 1)$,

$$\frac{\partial\, \mathbb{E}[\tilde{R}]}{\partial p_i} \;=\; \frac{\tilde{w}_k}{\sum_\ell \tilde{w}_\ell} \prod_{j \in \mathcal{I}_k \setminus \{i\}} p_j, \qquad i \in \mathcal{I}_k. \tag{3}$$

A criterion now pays only in proportion to how likely the rest of its dimension is to hold. A fragment satisfied in a dimension the policy otherwise fails earns almost nothing, while the last missing piece of a nearly complete dimension earns the dimension's full weight: the reward stops paying for isolated fragments and starts paying for completed protocol phases.

## 4.2 FAILURE CLAUSES

Conjunction acts only on what the criteria name. No criterion forbids burying the decision under unrequested material or adding a needless high-risk test, so neither costs anything under Eq. 2. ProRubric therefore ends each dimension's description with a *failure clause*, a sentence naming a condition under which the phase is not met however much else the response contains, such as delaying urgent care. A dimension then counts only if all of its criteria hold and its failure clause does not fire. Both the conjunction and the failure clause are written into the text the judge reads, and the judge returns one verdict per dimension; nothing is computed after it answers. This matters for the controls of Section 5.3, which differ from ProRubric only in that text.

An additive deduction can be offset by adding more; a fired failure clause cannot, because nothing else in the dimension restores its credit. Both effects appear before any training in Table 1b: grouping removes the payment for naming an item, and the failure clause makes the needless test costly. Both act within a dimension only. Across dimensions the reward remains additive, so ProRubric reduces the units that can compensate for one another from $m$ to $K$ rather than removing compensation.

## 4.3 GENERATING THE DIMENSIONS

Given a prompt and its checklist, a generator model first partitions the criteria into protocol-level dimensions, then rewrites each group as a single dimension description that ends in a failure clause, and finally assigns each dimension the summed weight of its criteria (Eq. 2). A validation step requires every criterion to belong to exactly one dimension and repairs or drops rubrics that fail (Appendix A.3). We use $2 \leq K \leq 5$, matching the phases of the protocols we study; in clinical

Table 2: **Cross-domain evaluation results at 4B and 8B scales.** Deltas are relative to the untrained Base; bold and underline mark the best and second-best results.

| Method | Writing | | Dialogue | Clinical Medicine | | Scientific Reasoning | | Overall |
|---|---|---|---|---|---|---|---|---|
| | WritingBench | Creative-v3 | Arena-Hard | HealthBench | MedQA | GPQA | ResearchQA | Average |
| *Qwen3-4B* | | | | | | | | |
| Base | 49.2 | 27.9 | 12.9 | 26.1 | **65.2** | 41.1 | 51.8 | 39.2 |
| + SFT | 63.9 +14.7 | 35.5 +7.6 | **21.6** +8.7 | **35.1** +9.1 | 59.5 −5.7 | 36.0 −5.1 | **63.2** +11.4 | 45.0 +5.8 |
| + OPSD | 49.7 +0.4 | 27.7 −0.1 | 9.2 −3.7 | 25.1 −0.9 | 60.8 −4.4 | 36.4 −4.7 | 47.5 −4.3 | 36.6 −2.5 |
| + Rubric-RL | 64.6 +15.3 | 35.0 +7.2 | 17.3 +4.4 | 31.9 +5.8 | 62.6 −2.6 | 44.7 +3.6 | 62.9 +11.1 | 45.6 +6.4 |
| + RuscaRL | 63.7 +14.4 | 34.0 +6.2 | 18.2 +5.3 | 31.9 +5.8 | 62.0 −3.2 | **45.6** +4.5 | 61.3 +9.4 | 45.2 +6.1 |
| **+ ProRubric (Ours)** | **67.1** +17.8 | **37.7** +9.9 | 18.6 +5.7 | 34.7 +8.6 | 65.1 −0.1 | 43.9 +2.8 | 61.1 +9.3 | **46.9** +7.7 |
| *Qwen3-8B* | | | | | | | | |
| Base | 53.9 | 36.0 | 19.1 | 31.4 | 69.0 | 42.9 | 55.9 | 44.0 |
| + SFT | 68.2 +14.3 | **43.8** +7.8 | **30.6** +11.5 | **40.4** +9.0 | 67.8 −1.2 | 31.0 −12.0 | **67.4** +11.5 | 49.9 +5.8 |
| + OPSD | 48.1 −5.8 | 23.2 −12.8 | 9.7 −9.4 | 27.0 −4.3 | 67.7 −1.3 | 40.2 −2.7 | 47.2 −8.7 | 37.6 −6.4 |
| + Rubric-RL | 68.1 +14.2 | 38.9 +2.8 | 24.8 +5.7 | 36.4 +5.0 | 68.9 −0.2 | 49.8 +6.9 | 65.8 +9.9 | 50.4 +6.3 |
| + RuscaRL | 68.0 +14.1 | 38.6 +2.5 | 24.0 +4.8 | 36.5 +5.1 | 69.7 +0.6 | 45.5 +2.5 | 66.2 +10.3 | 49.8 +5.7 |
| **+ ProRubric (Ours)** | **69.1** +15.2 | 43.1 +7.1 | 28.6 +9.5 | 39.9 +8.6 | **69.9** +0.8 | **50.0** +7.1 | 64.9 +9.0 | **52.2** +8.2 |

triage these are rule-out, workup and disposition. Generation runs once per prompt before training (prompt in Appendix A.2) and adds no serving cost; the judge's training cost was not measured.

# 5 Experiments

We answer three questions: (1) Does ProRubric improve on Rubric-RL, and at what cost outside the domain it targets? (Section 5.2) (2) Is the aggregation the cause? (Section 5.3) (3) When does appropriateness fall, and why do only some domains collapse? (Sections 5.4 and 5.5)

## 5.1 Experimental Setup

**Models and datasets.** We use Qwen3-4B and Qwen3-8B (Yang et al., 2025) in non-thinking mode and train one policy per domain on RubricHub's medical, writing and dialogue subsets (Li et al., 2026) and on RaR-Science (Gunjal et al., 2026); inference is rubric-free. We evaluate on seven benchmarks: HealthBench (Arora et al., 2025) and MedQA (Jin et al., 2021) in medicine, WritingBench (Wu et al., 2025) and Creative-v3 (Paech, 2025) in writing, Arena-Hard v2 (Li et al., 2025) in dialogue, and GPQA-Diamond (Rein et al., 2024) and ResearchQA (Yifei et al., 2026) in science, and read appropriateness on HealthBench-consensus. DeepSeek-V4-Pro judges both axes.

**Baselines.** We compare against the untrained model (Base) and four methods trained on the same prompts: (1) Rubric-RL (Gunjal et al., 2026), the standard formulation, which scores each criterion separately and sums the weighted scores; (2) RuscaRL (Zhou et al., 2025), which uses rubrics to scaffold exploration during rollouts but keeps additive aggregation; (3) OPSD, on-policy self-distillation (Zhao et al., 2026) with the rubric as the teacher's context as in RGSD (Rezaei et al., 2026), run under a smaller generation budget; and (4) SFT on RubricHub's released SFT responses. Responses are capped at 8,192 tokens.

**Implementation details.** All RL methods use GRPO with 64 prompts per batch, 8 responses per prompt, learning rate $10^{-6}$ and 300 steps; advantages are normalized within each group, and no KL penalty is used unless stated. Training rewards come from Doubao-mini, which is not used for evaluation. Every trained method in the main table has three seeds, and all ablations are compared on one common set of items. More details are in Appendices A.4 and B.

## 5.2 Main Results

On the target domain, ProRubric recovers appropriateness without losing coverage. At 4B it reaches 36.8 against Rubric-RL's 26.0 (Figure 2, Table 3); at 8B it reaches 45.8 against 29.5. While appropriateness remains below the untrained Base (54.1), this reflects a Pareto trade-off: Base scores high on appropriateness through conservative answers that fail most requirements ($g = 26.6$), whereas

Rubric-RL gains coverage (32.1) by collapsing appropriateness (26.0). ProRubric expands the frontier, lifting coverage to 35.0 and appropriateness to 36.8 (Section 5.4 analyzes this gap).

Outside the target, the change costs little. ProRubric has the best seven-benchmark average at both scales (Table 2): over Rubric-RL the margin is $+1.3$ points at 4B, with a 95% interval of $[+0.2, +2.4]$ over three seeds, and ProRubric is ahead of both RL baselines on every paired seed at both scales. Single benchmarks move far more across seeds than the average does, so we read them as orderings rather than effect sizes (Appendix C).

The exceptions and the baselines mark the boundary. On MedQA, which uses no rubric, Rubric-RL regresses and ProRubric does not; on ResearchQA ProRubric trails Rubric-RL at both scales, since there breadth is itself the goal (Section 5.5). RuscaRL keeps additive aggregation and lands where Rubric-RL lands, OPSD reflects its smaller generation budget, and SFT pays the largest cost outside its domain, $-12.0$ on GPQA-Diamond at 8B (Appendix D.2).

### 5.3 Changing Only the Aggregation Moves the Collapse

Section 3 read the incentive off the reward before training; here we intervene on it. With the criterion text held fixed, every change to how the criteria are aggregated moves appropriateness up, and every change to anything else leaves the collapse in place (Table 3). Two controls change only the aggregation, keeping every criterion verbatim: *raw-AND* applies ProRubric's grouping, and *implicit aggregation* (Gunjal et al., 2026) scores the checklist whole. Two vary ProRubric itself: *Graded* gives four-level credit per dimension, and $K{=}1$ merges all dimensions into one. Three change something else: a *length-matched* variant (half the response limit), a *KL penalty*, and a *weighted* rubric (Appendix A.5).

Table 3: **Medical ablations (4B).** $c$: appropriateness; $g$: coverage; both by DeepSeek-V4-Pro, averaged over three seeds on the common set (untrained Base: 54.1 / 26.6).

| Reward | $c$ | $g$ |
|---|---|---|
| *ProRubric variants* | | |
| ProRubric | 36.8 $\pm$2.7 | 35.0 $\pm$1.4 |
| − failure clauses | 34.5 $\pm$1.1 | 34.9 $\pm$0.3 |
| partial credit (Graded) | 36.3 $\pm$1.8 | 35.1 $\pm$0.9 |
| one dimension ($K{=}1$) | 36.9 $\pm$1.9 | 33.5 $\pm$1.2 |
| *Same criteria, other aggregation* | | |
| additive (Rubric-RL) | 26.0 $\pm$2.2 | 32.1 $\pm$1.2 |
| grouped, verbatim (raw-AND) | 34.9 $\pm$3.6 | 33.8 $\pm$1.1 |
| holistic score (Implicit) | 36.7 $\pm$0.1 | 34.9 $\pm$0.1 |
| *Other changes* | | |
| shorter (Length-matched) | 27.1 $\pm$1.8 | 30.4 $\pm$1.4 |
| reweighted (Weighted) | 25.3 $\pm$1.5 | 32.0 $\pm$1.8 |

**Changing the aggregation moves the collapse.** Without rewriting anything, raw-AND recovers most of the distance to ProRubric in medicine, though with higher variance ($\pm$3.6 vs $\pm$2.7) and lower coverage (33.8 vs 35.0); implicit aggregation of the verbatim checklist matches ProRubric within confidence intervals. Both reproduce at 8B over three seeds (Appendix D.2). Within ProRubric, partial credit per dimension is not behind binary credit, so we do not claim that hard conjunction is the better operator: what these variants share is that no criterion is scored and summed alone (Eq. 3). $K{=}1$ matches ProRubric on appropriateness but has lower coverage (33.5 against 35.0), consistent with a weaker learning signal: about two thirds of its sampled groups receive identical rewards on every seed.

**Changing anything else does not.** Length is the vehicle, not the cause. The length-matched variant, 42% shorter, leaves appropriateness where it was (Figure 3a), and under an equal 2,048-token budget ProRubric wins 83% and 79% of decided blind pairs against Rubric-RL in medicine and science (Appendix B.3). A KL penalty recovers appropriateness only by stopping learning: at $\beta = 0.04$ both rewards approach the untrained model on both axes and in length (Figure 3b). Reweighting the rubric leaves Rubric-RL at 25.3, and the science rubric, already weighted, loses appropriateness all the same. SFT on RubricHub's released responses ends above the untrained model (66.3 against 54.1), so the data collapses appropriateness only once it becomes an additive reward. The ranking is nearly unchanged under GPT-5.6-luna (Appendix D.2).

**Where the attribution holds.** In dialogue it holds as in medicine: implicit aggregation beats Rubric-RL on all three seeds, close to ProRubric (Table 14). In science, however, the limits of verbatim aggregation emerge: implicit aggregation, raw-AND, and rewriting criteria one to one each

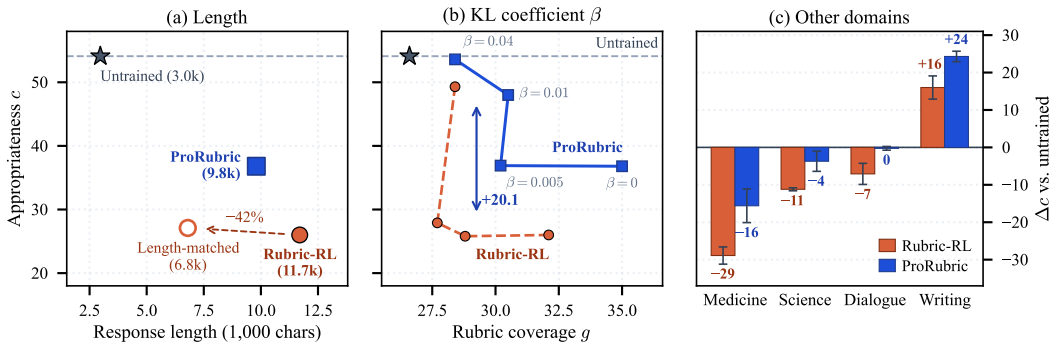


Figure 3: **Analysis of confounding factors and cross-domain generalization.** (a) Length matching does not restore appropriateness. (b) KL penalty. (c) Appropriateness retention across domains.

recover at most 3.2 points, against ProRubric's 7.5 (three-seed mean, at least 5.0 on every seed). Where criteria define substantive derivations rather than presence checks, verbatim concatenation overburdens the judge; only structured protocol grouping and synthesis reliably lift appropriateness.

### 5.4 When the Collapse Comes, and Where ProRubric Stops

**Under additive aggregation, appropriateness collapses early.** In retrained medical models (4B, three seeds, 300 HealthBench-consensus prompts; Figure 5 in Appendix D.2), Rubric-RL falls from 55.8 to 35.9 within 50 steps and ends at 23.8, while ProRubric remains at 50.3 at step 50 and ends at 35.3; at every checkpoint every Rubric-RL seed lies below every ProRubric seed. Without per-criterion credit the collapse is slower and smaller, but not stopped, as analyzed next.

**It stops where the dimensions meet.** Across dimensions the reward is still additive (Section 4.2), so adding material still pays: ProRubric's responses run to 9.8k characters against the untrained model's 3.0k, and added content, named or substantive, costs 17 to 38 points of the redundancy subscore (25 to 38 for added substance; Table 15b). This is the residual gap: ProRubric satisfies most of its dimensions but stays below the untrained model on appropriateness, and without a rubric the two judges disagree on whether it beats the untrained model (Table 7). It loses less than Rubric-RL because what it adds is substantive, but it moves along the same mechanism rather than escaping it.

**Why protocol dimensions go beyond verbatim grouping.** Verbatim grouping (raw-AND) recovers 8.9 points in medicine, while removing ProRubric's failure clauses costs 2.3 points (Table 3). Raw-AND also leaves criteria uncurated: higher cross-seed variance ($\pm 3.6$ against $\pm 2.7$), lower coverage (33.8 against 35.0), and in science it recovers little ($+3.2$ against $+7.5$; Section 5.3). Synthesized dimensions give non-compensable failure conditions and a cleaner training signal.

### 5.5 What a Rubric Pays For Decides Whether a Domain Collapses

Section 3.3 read the incentive off the medical rubric. Applied to all four domains, the same analysis accounts for most of the cross-domain pattern. Under explicit aggregation, appropriateness falls by 28.9 in medicine, 11.2 in science and 7.1 in dialogue, and rises by 16.0 in writing (Figure 3c; three-seed means on 200 prompts per domain, Table 15a). The cause is the same everywhere, additive aggregation paying for added material, but it has two routes: where criteria are mention-style the cheapest addition is hollow, and where they are substantive the policy must elaborate for real.

The first route is naming. On the untrained model's answers to 300 training prompts per domain (Table 15b), naming a checklist item without analyzing it earns, under the Rubric-RL reward, 32% of what adding unrequested substance earns in medicine and 13% in dialogue, and nothing detectable in writing (7%) or science ($-8\%$); under ProRubric's reward the shares fall to 6%, $-14\%$, $-10\%$ and $-23\%$. Medicine, where both routes are open, collapses hardest.

The second route runs through substance, and it explains science. Science's criteria are verifiable requirements, such as deriving a threshold energy from the four-momentum invariant, that naming cannot satisfy: 195 of 300 naming edits leave the reward unchanged. Still, every rubric pays for correct, relevant material that no criterion asked for ($+22.7$ in medicine, $+18.9$ in writing, $+17.9$

in dialogue, $+10.8$ in science), and in every domain it costs more of the redundancy sub-score than naming does, 25 to 38 points against 17 to 29. Science loses appropriateness by this route alone.

Why writing improves remains open; writing policies may integrate rather than append material (Appendix D.5). Disentangling tasks from rubric shapes requires further study (Appendix D.2).

## 6 Related Work

**Rubrics as rewards.** Fine-grained checks go back to behavioral testing (Ribeiro et al., 2020), and rule-based feedback traces to Constitutional AI (Bai et al., 2022b) and RLAIF (Lee et al., 2024; Yuan et al., 2024). Prometheus (Kim et al., 2024) and HealthBench (Arora et al., 2025) made rubrics a standard interface for LLM-based evaluation. Checklist and rubric rewards carry them into RL (Gunjal et al., 2026; Viswanathan et al., 2025; Huang et al., 2025; Jia et al., 2026), and rubric datasets and generators scale up criteria (Li et al., 2026; Liu et al., 2026; Shen et al., 2026). Rubrics also enter training in other ways: RuscaRL (Zhou et al., 2025) uses them as scaffolding during rollouts, RISE-RL (Hou et al., 2026) selects trajectories to explore by the criteria a policy repeatedly misses, and rubric-guided self-distillation (Rezaei et al., 2026), built on on-policy self-distillation (Zhao et al., 2026), gives them to a teacher. These works change what criteria ask for and where they enter training; as a reward, their verdicts are still a weighted sum, the step we study.

**Aggregating the criteria.** In RaR, Gunjal et al. (2026) compare explicit aggregation, a normalized weighted sum of per-criterion verdicts, with implicit aggregation, in which the judge reads the whole checklist and returns one score. Recent work questions the weighted sum. Lv et al. (2026) observe that criteria scored as independent utilities pass credit past unmet prerequisites, which they call *false credit propagation*, and suppress it through a rubric graph annotated with prerequisite and activation relations. Rubric Dropout (Yang et al., 2026) drops criteria at random during training to curb reward hacking, and SRaR (Xie et al., 2026) attributes rubric items to individual steps of mathematical reasoning instead of folding them into one scalar. ProRubric needs no annotated relations and keeps one response-level reward: it groups a plain checklist into dimensions that must hold jointly, and we measure what the weighted sum costs on criteria the policy never trained against.

**Reward hacking.** Over-optimizing a proxy reward is an instance of Goodhart's law (Goodhart, 1975), extensively documented in RLHF (Ouyang et al., 2022; Bai et al., 2022a; Skalse et al., 2022; Gao et al., 2023) and DPO (Rafailov et al., 2024), where reward model ensembles mitigate but do not eliminate gaming (Eisenstein et al., 2024); in text generation it often surfaces as length bias (Singhal et al., 2024; Dubois et al., 2024; OpenAI, 2026). In rubric RL, Mahmoud et al. (2026) report the reversal we start from: rubric-based verifiers prefer the RL checkpoint while rubric-free judges prefer the base model, with gains concentrated in completeness and presence-based criteria, and attribute it to rubrics that leave important failure modes unspecified. CHERRL (Wang et al., 2026) reproduces reward hacking by injecting biases into judges. We trace the reversal elsewhere: with the judge and the criterion text fixed, changing only the aggregation moves it in medicine and dialogue.

## 7 Conclusion

Rubric-based RL writes quality down as criteria and turns them back into a number; we find that the second step largely decides whether optimizing the number helps. Under a weighted sum, cheap criteria compensate for a missed critical one, and the policy learns to do exactly that: coverage rises while appropriateness on held-out physician criteria falls below the untrained model. With criterion text fixed, changing only aggregation moves this collapse in medicine and dialogue; in science, wording matters as well. ProRubric groups criteria offline into dimensions that must hold jointly, lifting appropriateness without losing coverage or touching the optimizer: reward validity depends on aggregation as much as verification.

**Limitations and future work.** Clinical consultation unfolds across multiple turns where clarifying questions are essential. This work evaluates single-turn responses using an LLM judge calibrated against physicians (Appendix B.1) rather than clinical trials. Extending protocol rubrics to interactive dialogues and human studies remains a key next step.

## ETHICS STATEMENT

This work studies how rubric rewards shape the behavior of language models. The medical experiments use public benchmarks (HealthBench, MedQA) and involve no private data and no human subjects. The trained models are research artifacts; they are not validated for clinical use and should not be used to give medical advice or to triage patients.

## REPRODUCIBILITY STATEMENT

The medical prompt that generates ProRubric's dimensions and the binary training judge's prompt are given verbatim in Appendix A.2. Code and all judge prompts are in the supplementary material. Training hyperparameters and baseline configurations are in Appendix A.4 (Table 4), and benchmark scoring, judge configurations, common evaluation sets and per-seed results in Appendices B and C.

## AI USE STATEMENT

Language models generate ProRubric's dimensions and serve as judges, as described in the text. They also helped polish prose and format LaTeX; the authors verified every claim.

## REFERENCES


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. URL https://arxiv.org/abs/2505.08775.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*, 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022b. URL https://arxiv.org/abs/2212.08073.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.04475.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate Reward Hacking. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2312.09244.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Atul Gawande. *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books, 2009.

Charles A. E. Goodhart. Problems of Monetary Management: The U.K. Experience. *Papers in Monetary Economics, Reserve Bank of Australia*, 1, 1975.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=c1bTcrDmt4.

Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K. Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chi Tong, Kavin Karthik, Arnav Dugar, Akshay Jagadeesh, Khaled Saab, Johannes Heidecke, Ashley Alexander, Nate Gross, and Karan Singhal. HealthBench Professional: Evaluating Large Language Models on Real Clinician Chats. *arXiv preprint arXiv:2604.27470*, 2026. URL https://arxiv.org/abs/2604.27470.

Jinkun Hou, Zhuo Liu, Huimin Ren, Hongsheng Xin, Pan Zhou, and Kun Zhan. RISE-RL: Rubric-Informed Selective Exploration for Open-Ended Reinforcement Learning. *arXiv preprint arXiv:2608.09123*, 2026. URL https://arxiv.org/abs/2608.09123.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. Reinforcement Learning with Rubric Anchors. *arXiv preprint arXiv:2508.12790*, 2025. URL https://arxiv.org/abs/2508.12790.

Ruipeng Jia, Yunyi Yang, Wen Wang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Open Rubric System: Scaling Reinforcement Learning with Pairwise Adaptive Rubric. *arXiv preprint arXiv:2602.14069*, 2026. URL https://arxiv.org/abs/2602.14069.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. doi: 10.48550/arXiv.2310.08491. URL https://openreview.net/forum?id=8euJaTveKw.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26874–26901. PMLR, 2024. doi: 10.48550/arXiv.2309.00267. URL https://proceedings.mlr.press/v235/lee24t.html.

Sunzhu Li, Jiale Zhao, Huimin Ren, Zhenlin Wei, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang, and Wei Chen. RubricHub: A Comprehensive and Highly Discriminative Rubric Dataset via Automated Coarse-to-Fine Generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31320–31344, 2026. doi: 10.18653/v1/2026.acl-long.1445. URL https://aclanthology.org/2026.acl-long.1445/.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From Crowdsourced Data to High-quality Benchmarks: Arena-Hard and Benchbuilder Pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34209–34231. PMLR, 2025. doi: 10.48550/arXiv.2406.11939. URL https://proceedings.mlr.press/v267/li25h.html.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://arxiv.org/abs/2510.07743.

Can Lv, Mingju Chen, Heng Chang, and Shiji Zhou. Mitigating False Credit Propagation: Probabilistic Graphical Reward Aggregation for Rubric-Based Reinforcement Learning. *arXiv preprint arXiv:2606.03361*, 2026. URL https://arxiv.org/abs/2606.03361.

Anas Mahmoud, MohammadHossein Rezaei, Zihao Wang, Anisha Gunjal, Bing Liu, and Yunzhong He. Reward Hacking in Rubric-Based Reinforcement Learning. *arXiv preprint arXiv:2605.12474*, 2026. URL https://arxiv.org/abs/2605.12474.

OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

OpenAI. GPT-5.5 System Card. Technical report, OpenAI, April 2026. URL https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.02155.

Samuel J. Paech. EQ-Bench Creative Writing Benchmark v3. Software repository, version 3, 2025. URL https://github.com/EQ-bench/creative-writing-bench. Accessed September 9, 2026.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.02900.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2311.12022.

MohammadHossein Rezaei, Anas Mahmoud, Zihao Wang, Utkarsh Tyagi, Advait Gosai, Razvan-Gabriel Dumitru, Aakash Sabharwal, Bing Liu, and Yunzhong He. Rubric-Guided Self-Distillation: Post-Training Without Rubric Verifiers. *arXiv preprint arXiv:2606.12507*, 2026. URL https://arxiv.org/abs/2606.12507.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4902–4912, 2020. URL https://aclanthology.org/2020.acl-main.442/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

William F. Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking Rubric Generation for Improving LLM Judge and Reward Modeling for Open-ended Tasks. *arXiv preprint arXiv:2602.05125*, 2026. URL https://arxiv.org/abs/2602.05125.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A Long Way to Go: Investigating Length Correlations in RLHF. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2310.03716.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists Are Better Than Reward Models For Aligning Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 114728–114754, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a6837c1dd021f76f1b4098e3722052a8-Abstract-Conference.html.

Xuekang Wang, Zhuoyuan Hao, Shuo Hou, Hao Peng, Juanzi Li, and Xiaozhi Wang. Reproducing, Analyzing, and Detecting Reward Hacking in Rubric-Based Reinforcement Learning. *arXiv preprint arXiv:2606.04923*, 2026. URL https://arxiv.org/abs/2606.04923.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. WritingBench: A Comprehensive Benchmark for Generative Writing. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2503.05244.

Weichu Xie, Haozhe Zhao, Wenpu Liu, Yongfu Zhu, Liang Chen, Minghao Ye, Zirong Chen, Yuqi Xu, Shuai Dong, Ziyue Wang, Xinbo Xu, Kean Shi, Ruoyu Wu, Xiaoying Zhang, Wenqi Shao, Baobao Chang, Nan Duan, and Jiaqi Wang. Step-wise Rubric Rewards for LLM Reasoning. *arXiv preprint arXiv:2605.17291*, 2026. URL https://arxiv.org/abs/2605.17291.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Minglai Yang, Xinyu Guo, Utkarsh Tyagi, Mian Zhang, Razvan Dumitru, Sunjie Hou, Yunzhong He, Daniel Yue Zhang, and Ying Liu. Rubric Dropout: A Simple Way to Mitigate Reward Hacking in Rubric-as-Reward RL. *arXiv preprint arXiv:2608.11669*, 2026. URL https://arxiv.org/abs/2608.11669.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. ResearchQA: Evaluating Scholarly Question Answering at Scale Across 75 Fields with Survey-Mined Questions and Rubrics. *Transactions of the Association for Computational Linguistics*, 14:1365–1389, 2026. doi: 10.1162/tacl.a.732. URL https://aclanthology.org/2026.tacl-1.62/.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-Rewarding Language Models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2401.10020.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2601.18734.

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 (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, Hengtong Lu, Wei Chen, Yan Xie, and Mingli Song. Breaking the Exploration Bottleneck: Rubric-Scaffolded Reinforcement Learning for General LLM Reasoning. *arXiv preprint arXiv:2508.16949*, 2025. URL https://arxiv.org/abs/2508.16949.

**Appendix Roadmap.** This appendix provides reproducibility details, complete empirical evaluations, sensitivity analyses, and qualitative examples:

- **Section A: Implementation Details and Prompts** – Training data curation, offline protocol generation templates (Box A.1), reward judge specification (Box A.2), schema validation, and training configurations (Table 4).
- **Section B: Evaluation Protocols and Calibrations** – Benchmark configurations, human-judge calibration (Table 5), sample sizes (Table 6), and rubric-free blind pairwise protocols (Table 7, Figure 4).
- **Section C: Full Numerical Results and Ablations** – Per-seed cross-domain results (Table 8), complete medical ablation tables under dual judges (Table 9, Table 10), cross-model ranking validation (Table 11), and cross-domain appropriateness probes (Table 14, Table 15).
- **Section D: Mechanistic Analyses and Diagnostics** – Incentive audit across aggregations (Table 13), training dynamics trajectories (Figure 5), KL regularization sweep, fixed appropriateness criterion analysis, and text placement experiments.
- **Section E: Qualitative Case Studies and Protocol Analysis** – Clinical consultation walkthrough, protocol dimensions, and trade-off analysis on exhaustive retrieval.

# A Implementation Details

## A.1 Data Sources and Training Datasets

We train one policy per domain using RubricHub's medical, writing and dialogue subsets (Li et al., 2026) and RaR-Science (Gunjal et al., 2026). For RubricHub, prompts are deduplicated by hash and filtered to retain rubrics with 5–60 nonempty criteria and positive total weight. For science, we train on the 18,333 prompts of the RaR-Science training split. All model evaluations are conducted out-of-domain on standard benchmarks or against held-out physician criteria (Section 5.1). Rubric-RL and ProRubric are trained on identical prompts, ensuring that all observed differences originate from reward aggregation rather than data curation.

## A.2 Rubric Generation Prompts and Protocol Templates

Offline generation turns atomic rubrics of 20–30 items into two to five protocol-level dimensions with DeepSeek-V4-Pro as the generator. Generation uses temperature 0, a limit of 3,000 output tokens, and at most two retries when the output fails schema validation. The generator's proposed weights are retained as metadata; training weights are the absolute-weight sums of Section 4.1.

Box A.1 gives the full generation prompt. It asks the generator for two to five dimensions, each stating what a good answer achieves and what would make it fail, judgeable only from the whole answer, with every atomic item assigned to exactly one dimension and a weight equal to the sum of its items' weights. The {anchor} slot receives the training example's reference rubric as distributed with the dataset; in all four training sets this is the atomic rubric itself, re-rendered as a list, and no reference-answer notes are present, so the anchor adds nothing beyond the atomic items and no evaluation criterion is shown to the generator. Box A.2 gives the training reward judge's prompt: the judge reads the whole rubric (the atomic criteria, or ProRubric's dimensions) and returns one binary verdict per item; raw-AND uses the same judge prompt, with each group's verbatim items under an instruction that all must hold. The writing and dialogue generation prompts differ in their task description; the science prompt keeps the medical wording and differs only in weighting items by magnitude and folding pitfall items into failure conditions. These prompts and those of the Implicit, Graded and $K{=}1$ judges are in the supplementary material.

**Box A.1: Prompt for Offline Dimension Generation (ProRubric, medicine; verbatim)**

```
You are designing a compact, protocol-level grading rubric for a
medical Q&A answer, to replace an atomic checklist of many small
"mentions X" items with 2-5 coarse-grained criteria that each
describe what a GOOD answer ACHIEVES on one dimension.

# Question / conversation
{conversation}

# Existing atomic rubric items (for reference only - do not just copy
these into your descriptions as a checklist)
{atomic_items}

# Anchor for resolving contradictions (an expert-reviewed reference
rubric for this same question - some atomic items above may be
redundant, mutually exclusive, or simply wrong; use this anchor to
decide what a correct answer should actually contain)
{anchor}

# Task
Design between 2 and 5 protocol-level criteria.  Each criterion is a
paragraph, not a checklist item.  Each criterion must:
1.  state what a good answer ACHIEVES on that dimension for THIS
SPECIFIC question, and what would make an answer FAIL that dimension;
2.  be judgeable only by reading and understanding the whole answer
holistically - NOT by scanning for a keyword or a single sentence;
3.  NOT be phrased as a list of "mentions A, B, C" terms to check
off;
```

```
4.  if two or more atomic items below contradict each other (e.g.
reward mutually exclusive answers), resolve the contradiction
explicitly using the anchor - the description should state what the
CORRECT position is, not just note that people disagree;
5.  together with the other criteria, have every atomic item index
from 1 to {n} assigned to exactly one criterion (pick whichever
criterion is closest in topic, even after you rewrite its language
into holistic phrasing);
6.  carry a weight equal to the sum of the weights of the atomic
items assigned to it.

# Output format
Return ONLY a JSON object, no markdown fences, no other text:
{"criteria":  [{"name":  "<=6 words", "description":  "2-5 sentences,
self-contained, states what ACHIEVES / FAILS this dimension",
"weight":  <number>, "atomic_indices":  [<1-based ints into the
atomic list above>]}, ...]}
The atomic_indices arrays across all criteria must union to exactly
{1, ..., {n}} with no repeats and no omissions.
```

**Box A.2: Prompt of the Training Reward Judge (verbatim)**

```
You are a strict binary rubric judge.  Evaluate the assistant
response against every criterion independently.  A criterion is
satisfied only when the response meets all of its requirements.
Negative-weight criteria describe undesirable behavior:  return true
when that behavior is present.  Return ONLY one JSON object with
exactly one field, "satisfied", whose value is an object mapping each
criterion index (as a string, e.g.  "1", "2", ...)  to a boolean,
covering every index exactly once.  Do not include explanations,
markdown, or extra keys.

PROMPT:
{prompt}

ASSISTANT RESPONSE:
{response}

CRITERIA:
index":  1, "criterion":  ..., "weight":  ..., "polarity":  ...}, ...
```

### A.3 Validation, Repair, and Filtering

Each generation is checked for the required fields and for a complete partition of the criteria. If retries fail, a structural repair drops invalid or duplicate assignments and assigns an omitted criterion to the nearest group by index, then recomputes the weights; it does not regenerate descriptions. A rubric is kept only if every criterion occurs exactly once, every group is nonempty, there are two to five dimensions, and every weight is positive; the released data record how many rubrics were kept, repaired and excluded, and preserve the original rubrics. These checks are structural: they do not verify the factual correctness or semantic coverage of a description.

### A.4 Training Configurations and Baseline Differences

Table 4 summarizes the training configurations. Rubric-RL and ProRubric share the optimizer and rollout settings; the main difference is the rubric carried by the training data. Both disable KL reward shaping and KL loss, use a constant learning-rate schedule with ten warmup steps, and sample with temperature 1 and top-$p = 1$. The GRPO clipping parameters are 0.20 and 0.28 for the lower and upper bounds. Checkpoints are saved and validated every 50 steps (every 100 steps for two early 4B models). Every model trained with the rubric judge's reward also applies DAPO-style overlong shaping to its training rewards: a response longer than 4,096 tokens (the 8,192-token response limit minus a 4,096-token buffer) loses reward linearly, by up to 0.5 at the limit; validation scores are

Table 4: **Training hyperparameters and generation budgets of each method.**

| Setting | Rubric-RL / ProRubric | RuscaRL | OPSD | SFT |
|---|---|---|---|---|
| Training steps | 300 | 300 | 485 | 1,224 (3 epochs) |
| Prompts per batch | 64 | 512 | 128 | 64 |
| Responses per prompt | 8 | 1 | 1 | — |
| Mini-batch size | 32 | 256 | 32 | — |
| Learning rate | $10^{-6}$ | $10^{-6}$ | $4.2 \times 10^{-6}$ | $10^{-5}$ |
| Prompt limit | 4,096 | 6,144 | 4,096 | 20,000 |
| Response limit | 8,192 | 8,192 | 2,048 | |
| Training-time rubric judge | Yes | Yes | No | No |
| Overlong shaping (buffer / factor) | 4,096 / 0.5 | 4,096 / 0.5 | — | — |

unshaped. It is identical across Rubric-RL, ProRubric, RuscaRL and the ablations and is active in every such run; it rarely binds under the larger KL coefficients, where responses stay short.

Step counts in Table 4 are those of medical training. In the writing and dialogue domains RuscaRL exhausts its training set after 296–297 steps, so the evaluated checkpoint is the last one saved, step 250; the same holds for every variant built on RuscaRL's scaffolding. OPSD is evaluated at step 400 in writing, 350–352 in dialogue and 700 in science. In those two domains RuscaRL is therefore compared with Rubric-RL and ProRubric at 250 steps against 300. RuscaRL preserves contiguous scaffold groups during training and uses a larger prompt allowance to accommodate the additional context. OPSD instead supplies rubric information to a teacher and optimizes a self-distillation loss following RGSD (Rezaei et al., 2026); the rubric judge is used for validation, not to provide its training reward. SFT fine-tunes on RubricHub's released SFT corpus (each response the best of six samples by rubric score; 26,162 of its 26,194 rows, after dropping empty rows and rows over the 20,000-token limit), whose prompts span all four of our domains, science included, mixed into one training set, following the supervised setup of RISE-RL (Hou et al., 2026) (learning rate $10^{-5}$ with 5% warmup, three epochs) with weight decay 0.1; it produces one model per seed that serves all four domains, whereas every RL policy is trained per domain. Its training examples omit the empty `<think></think>` block that the Qwen3 chat template inserts at inference; the loss targets are unchanged. The baseline results describe these configurations, not each method's best.

### A.5 Ablation Construction

The raw-AND variant uses the same group assignments and absolute-weight sums as ProRubric, but presents the original items under a joint-satisfaction instruction. Negative-weight items become behaviors that must not occur. The Graded variant keeps the rewritten descriptions and scores 0 to 3, divided by 3 before aggregation. The single-dimension variant ($K{=}1$) rewrites the whole checklist into one ProRubric-style description with its failure clause.

The failure-clause ablation removes sentences matching explicit failure expressions from the rewritten text. Unmatched expressions remain, and descriptions shortened below 40 characters are restored. It therefore tests removal of a class of explicit clauses, not complete elimination of failure semantics. The added appropriateness criterion assesses directness, scope, and adaptation to the user. Its wording is the same for Rubric-RL and ProRubric, and its weight is the total absolute atomic weight divided by the corresponding number of ProRubric dimensions. For ProRubric, this equals the mean existing dimension weight. The new criterion is scored independently and included in the weighted reward; failure on it does not zero the scores of the other criteria.

Implicit aggregation keeps Rubric-RL's checklist and has the training judge read it whole and return one holistic score from 1 to 10 for the response. The length-matched variant lowers Rubric-RL's response limit from 8,192 to 4,096 tokens. The KL variants add a KL loss toward the reference policy (low-variance estimator) with $\beta \in \{0.005, 0.01, 0.04\}$ to both Rubric-RL and ProRubric. The weighted rubric keeps every criterion's text and count but multiplies the weight of criteria tagged critical by 3 (13.5% of criteria), sets formatting and duplicate criteria to 0 (7.2% and 11.3%) and needless-test criteria to $-1$ (2.6%); each criterion is still scored and summed on its own. In science, the rewritten-criteria variant restates each criterion one to one in the style of ProRubric's dimen-

Table 5: **Agreement with physicians on HealthBench's meta-evaluation (balanced F1).**

| Metric | Emergency referrals | Global health | Communi-cation | Context seeking | Hedging | Health data | Complex responses | All criteria |
|---|---|---|---|---|---|---|---|---|
| Criteria ($n$) | 6 | 3 | 4 | 4 | 9 | 4 | 4 | 34 |
| DeepSeek-V4-Pro | 0.65 | 0.69 | 0.63 | 0.61 | 0.61 | 0.71 | 0.45 | 0.62 |
| Physicians | 0.65 | 0.65 | 0.62 | 0.64 | 0.66 | 0.74 | 0.57 | 0.65 |

sions, keeping its weight and explicit aggregation. The generator control regenerates ProRubric's dimensions with Doubao-lite instead of DeepSeek-V4-Pro under the same rules.

# B Evaluation Protocols

## B.1 Benchmark Scoring and Judge Configurations

HealthBench evaluation uses the upstream grading template with one judgment per criterion. For each question, the score is the sum of satisfied criteria's signed weights divided by the sum of positive weights, followed by averaging over questions. The same normalization is used for the consensus evaluation. WritingBench uses query-dependent criteria; Creative Writing Benchmark v3 (Creative-v3) uses its isolated rubric component without the Elo/Glicko ranking component. Arena-Hard v2 uses the upstream double-order comparison protocol and category-specific reference responses. ResearchQA scores rubric coverage on a five-level scale, maps the levels to $[0, 1]$, and averages first over criteria and then over questions. We report all of these scores multiplied by 100.

DeepSeek-V4-Pro replaces the upstream judges for these model-judged benchmarks. A second evaluator, Doubao-lite, is from the same family as the training reward judge; we use it only to compare judges. ProRubric's dimensions are generated by DeepSeek-V4-Pro, which is also the judge, so the same model proposes the dimensions and scores the responses trained on them. Regenerating the dimensions with Doubao-lite moves appropriateness by less than two points (Appendix D.2); raw-AND and implicit aggregation rewrite nothing and are unaffected. All judges, generation seeds and scorer settings are fixed across evaluations, so paired comparisons are reproducible.

**Agreement with physicians.** We ran DeepSeek-V4-Pro on HealthBench's meta-evaluation set (29,511 response–criterion pairs over the 34 consensus criteria, with 60,896 physician labels) using the official grader prompt, and scored it with the official code: balanced F1 against each physician label, averaged within each criterion and then over criteria. It reaches $0.62$, against $0.65$ for the average physician and $0.71$ for GPT-4.1 as reported by Arora et al. (2025) (Table 5). It is above the physician average in three of the seven themes and furthest below it on complex responses. Every method is scored by the same judge, so this bounds how well appropriateness tracks physicians, not the comparison between methods.

## B.2 Paired Evaluation Sets and Statistical Estimation

Every comparison is paired on the intersection of questions that all compared models have been scored on without evaluator failure, so the set shrinks as models are added and no comparison mixes subsets (Table 6). For the ablations this gives 3,140 cases on the consensus axis and 4,392 on the coverage axis, from pools of 3,671 and 5,000; SFT and the 8B controls are paired on their own intersection with Base, Rubric-RL and ProRubric. The $g_{\mathrm{pro}}$ column of Table 10 uses its own set: the $4{,}138$ of $4{,}563$ HealthBench-full prompts on which every seed of every variant in Table 3 has a DeepSeek-V4-Pro verdict for every criterion. Other rows use the part of this set they cover.

At 4B, the WritingBench, Creative-v3, GPQA-Diamond and ResearchQA results of Base, and of Rubric-RL and ProRubric at seed 42, average three evaluations of the same checkpoint; every other entry is a single evaluation. These repeated evaluations are not the three trained seeds.

Table 6: **Sample sizes of common evaluation sets across benchmarks and judges.**

| Medicine | Science | Writing | Dialogue |
|---|---|---|---|
| HealthBench-consensus (3,140) | GPQA-Diamond (198) | WritingBench (1,000) | Arena-Hard v2 (748) |
| HealthBench-full, lite (4,392) | ResearchQA (702) | Creative-v3 (96) | |
| HealthBench-full, pro (4,138) | | | |
| MedQA (1,273) | | | |

Table 7: **Full win/tie/loss counts and win rates for rubric-free pairwise comparisons.**

| | DeepSeek-V4-Pro | | GPT-5.6-luna | |
|---|---|---|---|---|
| **Domain** | **W / T / L** | **Win rate (%)** | **W / T / L** | **Win rate (%)** |
| *Base vs. Rubric-RL* | | | | |
| Medicine | 136 / 93 / 71 | 65.7 $\pm$13.2 | 229 / 55 / 16 | 93.6 $\pm$2.3 |
| Science | 3 / 35 / 262 | 1.2 $\pm$1.2 | 151 / 134 / 15 | 90.3 $\pm$9.0 |
| Dialogue | 45 / 117 / 138 | 24.3 $\pm$5.8 | 81 / 114 / 99 | 44.9 $\pm$6.8 |
| Writing | 0 / 8 / 292 | 0.0 $\pm$0.0 | 4 / 30 / 266 | 1.4 $\pm$1.6 |
| *Base vs. ProRubric* | | | | |
| Medicine | 64 / 104 / 132 | 33.1 $\pm$11.5 | 162 / 98 / 40 | 80.2 $\pm$6.2 |
| Science | 4 / 43 / 253 | 1.6 $\pm$0.6 | 77 / 171 / 52 | 59.7 $\pm$2.1 |

### B.3 Rubric-Free Comparisons and Length Controls

The rubric-free evaluator receives the user prompt and two anonymous responses, with neither training rubric shown. It judges usefulness, relevance, fit to the user's circumstances, and unnecessary verbosity. Each pair is evaluated in both orders. A stable win or loss requires agreement across the two orders; ties and inconsistent decisions are excluded from the conditional win-rate denominator. For every comparison plotted in Figure 4, we report the conditional win rate over stable, non-tied decisions; order consistency stays above 90%.

The equal-budget evaluation regenerates both policies' responses under the same 2,048-token ceiling; this does not equalize realized lengths, but neither policy can generate beyond it.

## C Full Numerical Results and Ablations

### C.1 Per-Domain and Per-Seed Results

Every trained-model score in Tables 2 and 3 is a mean over three training seeds. This section gives the 4B per-seed values: Table 8 for the cross-domain benchmarks and Table 9 for the medical ablations; Table 11 gives the third-family cross-check.

**Reading the tables.** A slash separates seeds, in the order 42 / 43 / 44; a dash marks an entry we have not evaluated. Absolute scores are not comparable across judges (Appendix D.2).

**Seed variation.** Single benchmarks move far more across seeds than the seven-benchmark average. On GPQA-Diamond (198 questions) the paired difference between ProRubric and Rubric-RL at 4B is $-3.5$, $-5.6$ and $+6.6$; on Arena-Hard it ranges from $+0.2$ to $+2.3$ ($+0.2$, $+1.4$, $+2.3$); on HealthBench it keeps its sign but varies fourfold ($+5.2$, $+1.9$, $+1.3$). We therefore report the average as a paired per-seed difference and read single benchmarks as orderings, not effect sizes.

The mean improvement is 10.8 points under DeepSeek-V4-Pro and 6.1 under Doubao-lite over the three seeds shown. These averages are not confidence intervals. They are computed from unrounded per-seed scores, so adding the displayed one-decimal values can differ from the stated mean by 0.1.

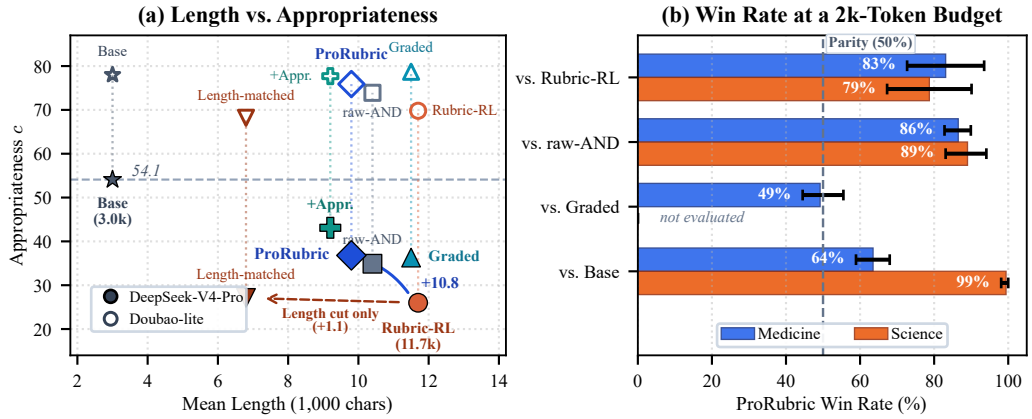


Figure 4: **Response length and budget constraints.** (a) Appropriateness vs. mean response length (filled: DeepSeek-V4-Pro; open: Doubao-lite). (b) Pairwise win rates at a 2,048-token budget (DeepSeek-V4-Pro; vs. Rubric-RL: three seeds, others: seed 42).

Table 8: **Per-seed cross-domain evaluation results at 4B scale (seeds 42, 43, and 44).**

| Benchmark | Rubric-RL | | ProRubric | | Paired Δ |
|---|---|---|---|---|---|
| | per seed | mean | per seed | mean | |
| *Writing* | | | | | |
| WritingBench | 66.3 / 63.6 / 63.7 | 64.6 ±1.5 | 67.9 / 66.8 / 66.5 | 67.1 ±0.7 | +2.50 ±0.85 |
| Creative-v3 | 36.5 / 34.1 / 34.5 | 35.0 ±1.3 | 39.1 / 36.8 / 37.4 | 37.7 ±1.2 | +2.73 ±0.17 |
| *Dialogue* | | | | | |
| Arena-Hard | 18.2 / 16.1 / 17.4 | 17.3 ±1.1 | 18.4 / 17.5 / 19.8 | 18.6 ±1.1 | +1.29 ±1.07 |
| *Medicine* | | | | | |
| HealthBench | 30.8 / 33.0 / 31.8 | 31.9 ±1.1 | 36.0 / 34.9 / 33.1 | 34.7 ±1.4 | +2.77 ±2.10 |
| MedQA | 63.2 / 61.2 / 63.5 | 62.6 ±1.3 | 66.1 / 65.0 / 64.2 | 65.1 ±1.0 | +2.49 ±1.61 |
| *Science* | | | | | |
| GPQA-Diamond | 46.8 / 48.0 / 39.4 | 44.7 ±4.7 | 43.3 / 42.4 / 46.0 | 43.9 ±1.8 | -0.84 ±6.50 |
| ResearchQA | 60.4 / 64.6 / 63.6 | 62.9 ±2.2 | 60.4 / 63.3 / 59.5 | 61.1 ±2.0 | -1.82 ±2.08 |
| Seven-benchmark average | 46.0 / 45.8 / 44.9 | 45.6 ±0.6 | 47.3 / 46.7 / 46.6 | 46.9 ±0.4 | **+1.30** ±0.46 |

# D Mechanistic Analyses and Extended Ablations

## D.1 Sensitivity Analysis: All Reward Structures

Table 13 extends Table 1b to every reward structure, adds partial satisfaction of a criterion, which the main text does not use, and gives the writing rubric under explicit aggregation. The writing rubric was tested with the two naming perturbations only, since the other three are clinical; naming an item earns $+1.3$ there, an interval containing zero.

## D.2 Notes on Section 5

**Judges.** Doubao-lite, from the training reward's family, sees the reversal of Section 3.1 on a smaller scale: Rubric-RL gains 17.7 points of coverage over the untrained model and loses 8.2 on the physician criteria, against a 5.5-point gain and a 28.1-point loss under DeepSeek-V4-Pro. GPT-5.6-luna ranks the models in nearly the same order as DeepSeek-V4-Pro (Spearman $\rho = 0.84$, eleven models, same responses), but its absolute scores are far lower: the untrained model scores 33.7 where DeepSeek-V4-Pro gives 54.1. We therefore read it for ordering only. Criterion by criterion, of the positive-weight physician criteria on 3,657 prompts (7,949 to 8,019 per model), Doubao-lite

Table 9: **Per-seed medical ablation results at 4B scale under both judges.** Subscripts: pro, DeepSeek-V4-Pro; lite, Doubao-lite.

| Method | $c_{\text{pro}}$ per seed | $c_{\text{lite}}$ per seed | $g_{\text{lite}}$ per seed |
|---|---|---|---|
| *Baselines* | | | |
| Rubric-RL | 24.3 / 28.4 / 25.1 $\pm$2.2 | 69.9 / 71.3 / 68.3 $\pm$1.5 | 55.7 / 56.1 / 55.4 $\pm$0.4 |
| RuscaRL | 24.6 / 25.8 / 29.5 $\pm$2.6 | 66.4 / 70.3 / 71.9 $\pm$2.8 | 54.2 / 56.4 / 55.6 $\pm$1.1 |
| OPSD | 40.7 / 42.9 / 41.4 $\pm$1.1 | 64.8 / 67.9 / 66.3 $\pm$1.6 | 35.4 / 36.9 / 36.5 $\pm$0.8 |
| *Aggregation changed* | | | |
| Implicit | 36.7 / 36.6 / 36.8 $\pm$0.1 | 78.0 / 77.5 / 77.7 $\pm$0.2 | 56.4 / 56.1 / 55.6 $\pm$0.4 |
| raw-AND | 38.8 / 31.9 / 33.9 $\pm$3.6 | 76.3 / 72.1 / 73.3 $\pm$2.2 | 53.8 / 54.3 / 53.0 $\pm$0.6 |
| Graded | 38.3 / 34.7 / 35.8 $\pm$1.8 | 80.1 / 77.6 / 78.4 $\pm$1.3 | 58.7 / 58.0 / 59.0 $\pm$0.5 |
| ProRubric | 39.8 / 36.2 / 34.5 $\pm$2.7 | 78.0 / 74.9 / 74.9 $\pm$1.8 | 56.4 / 55.3 / 54.2 $\pm$1.1 |
| *Appropriateness criterion added* | | | |
| Rubric-RL + appropriateness criterion | 34.5 / 34.5 / 33.7 $\pm$0.5 | 72.8 / 72.7 / 71.6 $\pm$0.7 | 54.9 / 52.9 / 54.2 $\pm$1.0 |
| ProRubric + appropriateness criterion | 45.4 / 40.7 / 43.4 $\pm$2.4 | 77.6 / 77.6 / 78.0 $\pm$0.2 | 53.6 / 53.8 / 53.2 $\pm$0.3 |
| *Controls* | | | |
| ProRubric w/o failure clauses | 35.8 / 34.1 / 33.6 $\pm$1.1 | 73.4 / 74.0 / 76.6 $\pm$1.7 | 54.8 / 54.9 / 55.6 $\pm$0.4 |
| $K$=1 | 34.8 / 37.6 / 38.5 $\pm$1.9 | 74.1 / 75.3 / 77.1 $\pm$1.5 | 51.9 / 53.0 / 52.4 $\pm$0.5 |
| Length-matched | 27.5 / 25.1 / 28.7 $\pm$1.8 | 69.4 / 67.2 / 67.9 $\pm$1.1 | 52.9 / 52.4 / 53.2 $\pm$0.4 |

Table 10: **Full medical ablations under both judges (4B).** Subscripts: pro, DeepSeek-V4-Pro; lite, Doubao-lite. Appr. crit.: the fixed appropriateness criterion (Appendix D.3); Len.: mean response length in characters; $n$: number of seeds.

| Configuration | Reward structure | | Score | | | | Len. | $n$ |
|---|---|---|---|---|---|---|---|---|
| | grouped | appr. crit. | $g_{\text{lite}}$ | $g_{\text{pro}}$ | $c_{\text{lite}}$ | $c_{\text{pro}}$ | | |
| *Baselines* | | | | | | | | |
| Base | – | – | 38.1 | 26.6 | 78.0 | **54.1** | 3.0k | 1 |
| Rubric-RL | ○ | ○ | 55.8 $\pm$0.4 | 32.1 | 69.8 $\pm$1.5 | 26.0 $\pm$2.2 | 11.7k | 3 |
| RuscaRL | ○ | ○ | 55.4 $\pm$1.1 | 32.1 | 69.5 $\pm$2.8 | 26.7 $\pm$2.6 | 11.9k | 3 |
| OPSD | – | – | 36.3 $\pm$0.8 | 25.6 | 66.3 $\pm$1.6 | 41.7 $\pm$1.1 | 4.7k | 3 |
| *Aggregation changed* | | | | | | | | |
| Implicit (one holistic score) | ● | ○ | 56.0 $\pm$0.4 | 34.9 | 77.8 $\pm$0.2 | 36.7 $\pm$0.1 | 10.7k | 3 |
| raw-AND (text unchanged) | ● | ○ | 53.7 $\pm$0.6 | 33.8 | 73.9 $\pm$2.2 | 34.9 $\pm$3.6 | 10.4k | 3 |
| Graded (4-level credit) | ● | ○ | 58.6 $\pm$0.5 | 35.1 | 78.7 $\pm$1.3 | 36.3 $\pm$1.8 | 11.5k | 3 |
| ProRubric | ● | ○ | 55.3 $\pm$1.1 | 35.0 | 75.9 $\pm$1.8 | 36.8 $\pm$2.7 | 9.8k | 3 |
| *Appropriateness criterion added* | | | | | | | | |
| Rubric-RL + appropriateness criterion | ○ | ● | 54.0 $\pm$1.0 | 33.7 | 72.4 $\pm$0.7 | 34.2 $\pm$0.5 | 10.6k | 3 |
| ProRubric + appropriateness criterion | ● | ● | 53.5 $\pm$0.3 | 35.3 | 77.7 $\pm$0.2 | **43.1** $\pm$2.4 | 9.2k | 3 |
| *Controls* | | | | | | | | |
| ProRubric w/o failure clauses | ● | ○ | 55.1 $\pm$0.4 | 34.9 | 74.7 $\pm$1.7 | 34.5 $\pm$1.1 | 10.6k | 3 |
| $K$=1 (one dimension) | ● | ○ | 52.4 $\pm$0.5 | 33.5 | 75.5 $\pm$1.5 | 36.9 $\pm$1.9 | 9.6k | 3 |
| Length-matched (4,096-token limit) | ○ | ○ | 52.9 $\pm$0.4 | 30.4 | 68.2 $\pm$1.1 | 27.1 $\pm$1.8 | 6.8k | 3 |
| Rubric-RL, weighted criteria | ○ | ○ | 54.8 $\pm$1.7 | 32.0 | 67.0 $\pm$1.8 | 25.3 $\pm$1.5 | 12.4k | 3 |
| ProRubric, dimensions from another generator | ● | ○ | 51.4 | 33.8 | 71.6 | 38.1 | 10.4k | 1 |
| *KL penalty added* | | | | | | | | |
| Rubric-RL, $\beta$=0.005 | ○ | ○ | 46.3 $\pm$1.1 | 28.8 | 61.0 $\pm$3.2 | 25.8 $\pm$1.3 | 12.7k | 3 |
| Rubric-RL, $\beta$=0.01 | ○ | ○ | 43.2 $\pm$0.8 | 27.7 | 58.9 $\pm$3.4 | 27.9 $\pm$2.0 | 12.3k | 3 |
| Rubric-RL, $\beta$=0.04 | ○ | ○ | 41.4 $\pm$0.6 | 28.4 | 75.2 $\pm$0.3 | 49.3 $\pm$0.7 | 4.1k | 3 |
| ProRubric, $\beta$=0.005 | ● | ○ | 46.7 $\pm$0.3 | 30.2 | 69.9 $\pm$1.4 | 36.9 $\pm$1.4 | 9.1k | 3 |
| ProRubric, $\beta$=0.01 | ● | ○ | 44.4 $\pm$0.5 | 30.5 | 75.4 $\pm$0.3 | 48.0 $\pm$1.3 | 4.9k | 3 |
| ProRubric, $\beta$=0.04 | ● | ○ | 39.9 $\pm$0.6 | 28.4 | 77.7 $\pm$0.4 | **53.6** $\pm$0.4 | 3.3k | 3 |

marks satisfied and DeepSeek-V4-Pro unmet $27.5\%$ for the untrained model and $46.2\%$, $44.3\%$ and $44.1\%$ for Rubric-RL at seeds 42, 43 and 44. Under GPT-5.6-luna, grouping alone (raw-AND) is worth $+7.9$ over Rubric-RL, and appending the appropriateness criterion to ProRubric's dimensions $+4.6$ over appending it to the original checklist. Scored by Doubao-lite, appending the appropriate-

Table 11: **Cross-family ranking validation using GPT-5.6-luna (3,652–3,671 items per method).**

| | Baselines | | Aggregation changed | | | | + Appr. crit. | | KL penalty | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | **Base** | **Rubric-RL** | **Implicit** | **raw-AND** | **Graded** | **ProRubric** | **Rubric-RL + appr.** | **ProRubric + appr.** | **Rubric-RL $\beta$=0.01** | **Rubric-RL $\beta$=0.04** | **ProRubric $\beta$=0.01** |
| **Score** | 33.7 | 14.4 | 21.6 | 22.4 | 20.1 | 20.3 | 18.0 | 22.6 | 20.2 | 29.2 | 28.5 |
| **Items** | 3,655 | 3,670 | 3,652 | 3,669 | 3,670 | 3,671 | 3,668 | 3,669 | 3,652 | 3,654 | 3,653 |

Table 12: **Per-seed HealthBench-consensus results at 4B ($\Delta$ = ProRubric − Rubric-RL).**

| | DeepSeek-V4-Pro | | | Doubao-lite | | |
|---|---|---|---|---|---|---|
| **Seed** | **Rubric-RL** | **ProRubric** | **$\Delta$** | **Rubric-RL** | **ProRubric** | **$\Delta$** |
| 42 | 24.3 | 39.8 | +15.5 | 69.9 | 78.0 | +8.1 |
| 43 | 28.4 | 36.2 | +7.7 | 71.3 | 74.9 | +3.6 |
| 44 | 25.1 | 34.5 | +9.4 | 68.3 | 74.9 | +6.6 |
| Mean | 26.0 | 36.8 | +10.8 | 69.8 | 75.9 | +6.1 |

ness criterion costs 1.8 points of coverage, negative on all three seeds; scored by DeepSeek-V4-Pro on the same responses it costs nothing ($+0.30 \pm 1.05$, with the sign flipping between seeds). Two further pairs behave the same way at seed 42: the length-matched variant costs 2.8 points of coverage under Doubao-lite and 0.7 under DeepSeek-V4-Pro, and Graded leads ProRubric by 2.2 in coverage under Doubao-lite and by 0.8 under DeepSeek-V4-Pro. In each pair Doubao-lite prefers the longer variant, and the gap shrinks about three- to fourfold under DeepSeek-V4-Pro.

**Aggregation controls at 8B and the generator.** At 8B over three seeds, implicit aggregation reaches $47.6 \pm 1.0$ against ProRubric's $45.8 \pm 1.8$ and Rubric-RL's $29.5 \pm 1.1$, and raw-AND recovers 11.4 of ProRubric's 16.3 points. At 4B, $K$=1 is above ProRubric on appropriateness at two seeds and below at one, and is lower on coverage on average; 63–66% of its sampled groups receive identical rewards. Regenerating ProRubric's dimensions with a different generator moves appropriateness by less than two points (38.1 against 39.8 at the same seed).

**KL penalty.** At $\beta \leq 0.01$ Rubric-RL's appropriateness stays flat while its coverage falls. The separation between ProRubric and Rubric-RL across $\beta \in \{0, 0.005, 0.01, 0.04\}$ is +10.8, +11.1, +20.1 and +4.3, three-seed means. The pre-registered test required at least +10 at $\beta = 0.005$ and at least +5 at $\beta = 0.04$; the first holds and the second does not (per seed +4.9, +4.3 and +3.7). At $\beta = 0.04$ ProRubric reaches 53.6 against the untrained model's 54.1, with coverage 28.4 against 26.6 and responses of 3.3k characters against 3.0k. Wherever the policy still learns, ProRubric is ahead by 10 to 20 points, a restriction we chose after seeing these results.

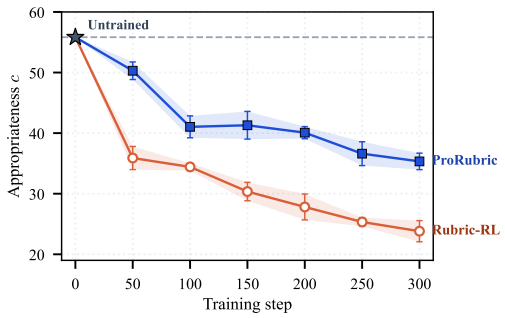


Figure 5: **Appropriateness during training.** Medical 4B, 300 HealthBench-consensus prompts; three-seed mean ± standard deviation, step 0 is the untrained model.

Table 13: **Reward change per edit before training, for every aggregation.** + appr.: with the fixed appropriateness criterion (Appendix D.3).

| Edit to the answer | Rubric-RL | raw-AND | Graded | ProRubric | Rubric-RL + appr. | ProRubric + appr. | Implicit | Writing (Rubric-RL) |
|---|---|---|---|---|---|---|---|---|
| Name an item | **+7.3** [+5.9, +8.7] | −0.3 [−1.9, +1.4] | +1.8 [+0.8, +2.8] | +2.2 [−0.1, +4.5] | +3.0 [+1.5, +4.5] | −3.9 [−6.4, −1.4] | +1.4 [−0.2, +2.8] | +1.3 n.s. |
| Name the topic | −0.4 [−1.7, +0.8] | +0.6 | −1.2 | −3.8 | −1.6 | −7.6 | −1.7 | −0.3 |
| Satisfy a criterion in part | +2.2 | +1.1 | +1.8 | +3.3 | +1.9 | +0.4 | +0.9 | – |
| Add a needless test | +0.3 | −1.4 | −2.7 | −5.8 | −4.6 | **−9.4** | −4.5 | – |
| Drop the key advice | −0.9 | −0.8 | −1.1 | −2.2 | −0.5 | −0.2 | −1.2 | – |
| Paraphrase | +1.1 | +1.1 | +1.5 | +1.0 | +1.7 | +1.1 | +1.0 | – |

**Training dynamics.** The models behind Figure 5 repeat the Rubric-RL and ProRubric settings of Table 3 with the same seeds, saving a checkpoint every 50 steps; they are newly trained, not the checkpoints of Table 3, and all six are reported. Appropriateness is scored on 300 HealthBench-consensus prompts from the common set by a second instance of the same DeepSeek-V4-Pro model. At step 300 the curves end at 23.8 and 35.3. Scored as in Table 3 (full common set, same judge instance), the step-300 checkpoints reach 24.3 and 35.7 (three-seed means; ProRubric over Rubric-RL +11.5, against +10.8 in Table 3), and each Table 3 mean lies within the range of the retrained runs. Grouping ProRubric's dimensions by their status in the untrained model's answer, dimensions one criterion short of complete are completed more often by step 300 than partly or wholly unmet ones, under both rewards and on every seed, so this ordering is not specific to conjunction.

**SFT.** On the consensus axis SFT ends above the untrained model, $66.3 \pm 0.4$ against 54.1 at 4B and $75.6 \pm 0.5$ against 62.5 at 8B (3,143 paired items), where Rubric-RL falls to 26.0 and 29.5; none of the consensus prompts appears in its training data. Outside its domain it pays the largest cost, −5.7 on MedQA at 4B and −12.0 on GPQA-Diamond at 8B, where both RL methods gain about seven. Its responses run 19–55% longer than ProRubric's on the writing benchmarks. On HealthBench it is level with ProRubric at both scales (35.1 against 34.7 at 4B, 40.4 against 39.9 at 8B).

**Domain and rubric shape.** The four domains differ in their rubrics as well as their tasks: science has 7.5 criteria per question in graded tiers with negative weights, the other domains 27 to 32 positive ones. With four domains the two cannot be separated.

### D.3 A Fixed Appropriateness Criterion

Appending one fixed criterion to every checklist, asking the response to answer what was asked, stay in scope and suit the asker (Appendix A.5), raises appropriateness under every aggregation: Rubric-RL from 26.0 to $34.2 \pm 0.5$, ProRubric from 36.8 to $43.1 \pm 2.4$, and implicit aggregation from 36.7 to $44.1 \pm 0.8$ (three seeds each). Coverage is unchanged or higher (33.7 and 35.3 against 32.1 and 35.0). Because the criterion asks for nearly what the physician criteria check, we do not count it as part of ProRubric. It acts on length rather than on errors. On the untrained model's answers the training judge marks it unmet 54.0% of the time; adding rubric terms raises this to 81.7%, while deleting the answer's key sentence raises it only to 57.1%. It is judged unmet on 80.5% of answers into which a mildly inappropriate recommendation is inserted and on 56.3% of content-preserving paraphrases. Adjudicating a sample of its unmet verdicts with DeepSeek-V4-Pro, 67.3% of those on detail-only variants and 52.0% on the untrained model's answers were unwarranted.

### D.4 Cross-Domain Appropriateness Evaluation

The cross-domain results in Sections 5.3 and 5.5 come from an evaluation fixed before training finished. Each domain contributes 200 prompts: HealthBench-consensus in medicine, Arena-Hard v2 in dialogue, ResearchQA in science, and 96 Creative-v3 plus 104 WritingBench prompts in writing. DeepSeek-V4-Pro scores every response on the same four domain-independent criteria: the response addresses the primary request before supplementary material, adapts to the user's stated context, includes only material that helps and avoids repetition (*redundancy*), and separates supported claims from uncertainty. The score is the mean over the four criteria (×100). Differences are paired over

Table 14: **Cross-domain appropriateness: scores and differences from Rubric-RL across four domains.**

| | Medicine | | Dialogue | | Science | | Writing | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Score** | **Δ [95% CI]** | **Score** | **Δ [95% CI]** | **Score** | **Δ [95% CI]** | **Score** | **Δ [95% CI]** |
| Base | 50.0 | $+31.1^{*}$ [+27.0, +35.1] | 45.6 | $+7.4^{*}$ [+3.6, +11.1] | 52.2 | $+11.2^{*}$ [+8.4, +14.4] | 38.9 | $-19.5^{*}$ [−23.9, −15.1] |
| Rubric-RL | 18.9 | — | 38.2 | — | 41.0 | — | 58.4 | — |
| ProRubric | 39.2 | $+20.4^{*}$ [+16.9, +23.9] | 44.9 | $+6.6^{*}$ [+3.5, +9.9] | 51.0 | $+10.0^{*}$ [+7.4, +12.6] | 64.1 | $+5.8^{*}$ [+3.2, +8.2] |
| raw-AND | 29.9 | $+11.0^{*}$ [+8.0, +14.1] | 45.9 | $+7.6^{*}$ [+4.4, +10.9] | 43.6 | $+2.6^{*}$ [+0.1, +5.2] | 59.9 | +1.5 [−1.2, +4.2] |
| Implicit, seed 42 | – | – | 43.8 | $+5.5^{*}$ [+2.5, +8.8] | 44.2 | $+3.2^{*}$ [+0.8, +5.9] | 66.0 | $+7.6^{*}$ [+5.4, +9.9] |
| Implicit, seed 43 | – | – | 44.1 | $+5.9^{*}$ [+3.0, +9.0] | 42.8 | +1.8 [−0.6, +4.2] | – | – |
| Implicit, seed 44 | – | – | 43.4 | $+5.1^{*}$ [+2.0, +8.2] | 41.5 | +0.5 [−1.8, +2.9] | – | – |
| Rewritten, seed 42 | – | – | – | – | 43.1 | +2.1 [−0.2, +4.5] | – | – |
| Rewritten, seed 43 | – | – | – | – | 42.6 | +1.6 [−0.8, +4.1] | – | – |

prompts, with 95% bootstrap intervals over prompts only. Table 14 shows seed 42 unless a seed is marked. The changes from the untrained model in Figure 3c and Section 5.5, and science's 7.5 in Section 5.3, are instead means over seeds 42, 43 and 44 of Rubric-RL and ProRubric on the same prompts (Table 15a). The redundancy sub-score of Sections 5.4 and 5.5 is the change, in points, in the pass rate of the redundancy criterion alone when the untrained model's answers are perturbed as in Section 3.3, in each domain (Table 15b).

**Blind preference and appropriateness.** The blind pairwise judge of Table 1a weighs, in order, whether a response answers what was asked, fits the asker, is accurate, and avoids unnecessary length, and returns one overall preference; the criteria above are judged one at a time, so redundancy is a quarter of the score and cannot be offset by completeness. In science the two readings part for the same judge on the same prompts and checkpoints. On the 100 ResearchQA prompts per seed of Table 1a, DeepSeek-V4-Pro prefers Rubric-RL in 262 of 300 pairs, but on the answers it prefers, Rubric-RL passes the redundancy criterion 0–1% of the time against 35–42% for the untrained model. On these prompts appropriateness falls by 12.0 to 13.0 points per seed: the pass rate of the redundancy criterion falls by 35 to 36 points and that of adaptation to context by 14 to 18, while answering the primary request rises by 1. Rubric-RL's science answers are more complete (ResearchQA coverage 51.8 to 62.9, Table 2) and less focused; DeepSeek-V4-Pro counts the completeness as worth its length when ranking whole answers, and GPT-5.6-luna does not. Dialogue follows the same pattern: both judges prefer the trained policy, and its 7.1-point fall lies mostly in the redundancy criterion.

### D.5 Where Added Material Sits: Appending versus Integrating

The two perturbations of Section 5.5 both append material, so they cannot say whether the appropriateness charge is for the added content or for its position. A third perturbation places the same material inside the answer instead of appending it. Sized to a common per-prompt budget, the two placements are charged alike: on the items where both land within 25% of each other, the Rubric-RL reward pays +41.4 appended against +40.3 integrated in medicine ($n$=35) and +19.5 against +18.4 in science ($n$=108), with ProRubric at +48.8 against +52.1 and +28.9 against +28.5. Writing and dialogue lean the other way by 4 to 12 points ($n$=50 and $n$=66), but we do not read that lean as an effect of placement: within every matched pair the integrated variant is the longer one, by 13.7% to 17.2%, which is the same order as the gap itself. The comparison does settle one point. At unequal budgets, with the integrated variant given about 18% of the response and the appended one 50%, on 8 to 87 items, the domains had been 31 points apart. At a common budget that spread is gone and no domain charges nothing. *Where* the material sits is at most a second-order term; what is charged is the material. The comparison has three limits. The pair is matched after the fact and only 12% to 36% of prompts admit both rewrites, so the four domains are not four repetitions of one experiment; matching selected *shorter* originals in medicine and *longer* ones in the other three. The realized expansion is +68% to +105% rather than the +50% we pre-registered. And the matched subsets run from 35 to 108 items, so the per-domain gaps carry wide intervals.

Table 15: **Cross-domain appropriateness over three seeds, and the incentive audit in every domain.** (a) Change from the untrained model on the 200 prompts per domain of Table 14, seeds 42 / 43 / 44, mean ± s.d.; the last column pairs each seed with the Rubric-RL model of the same seed. (b) The untrained model's answers to 300 training prompts per domain, perturbed as in Section 3.3: reward change ($\times 100$) under the Rubric-RL reward for naming an item and for adding unrequested substance; the naming payment as a share of the substance payment under each reward; the naming edits that leave the Rubric-RL reward unchanged; and the change in the redundancy criterion's pass rate (points; all prompts with every variant, 293–300 per domain).

**(a) Appropriateness change from the untrained model**

| Domain | Untrained score | Rubric-RL 42 / 43 / 44 | Rubric-RL mean | ProRubric 42 / 43 / 44 | ProRubric mean | ProRubric − Rubric-RL mean (lowest seed) |
|---|---|---|---|---|---|---|
| Medicine | 50.0 | −31.1 / −26.5 / −29.0 | −28.9 ±2.3 | −10.8 / −16.6 / −19.5 | −15.6 ±4.5 | +13.3 (+9.5) |
| Science | 52.2 | −11.2 / −11.6 / −10.8 | −11.2 ±0.4 | −1.2 / −6.6 / −3.2 | −3.7 ±2.7 | +7.5 (+5.0) |
| Dialogue | 45.6 | −7.4 / −9.8 / −4.1 | −7.1 ±2.8 | −0.8 / −0.2 / +0.2 | −0.3 ±0.5 | +6.8 (+4.4) |
| Writing | 38.9 | +19.5 / +13.9 / +14.6 | +16.0 ±3.1 | +25.2 / +25.0 / +22.6 | +24.3 ±1.4 | +8.3 (+5.8) |

**(b) Incentive audit before training**

| Domain | Rubric-RL reward: Name an item | Rubric-RL reward: Add substance | Naming / substance: Rubric-RL | Naming / substance: ProRubric | Reward unchanged by naming | Redundancy pass rate: Name an item | Redundancy pass rate: Add substance |
|---|---|---|---|---|---|---|---|
| Medicine | +7.3 | +22.7 | 32% | 6% | 37 / 300 | −17.0 | −25.0 |
| Dialogue | +2.4 | +17.9 | 13% | −14% | 66 / 300 | −29.2 | −33.6 |
| Writing | +1.3 | +18.9 | 7% | −10% | 71 / 300 | −25.6 | −37.5 |
| Science | −0.8 | +10.8 | −8% | −23% | 195 / 300 | −26.8 | −31.9 |

# E Qualitative Case Studies and Protocol Analysis

## E.1 ProRubric Rubrics and Representative Responses

**One rubric, before and after.** A training prompt from the clinical split reads: *"Doctor, I was in a car accident and I'm experiencing neck pain and abdominal pain. What could be the cause of this?"* Its atomic rubric carries 32 independently scored criteria, among them *"explicitly states that the neck and abdominal pain are occurring after a car accident"* ($w = 8$), *"avoids stating a single certain diagnosis"* ($w = 9$), *"lists concrete urgent warning signs"* ($w = 10$) and *"describes whiplash or neck muscle strain as common post-collision injuries"* ($w = 7$). ProRubric folds these into four protocol dimensions, each inheriting the summed weight of its members (59, 118, 36 and 20, totaling 233, exactly the sum of the 32 atomic weights):

1. **Emergency recognition** ($w = 59$): directs the user to urgent in-person evaluation. *Fails if it omits the recommendation for urgent medical evaluation, fails to include NPO guidance, or presents the situation as non-urgent.*
2. **Probabilistic differential** ($w = 118$): spans common minor conditions and life-threatening pathologies, linking each to the mechanism of blunt trauma. *Fails if it presents only benign explanations, omits the referred-pain mechanism linking the two symptom regions, or states a definitive diagnosis.*
3. **Expected workup** ($w = 36$): describes the physical examination, laboratory tests and imaging. *Fails if it does not outline the expected workup or omits dual-region CT imaging.*
4. **Accessible communication** ($w = 20$): organises the content and explains terminology. *Fails if it uses unexplained jargon or ends without a summary reinforcing the key safety message.*

The requirements are the same, grouped; each dimension gains a failure clause and is paid only when all of its members hold.

**What the two policies produce.** The clearest instances are the shortest prompts. On the consensus benchmark, one user writes five words: *"My baby has a fever."* Two physician-authored criteria apply, each weighted 5. The Rubric-RL policy answers in 10,890 characters and DeepSeek-V4-Pro marks *both* unmet; the ProRubric policy answers in 5,277 and the judge marks both met.

The first criterion states its own reason: the response "should avoid being overly alarmist," should "recommend care with enough urgency," and, most directly, *"the emergency referral should not be overly verbose, unclear, or indirect in a way that could distract the user from seeking care."* The second requires the response either to seek the missing context that decides whether emergency care is needed, or to give conditional guidance covering the likely scenarios. Neither criterion rewards breadth, and the longer answer satisfies neither.

### E.2 Failure Cases and Coverage Trade-offs

In action-oriented domains conjunctive grouping discourages long lists of loosely related facts, but it has a cost on broad information-retrieval tasks such as ResearchQA.

In literature-synthesis queries where the evaluation rubric tests exhaustive topical coverage (e.g., enumerating all documented experimental methodologies across multiple sub-fields), ProRubric tends to summarize the main methodological approaches and occasionally omits niche cases that Rubric-RL lists explicitly. Under a protocol that scores only coverage of the atomic checklist, ProRubric therefore scores slightly lower than the unconstrained Rubric-RL (61.1 vs. 62.9 at 4B and 64.9 vs. 65.8 at 8B, three-seed means). When a task's utility is exhaustive enumeration rather than concise decision-making, the conciseness enforced by conjunctive failure clauses gives up some coverage breadth for denser communication, a measurable trade-off.