Rethinking Rubric Generation for Improving LLM Judge and Reward Modeling for Open-ended Tasks

William F. Shen[*] [†] **Xinchi Qiu**[*]
Chenxi Whitehouse **Lisa Alazraki** **Shashwat Goel** **Francesco Barbieri**
Timon Willi **Akhil Mathur** **Ilias Leontiadis**
Meta Superintelligence Labs

Abstract

Recently, rubrics have been used to guide LLM judges in capturing nuanced, multi-dimensional human preferences, and have further been extended as reward signals for reinforcement fine-tuning (RFT). However, open-ended generation typically lacks a unique canonical target, and its evaluation rubrics are latent rather than directly observed. This makes rubric generation under-determined and difficult to control: rubrics often lack coverage, conflate dimensions, misalign preference direction, and contain highly correlated criteria, degrading judge accuracy and producing suboptimal rewards during RFT. We propose RRD, a practical framework for rubric refinement built on a recursive decompose–filter cycle. RRD decomposes coarse rubrics into fine-grained, discriminative criteria, expanding coverage while sharpening separation between responses. A complementary filtering mechanism removes misaligned and redundant rubrics, and a correlation-aware weighting scheme to prevent over-representing highly correlated criteria, yielding rubric sets that are informative, comprehensive, and non-redundant. Empirically, RRD delivers large, consistent gains across all evaluation and training: it improves preference-judgment accuracy on JudgeBench and PPE for both GPT-4o and Llama3.1-405B judges, achieving top performance in all settings with up to **+17.7** points on JudgeBench. When used as the reward source for RFT on WildChat, it yields substantially stronger and more stable learning signals, boosting reward by up to **160%** (Qwen3-4B) and **60%** (Llama3.1-8B) versus $\sim$10–20% for prior rubric baselines, with gains that transfer to HealthBench-Hard and BiGGen Bench. Overall, RRD establish recursive rubric refinement as a scalable and interpretable foundation for LLM judging and reward modeling in open-ended domains.

1 Introduction

Large language models (LLMs) are increasingly used as judges (LLM judge) to evaluate open-ended generations (e.g., creative writing and planning) [15]. However, because quality in these settings is subjective and inherently multi-attribute, LLM-based judging remains brittle: even frontier models can be near chance on preference benchmarks, and real-world use is further limited by bias, inconsistency, and limited transparency, especially when evaluation criteria are implicit or underspecified [50, 49, 37, 42, 64, 18, 24, 14].

In parallel, recent work extends Reinforcement Learning from Verifiable Rewards (RLVR) beyond domains with objectively checkable outcomes (e.g., math and coding) to open-ended tasks under non-verifiable rewards [8, 17, 26]. A common approach converts binary correctness into pairwise preferences [22], but this approach not only inherits the limitations of standard LLM judge labeling,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

[*]Equal contribution.
[†]Correspondence to: fs604@cam.ac.uk

but also introduces a key scaling bottleneck: generating preference supervision that incorporates diverse and robust criteria matching the complexity of real-world reasoning [36, 5, 47, 52, 60].

One promising direction is rubric-based judging, where LLMs generate explicit, structured rubrics and use them to ground assessments, improving reliability and interpretability relative to holistic judgments [19]. These rubric-level signals also integrate naturally into reinforcement fine-tuning (RFT) as structured rewards, motivating the Rubrics-as-Rewards paradigm for aligning models on complex open-ended tasks [16].

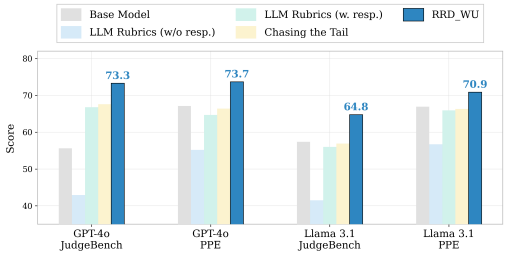


Figure 1: RRD consistently outperforms all baselines on both JudgeBench and PPE for both proprietary (GPT-4o) and open-weights (Llama3.1-405B) judges, delivering substantial gains in preference-judgment accuracy.

However, existing rubric generation methods face two key limitations: (1) *coverage deficiency*, where the rubric set fails to comprehensively capture the diverse and nuanced dimensions of generation quality; (2) *noisy evaluation outcomes*, where misaligned, duplicative, or highly correlated rubrics introduce unreliable signals. Consequently, rubric-based judges often show weak alignment with human preferences, reducing evaluation accuracy, and when used as rewards for RFT, yield limited improvements in learning preference-aligned behavior across diverse model outputs. These flaws are severe: we show that naively generated rubrics degrade GPT-4o's judgment accuracy from 55.6% to 42.9% (Figure 1 on JudgeBench [49] – 13 points below using no rubrics at all).

In this work, we propose Recursive Rubric Decomposition (RRD), a principled framework that improves both the accuracy of rubric-based judges and their effectiveness as reward models for downstream RFT. RRD expands coverage by decomposing high-level rubric items (i.e., broad criteria satisfied by many responses) into finer-grained, nuanced dimensions. This yields more comprehensive and discriminative evaluations, helping LLM judges better capture subtle yet consequential quality differences when comparing candidate responses. To complement rubric expansion, we introduce an aggregation mechanism to mitigate noise by (i) filtering rubrics that produce misaligned signals, and (ii) removing duplicative rubrics while down-weighting highly correlated ones. The first step guards against criteria that yield grossly incorrect judgments, while the second prevents overlapping perspectives from being overrepresented. Together, these components produce more stable and informative rubric-based assessments.

This process mirrors many real-world evaluations which follow a structured assessment rather than a single holistic verdict. Consider a physician diagnosing ambiguous symptoms: they don't render a holistic verdict, but forms hypotheses and orders discriminating tests. When a test result is consistent with multiple conditions, she orders more specific tests that distinguish between them. For example, a positive result for "inflammatory markers" doesn't determine whether the cause is autoimmune, infectious, or malignant - she must decompose further. The process terminates when tests discriminate: the remaining diagnosis is the only one consistent with all evidence. Crucially, she also recognizes correlated indicators – elevated CRP and ESR both signal inflammation but shouldn't be double-counted as independent evidence. RRD instantiates this diagnostic logic: rubrics satisfied by multiple responses are insufficiently discriminative and are recursively decomposed; the process adapts naturally to case complexity, and correlated criteria are appropriately down-weighted.

We empirically demonstrate the performance of RRD two-fold. We first evaluate RRD on two widely used preference-judgment benchmarks: JudgeBench [49] and Preference Proxy Evaluation (PPE) Preference [13]. Across both GPT-4o and Llama3.1-405B judges, RRD consistently yields stronger agreement with human pairwise preferences, improving accuracy over the base judges and prior rubric-based baselines. In particular, our best variant (RRD $_{\text{WU}}$) achieves the top score on all settings achieving up to **17.7%** improvement on JudgeBench for GPT-4o (Figure 1 and Figure 3).

In addition, we demonstrate the effectiveness of RRD in RFT by training open-source policies, Qwen3-4B [59] and Llama3.1-8B [10], with rubric-based judges on WildChat [63] as the reward signal. Compared to prior LLM-judge-based and rubrics-based reward baselines, RRD yields substantially stronger and better-calibrated rewards, translating into faster learning and higher final reward (Figure 4). Concretely, our method boosts reward by up to **160%** for Qwen3-4B and **60%**

for Llama3.1-8B, versus only ∼10–20% for LLM Rubrics and Chasing the Tail (Figure 4). These gains also carry over to downstream evaluations, where the resulting policies consistently improve on BiGGen Bench [25] and HealthBench-Hard [2] for both model families.

Overall, our results show that recursive rubric decomposition (RRD) is a key enabler of effective rubric-based judging and reward modeling for open-ended tasks, turning brittle rubrics into reliable signals that support improved alignment in open-ended language generation.

2 RRD Framework

We first review rubric-based judges in §2.1. We then motivate RRD by identifying three desiderata for useful rubric sets: *informativeness*, *comprehensiveness*, and *non-redundancy* in §2.2. However, open-ended generation typically lacks a unique canonical target, and its evaluation rubrics are latent: individual rubric edges $\boldsymbol{\mu}$ are not directly identifiable from unlabeled data. In §2.3, we present RRD, a practical framework that addresses this challenge and improves rubric generation, leading to stronger empirical performance for both LLM judges and downstream RFT.

2.1 Rubric-based Judge Overview

Let $\mathcal{P}$ denote the prompt space and $\mathcal{R}$ the response space. For each $P \in \mathcal{P}$, consider a finite candidate set $R(P) = \{R_1, \ldots, R_M\} \subseteq \mathcal{R}$. An LLM judge outputs a preference verdict $\mathcal{V}$ over $R(P)$.

A *rubric-based* judge conditions its evaluation on rubric predicates. Each rubric, assessed separately, is a measurable map $g : \mathcal{P} \times \mathcal{R} \to \{0, 1\}$, where $g(P, R) = 1$ indicates that response R satisfies the criterion under prompt P. Given a rubric family $\mathcal{G} = (g_1, \ldots, g_m)$ and nonnegative weights $\boldsymbol{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m_+$, we define the rubric reward

$$f_{\boldsymbol{w},\mathcal{G}}(P, R) \;:=\; \sum_{k=1}^{m} w_k \, g_k(P, R)\,. \tag{1}$$

The constraint $w_k \geq 0$ is without loss of generality: any negatively polarized rubric can be rewritten as a positive, contrastive criterion. For two responses $R_i, R_j \in R(P)$, the judge prefers R_i over R_j, denoted $R_i \succ R_j$, iff $f_{\boldsymbol{w},\mathcal{G}}(P, R_i) > f_{\boldsymbol{w},\mathcal{G}}(P, R_j)$.

2.2 Motivation: Desiderata for Useful Rubrics

Rubrics directly govern how LLM judges compare responses and, consequently, the reward signals used in RFT. Their quality is therefore a primary determinant of both judgment accuracy and reward fidelity. We motivate RRD through three desiderata for effective rubric sets: *informativeness*, *comprehensiveness*, and *non-redundancy*. We support this motivation by analyzing the probability that aggregated rubric-level verdicts misclassify the true preference label and derive an exponential upper bound on this error probability. Minimizing this bound gives rise to a principled objective for rubric generation and correlation-aware weighting, directly operationalizing the three desiderata.

Assume each rubric $g_k \in \mathcal{G}$ is weakly informative, i.e., a positive "edge" over random guessing (A1), and its noise, the deviation of its verdicts from conditional mean given the true label, is sub-Gaussian with bounded dependence (A2):

- (A1) There exist $\mu_k > 0$ such that $\mathbb{E}[\widehat{Y}_k \mid Y = +1] = +\mu_k$, $\mathbb{E}[\widehat{Y}_k \mid Y = -1] = -\mu_k$, where $Y \in \{\pm 1\}$ is the ground-truth preference label, $\widehat{Y}_k$ is the verdict of the kth rubric.
- (A2) Letting $Z_k = \widehat{Y}_k - \mu_k Y$, the vector $Z = (Z_1, \ldots, Z_m)$ is mean-zero sub-Gaussian with covariance Σ_y satisfying $\mathrm{Var}(Z_k) \leq \sigma_k^2$ and $|\mathrm{Corr}(Z_i, Z_j)| \leq \rho < 1$ for $i \neq j$.

In simple words, (A1) assumes that each rubric, with a positive edge ($\mu_k > 0$), contributes positively toward distinguishing the preferred response from the inferior one; (A2) assumes that individual rubric's noise is bounded and rubrics are not repetitive (bounded pairwise correlation $\rho < 1$). Here, the noise $Z_k = \widehat{Y}_k - \mu_k Y$ captures rubric-specific randomness or systematic mismatch – i.e., the part of a rubric's verdict not explained by the true label, arising from ambiguous criteria, imperfect judge execution, or instance-specific idiosyncrasies. This enables standard concentration bounds over the aggregated, non-redundant decisions.

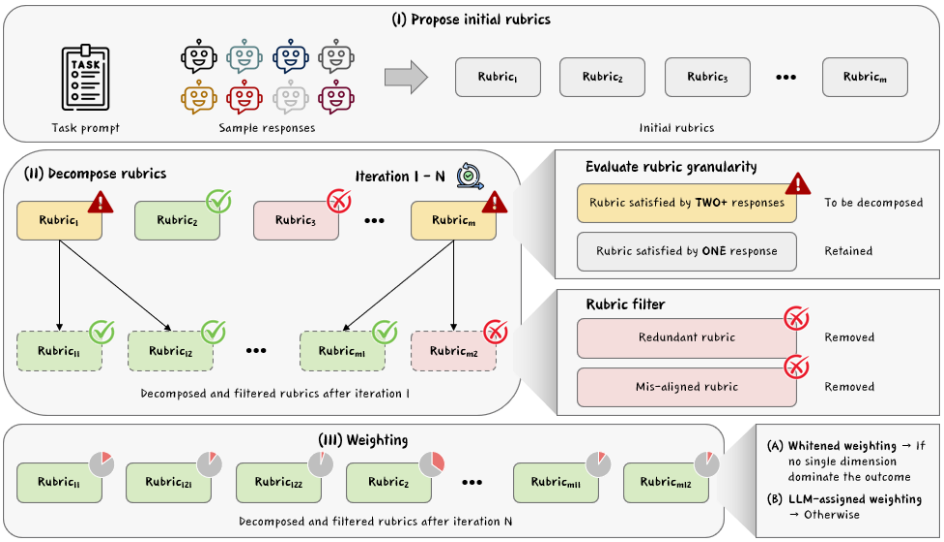


Figure 2: **Overview of RRD framework**: RRD consists of three stages: **(I) Initial Rubric Proposal.** LLM proposes initial candidate rubrics (conditioned on the task prompt and sample responses) for optimization. **(II) Recursive Decomposition and Filtering.** Recursively decompose coarse rubric into finer dimensions to enhance coverage and discrimination, while performing sanity checks for gross misalignment and duplication. The cycle terminates when discarded rubrics exceeds N, indicating saturation in novel, non-redundant, and valid rubrics. **(III) Rubric Weight Assignment.** For open-ended tasks where preference signal is distributed, assign whitened uniform (WU) weights to account for correlation structure and prevent over-representation of highly correlated rubrics. Empirically, WU weighting yields higher LLM judge accuracy and improves the effectiveness of rubrics as generative rewards in RFT.

The probability that the rubric-based judge, after aggregating decisions across all rubric items (Eq.1), produces an incorrect verdict (misclassification probability) $\widehat{Y} \neq Y$ is then upper bounded by:

$$\mathbb{P}(\widehat{Y} \neq Y) \;\leq\; \exp\Big(-\tfrac{1}{2} \min\{\Delta_m^2/V_m(+1),\, \Delta_m^2/V_m(-1)\} \Big) \tag{2}$$

where $\Delta_m := \boldsymbol{w}^\top \boldsymbol{\mu}$ and $V_m := \boldsymbol{w}^\top \boldsymbol{\Sigma}\, \boldsymbol{w}$ is the variance proxy of the weighted residuals (App. D.1).

Eq. 2 implies that tightening the misclassification bound amounts to maximizing $\Xi = \frac{(\boldsymbol{w}^\top \boldsymbol{\mu})^2}{\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}}$. This perspective suggests a high-level prescription:

1. **Decomposition**: Decompose broad rubrics into finer dimensions to enhance coverage and improve discriminative power among responses.
2. **Positivity**: Remove misaligned rubrics to prevent negative edge and maintain constructivity.
3. **Non-redundancy**: Remove duplicates to ensure each rubric contributes unique perspectives.
4. **Weighting.**: Down-weight highly correlated rubrics to mitigate over-representation at aggregation.

In tandem, (1) and (2) expand the rubric set to maximize aggregate edge, while (3) and (4) minimize duplication and correlation within the denominator. Together, these yield an exponentially decaying misclassification probability and motivates the three desiderata of effective rubric sets.

2.3 Methodology: Recursive Rubric Decomposition (RRD)

Now, we introduce *Recursive Rubric Decomposition* (RRD), a practical rubric construction framework that operationalizes the motivation in §2.2. The full RRD procedure is summarized in Figure 2 and Algorithm 1.

RRD consists of three stages. First, we let LLM propose initial task-specific rubrics conditioned on the task prompt and m sample responses (we set $m = 8$ in all experiments), yielding candidate rubrics for optimization.

Second, we start the recursive decomposition and filtering cycle. We exploit the fact that distinct responses necessarily differ in some respects: a rubric satisfied by many responses is too broad-brush and insufficiently discriminative. It can be decomposed into finer sub-dimensions that capture more nuanced aspects of quality. We operationalize this by instructing an LLM-based rubric generator to recursively decompose any rubric that applies to more than n rollouts (we use $n = 2$ in all experiments), while pairing it with a checker that verifies decomposed rubrics introduce finer-grained evaluation dimensions, without discarding information contained in the original rubric. Lastly, decomposed rubrics are filtered for potential misalignment or duplication as a sanity check. In the absence of human ground-truth labels, detecting misalignment inevitably requires a preference proxy [11, 7, 35]. In our experiments, we instantiate this by discarding rubrics that favor a weaker model (Llama3-8B) over a stronger model (GPT-4o), consistent with the use of strong-model preferences as supervision signals in prior work on LLM evaluation and preference learning [64, 30, 3, 27]. We show this simple filtering step yields tangible benefits in ablation study in §3.1, while noting that the gain is on a system level and can be further improved with more sophisticated preference-estimation procedures (more discussions in Appendix E.1).

We repeat this decomposition–filtering loop until the proposer struggles to produce novel, valid, non-redundant items. For efficiency, we use an early-stopping rule: if the number of accumulated rejected proposals exceeds a termination threshold, we stop the loop, since further iterations are unlikely to yield use new rubrics. This threshold is treated as a tunable hyperparameter, as is commonly used in iterative refinement procedures across domains [28, 34, 38, 40]. The resulting rubric set is task-adaptive: its size emerges from the intrinsic complexity of the prompt rather than being fixed *a priori*.

Finally, rubric weights are set to prevent over-representation of highly correlated rubrics. In practice, this is particularly challenging because the ground-truth preference labels needed to estimate the rubric edges $\boldsymbol{\mu}$ are unavailable. Prior work therefore instructs an LLM to assign rubric weights using a heuristically chosen scale [16, 51, 61]. This strategy is effective only if there exists a clearly *dominant* rubric and the LLM can reliably identify it, so that $(\boldsymbol{w}^\top \boldsymbol{\mu})^2$ dominates the covariance term, effectively suppressing the impact of rubric correlations. In open-ended tasks, quality is inherently multi-dimensional, subjective, and nuanced, so the class-separating signal is often *distributed* across many criteria rather than concentrated in single dominant rubric. This helps explain why existing rubric-based judges remain imperfect in practice, and why expanding coverage with additional valid rubrics further disperses the signal. To sidestep the need for labeled edges, we instead minimize the misclassification upper bound by *homogenizing signal in a whitened space*. This is feasible because the second-order redundancy structure of rubric scores, captured by $\boldsymbol{\Sigma}$, is an intrinsic property of the rubric set, not of any particular response pair, and, thus, can be estimated from unlabeled data (Appendix D.2). Concretely, we choose weights that *whiten* the rubric space via $\Sigma^{-1/2}$ (i.e., removing correlations while applying equal weighting in the whitened coordinates). This yields a simple, label-free, and correlation-aware weighting scheme that stabilizes aggregation when signals are spread across many rubric dimensions.

3 RRD-based LLM Judge Results

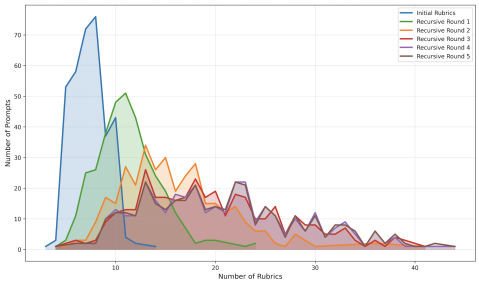
Having introduced RRD as a novel and practical framework for rubric generation in §2, we now evaluate whether it delivers consistent empirical gains across comprehensive experiments. In §3, we first test whether RRD improves LLM judge accuracy on open-form pairwise judgment tasks.

Dataset. We evaluate the accuracy of LLM judge using (1) JudgeBench [49], which consists of challenging open-form response pairs spanning knowledge, reasoning, mathematics, and coding tasks. Following Tan et al. [49], we report results on the subset where both responses are generated by GPT-4o (350 preference pairs), and (2) Preference Proxy Evaluations (PPE) [13], a large-scale benchmark containing 10.2K human preference pairs from Chatbot Arena covering 20 LLMs across 121+ languages.

Baselines. We compare RRD with the base judge (preference labeling without explicit rubrics) and rubric-based baselines spanning prompt-only rubric generation, response-grounded rubric generation, and prior automatic rubric-construction methods, including Chasing the Tail [61], Auto-Rubric [57], and InfiMed-ORBIT [53] (refer to Appendix F.1 for details). We also evaluate three RRD variants:

Method	*GPT-4o*		*Llama3.1-405B*	
	JudgeBench	**PPE**	**JudgeBench**	**PPE**
Base Model	55.6	67.1	57.4	66.9
LLM Rubrics (w/o resp.)	42.9	55.2	41.5	56.7
LLM Rubrics (w. resp.)	66.8	64.7	56.0	65.9
Chasing the Tail	67.6	66.4	56.9	66.3
Auto Rubric	51.1	52.0	54.3	57.0
InfiMed-ORBIT	60.0	61.3	48.5	50.1
$\text{RRD}_{\text{uniform}}$	70.0	71.2	61.1	68.2
RRD_{LLM}	72.4	72.6	63.5	69.7
RRD_{WU}	**73.3**	**73.7**	**64.8**	**70.9**

(a)



(b)

Figure 3: (a) Accuracy on JudgeBench and PPE Preference datasets for base model and different rubric-generation strategies. While basic LLM-generated rubrics (unconditioned on sample responses) can degrade performance, RRD yields consistent improvements over the baselines. Notably, RRD_{WU} delivers the largest gains and scales reliably across both proprietary (GPT-4o) and open-weights (Llama3.1-405B) judges. (b) Rubric-count dynamics on JudgeBench. Starting from an average of 7.4 rubrics, the count rises to ~ 20, while the increasing variance across tasks indicates that the recursive procedure adapts evaluation depth to instance complexity.

$\text{RRD}_{\text{uniform}}$, RRD_{LLM}, and RRD_{UW}, corresponding to uniform, LLM-assigned, and whitened weights, respectively. Unless otherwise stated, we use GPT-4o [21] and Llama-3.1-405B [10] as rubric generators and final judges, and GPT-4o and Gemini 2.5 Pro [6] as sample-response generators.

3.1 Results

We summarize the primary performance results in Figure 3(a) and analyze the dynamics of the recursive generation process in Figure 3(b).

As shown in Figure 3(a), RRD variants consistently outperform baselines across both models and datasets. On JudgeBench, RRD_{WU} improves GPT-4o's accuracy from 55.6% to 73.3% ($+17.7$ points). Similarly, Llama-3.1-405B sees a 6.6 point boost. Standard deviations across three runs for both base models are below 0.46. This suggests that the recursive discovery of latent rubrics provides critical signal that the initial rubric sets miss in a single pass. Interestingly, simple LLM Rubrics (w/o resp.) actually degrade performance compared to the base model as LLM judge. This highlights the failure mode where generic rubrics introduce noise or distract the judge. RRD mitigates this by grounding the rubrics in recursive residuals, ensuring they capture meaningful differences. Additionally, whitened weighting (RRD_{WU}) consistently outperforms the uniform and LLM-assigned weighting. This validates our theoretical framework: for open-ended tasks involving multiple non-dominating dimensions, accounting for rubric correlation yields a more robust aggregation of judge "votes" than simple averaging or LLM-based self-weighting.

Figure 3(b) shows the evolution of rubric counts across recursive rounds on JudgeBench. Two key phenomena emerge: first, the initial average of 7.4 rubrics grows rapidly and plateaus at approximately 20 by iteration 3, suggesting that RRD captures a comprehensive set of dimensions quickly before reaching a "saturation" point in novelty. Second, the framework is inherently adaptive, increasing the variance in rubric counts as it progresses. Consequently, recursion terminates quickly for simpler tasks but generates deeper criteria for complex tasks to resolve residual quality differences. A qualitative example of simpler tasks entailing fewer rubrics are provided in the Appendix I.

Ablations. We organize our ablations around three questions: whether RRD improves over simply generating more rubrics, whether each filtering component is helpful, and whether rubric generation depends on using the same model as the final judge.

First, Table 1(a) shows the value of RRD is beyond more rubric count. One-shot rubric generation with higher rubric budget of 30 improves the base judge but remains far below RRD_{WU} accuracy, indicating that recursive refinement discovers more discriminative criteria than a single-pass expansion. Second, Table 1(b) confirms that each filtering stage contributes meaningfully to the strong overall performance of RRD. Disabling any single component, such as the misalignment or duplication filter, still allows RRD to outperform most baselines, suggesting that its robustness does not depend on any

Table 1: Additional JudgeBench ablations. GPT-4o is used as the final judge throughout. In (a), rubric-based variants use the same budget of 30 rubrics. In (b), each row removes one filtering stage from the full pipeline. In (c), only the rubric proposer model is varied.

(a) Rubric Count

Variant	Acc.
Base	55.6
One-shot rubric generation	65.1
RRD_{WU}	**73.3**

(b) Filter Components

Variant	Acc.
w/o conflict filter	64.0
w/o misalignment filter	66.8
w/o duplication filter	68.6
RRD_{WU}	**73.3**

(c) Rubric Proposer

Proposer	Acc.
GPT-4o	73.3
Gemini	71.6
Claude	72.0
Qwen-235B	74.4

individual filter. That said, the full pipeline performs best, indicating that these filtering stages yield complementary system-level gains when used in combination. Third, Table 1(c) shows that, across settings, RRD maintains accuracy above 71 and substantially outperforms the baselines. It not only demonstrates that RRD delivers consistent gains across model families, but also shows that its benefits do not rely on proposer–judge self-consistency.

We report additional sensitivity analyses on recursion depth and sample-response generation in Appendix G.1. In particular, Table 4 ablates two implementation choices on JudgeBench: the recursion termination threshold and the sample-response generator used to condition rubric proposal. These results complement the main mechanistic ablations by showing that RRD remains robust under deeper recursion, while also confirming that high-quality and diverse sample responses are important for effective rubric decomposition.

4 RRD-based RFT Results

Previously in §3, we showed that RRD improves the accuracy of LLM judges on open-ended tasks by producing more comprehensive, discriminative and fine-grained rubrics. In this section, we ask whether these gains translate into learning gains when judges are used as generative reward models for reinforcement fine-tuning (RFT). RFT is the natural stress test for open-ended judging: even small systematic scoring biases can be amplified by the optimization loop, shaping model behavior in unintended ways. A reward model therefore must do more than standalone preference judges, it must provide stable, informative gradients that consistently promote the desired trade-offs across diverse prompts. We therefore use RFT to convert judge quality into an end-to-end signal, evaluating (i) reward reliability during training, and (ii) downstream performance of the resulting policies on in- and out-of-domain tasks.

Dataset and Training. We conduct RFT on 4K English, non-toxic, de-duplicated prompts sampled from WildChat [63], representing natural user–AI interactions. We employ Dr.GRPO [31] as our RFT algorithm for all settings (see details in Appendix F.2).

Then after training, to demonstrate the performance of resulting models, we evaluate the resulting policies on two rubric-based, open-ended generation benchmarks: BiGGen Bench [25] and HealthBench-Hard [2]. BiGGen Bench is a free-form generation benchmark spanning multiple core capabilities, and is evaluated with instance-specific criteria that capture what a good answer should contain for each prompt. HealthBench is a domain-specific benchmark containing 5K multi-turn clinical dialogues, scored using conversation-specific rubrics authored by physician experts, with importance-weighted criteria to reflect clinical priorities; the Hard split is designed to remain challenging and unsaturated.

We choose these two benchmarks because they provide complementary coverage of (i) in-distribution general assistant behavior and (ii) out-of-domain, high-stakes generalization. BiGGen Bench is closer to our training distribution (WildChat): both emphasize broad, real-world, open-ended assistant tasks, making BiGGen a natural test of whether RFT improves general helpfulness and capability across the kinds of prompts seen in the wild. In contrast, HealthBench-Hard focuses on clinical multi-turn dialogue with physician-defined notions of correctness, safety, and communication. It is substantially more domain-specific than open-domain chat, thereby stress-testing whether the learned policy maintain reliability in a high-stakes setting. The details of how scores for each dataset are computed can be found in Appendix F.2.

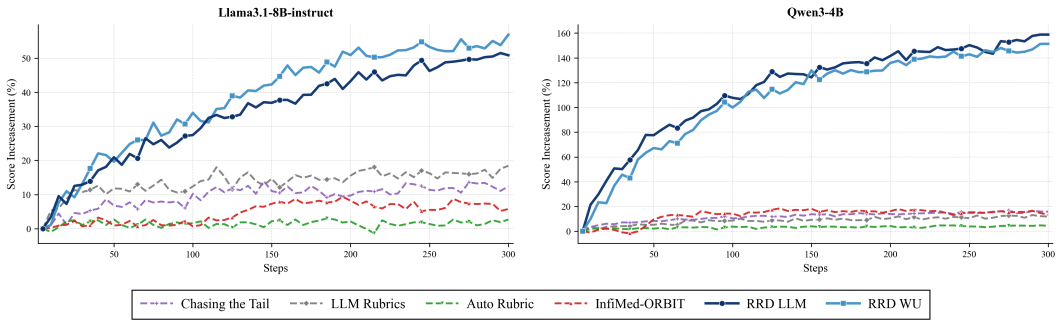


Figure 4: Validation reward improvement during training of Llama3.1-8B-instruct and Qwen3-4B models using various rubric generation methods. Both RRD_{WU} and RRD_{LLM} provide a significantly stronger reward signal than traditional rubric-based or iterative baselines. RRD_{WU}, in particular, shows superior training stability and higher cumulative reward gains across both architectures, indicating a more robust and granular supervision signal for RFT.

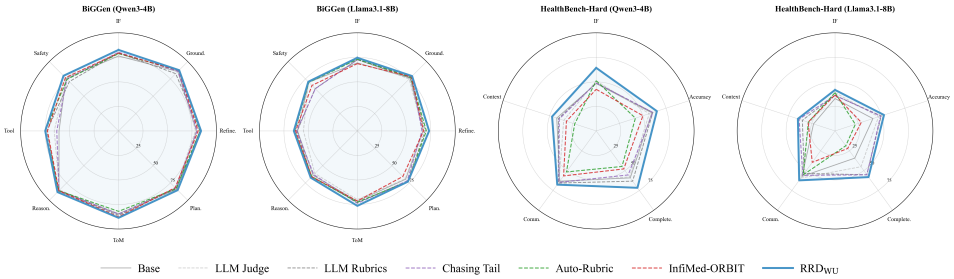


Figure 5: Multi-dimensional evaluation (scores in percentage) on BiGGen Bench (left) and HealthBench-Hard (right). Comparison of RRD_{WU} against five baseline methods using Llama-3.1 and Qwen3 base models. RRD_{WU} (solid blue) demonstrates robust improvements across all axes, particularly in Instruction Following (IF) and Completeness.

Models We fine-tune Qwen3-4B [59] and Llama3.1-8B [10] as policy models. GPT-4o serves as the rubric proposer and GPT-OSS-120B [1] determines rubric satisfaction for rollout responses during RFT.

4.1 Results

Reward Dynamics during Training. Figure 4 shows reward improvement during RFT for Qwen3-4B (left) and Llama3.1-8B-Instruct (right) under different rubric-generation strategies. Across both models, RRD produces a substantially stronger learning signal: rewards climb rapidly in the first $\sim$50–100 steps and continue improving throughout training, while prior baselines plateau early at low gains. Concretely, on Qwen3-4B, RRD_{WU} reaches roughly 150–160% reward improvement by the end of training, compared to only $\sim$10–20% for LLM Rubrics and Chasing the Tail; on Llama3.1-8B-Instruct, it attains about 55–60% versus $\sim$10–20% for baselines. Both RRD variants outperform alternatives, with RRD_{WU} also showing the smoothest, most stable curves. Overall, the large absolute gaps (often $\geq$3–10$\times$ higher improvements) indicate that RRD yields more discriminative and better-calibrated rewards, enabling sustained optimization rather than early saturation on open-ended tasks.

Policy Performance on BiGGen Bench. The superior reward signal provided by RRD translates into state-of-the-art performance for the resulting policies across diverse generation capabilities, as shown in the radar chart (Figure 5) with more details breakdown in Table 5 (Appendix G.3). For both Qwen3-4B and Llama3.1-8B backbones, RRD_{WU} achieves the highest overall scores of 82.8% and 71.1%, respectively. The method is particularly effective at improving instruction following (IF), refinement, and reasoning, all while maintaining robust performance on safety axes.

Generalization to High-Stakes Domains. Performance gain by RRD is further confirmed on HealthBench as shown in the same radar chart (Figure 5) with more details breakdown in Table 6

(Appendix G.3). These results confirm that the benefits of RRD transfer effectively to high-stakes, domain-specific tasks, such as the medical field. RRD_{WU} consistently yields the best policy performance, particularly in IF, accuracy, and completeness ($+16.0\%$, $+5.5\%$, and $+12.5\%$ points respectively on the Qwen3-4B backbone). These results underscore the framework's ability to provide granular supervision that aligns with complex, physician-authored evaluation criteria.

5 Related Works

LLM-as-a-Judge. Holistic LLM judges can approximate human preferences [64, 30, 12, 55, 4], but often suffer from bias [37, 42], inconsistency [64, 18], and opacity [24, 14], particularly for subjective and multidimensional open-ended tasks [50]. Rubric-assisted judges address some of these issues [24, 19, 23, 2], but many rely on static or heuristic rubrics that lack scalability or prompt-specific nuance. We instead propose a dynamic approach that recursively decomposes criteria for broad coverage and discriminative power.

Rubrics-based Rewards Structured rubrics are increasingly used not only for evaluation but also as reward signals for training [16]. Recent work improves rubric quality through Propose–Evaluate–Revise pipelines [58], score-differential optimization [61], and domain-conditioned generation with retrieved exemplars [54]. Others use rubrics as priors for data generation or policy learning [20, 65], or study online rubric generation [41]. RRD differs from prior work by introducing explicit mechanisms that enforce rubric *informativeness*, *comprehensiveness*, and *non-redundancy* in LLM-generated criteria, and shows its effectiveness as a source of high-quality RFT reward signals.

Reinforcement Fine-Tuning (RFT) for Open-Ended Tasks. Reinforcement Learning from Verifiable Rewards (RLVR) has achieved strong results in domains with objective ground truth, such as mathematics and coding, as shown by DeepSeek-R1 and process-reward methods [17, 48, 56]. Extending RLVR to open-ended tasks, such as creative writing and brainstorming, remains challenging because ground-truth verifiers are unavailable [62, 46]. Recent work addresses this by using LLM judges as proxy reward models for algorithms such as GRPO. Our work contributes to this frontier by providing a "Generative Verifier" via RRD that offers the granularity and reliability required to stabilize RFT in subjective and open-ended domains.

6 Conclusion

In this work, we introduced Recursive Rubric Decomposition (RRD), a practical framework for generating informative, comprehensive and non-redundant rubrics for rubric-based LLM judges and reward modeling for open-ended tasks. RRD operates by recursively decomposing high-level evaluation criteria into granular, discriminative rubrics, filtering for positive edge and non-redundancy to ensure reliability while optimizing weights to account for correlation structures and prevent the over-representation of highly correlated metrics. Empirically, we demonstrate that RRD significantly improves the accuracy of LLM judges in a training-free setting, and establishes its effectiveness as a high-fidelity reward model for RFT. Policies trained with RRD-derived rewards exhibit stronger alignment with human preferences on complex, open-ended generation tasks. These results highlight the importance of structured, granular, and statistically grounded rubric generation as a critical pathway toward scalable, interpretable, and reliable alignment for the next generation of LLMs.

Acknowledgments and Disclosure of Funding

Funding. This work was conducted at Meta and supported by Meta.

Competing interests. The authors' Meta affiliation is disclosed on the title page. The authors declare no other competing financial interests related to this work.

References

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[2] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel,

et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, 2025.

[5] Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024.

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[7] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[8] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

[9] Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11(4), 2010.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[11] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

[12] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[13] Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf. *arXiv preprint arXiv:2410.14872*, 2024.

[14] Jasmina Gajcin, Erik Miehling, Rahul Nair, Elizabeth Daly, Radu Marinescu, and Seshu Tirupathi. Interpreting llm-as-a-judge policies via verifiable global explanations. *arXiv preprint arXiv:2510.08120*, 2025.

[15] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[16] Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

[17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[18] Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in llm-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004, 2025.

[19] Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*, 2024.

[20] Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025.

[21] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[22] Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. In *Advances in Neural Information Processing Systems*, 2024.

[23] Ahrii Kim. Rubric-mqm: Span-level llm-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, 2025.

[24] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[25] Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models, 2024. URL `https://arxiv.org/abs/2406.05761`.

[26] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[27] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[28] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of machine learning and systems*, 2:230–246, 2020.

[29] Sunzhu Li, Jiale Zhao, Huimin Ren, Zhenlin Wei, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang, and Chen Wei. 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)*, pages 31320–31344, 2026.

[30] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[31] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[32] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

[35] Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7162–7200, 2025.

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[37] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[38] Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7957–7968, 2023.

[39] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[40] Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, et al. A systematic survey of automatic prompt optimization techniques. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33066–33098, 2025.

[41] MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Yunzhong He, and Afra Feyza Akyürek. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*, 2025.

[42] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

[43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[44] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[45] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[46] Toby Simonds, Kevin Lopez, Akira Yoshiyama, and Dominique Garmier. Self rewarding self improving. *arXiv preprint arXiv:2505.08827*, 2025.

[47] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

[48] Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.

[49] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*, 2024.

[50] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.

[51] Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. *arXiv preprint arXiv:2507.18624*, 2025.

[52] Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024.

[53] Pengkai Wang, Pengwei Liu, Zhijie Sang, Congkai Xie, Hongxia Yang, et al. Infimed-orbit: Aligning llms on open-ended complex tasks via rubric-based incremental training. *arXiv preprint arXiv:2510.15859*, 2025.

[54] Pengkai Wang, Qi Zuo, Pengwei Liu, Zhijie Sang, Congkai Xie, and Hongxia Yang. Infimed-orbit: Aligning llms on open-ended complex tasks via rubric-based incremental training. *arXiv preprint arXiv:2510.15859*, 2025.

[55] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023.

[56] Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

[57] Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, et al. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling. *arXiv preprint arXiv:2510.17314*, 2025.

[58] Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, et al. Auto-rubric: Learning to extract generalizable criteria for reward modeling. *arXiv preprint arXiv:2510.17314*, 2025.

[59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[60] Zihuiwen Ye, Fraser David Greenlee, Max Bartolo, Phil Blunsom, Jon Ander Campos, and Matthias Gallé. Improving reward models with synthetic critiques. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4506–4520, 2025.

[61] Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. *arXiv preprint arXiv:2509.21500*, 2025.

[62] Mengyu Zhang, Xubo Liu, Siyu Ding, Weichong Yin, Yu Sun, Hua Wu, Wenya Guo, and Ying Zhang. Auditable-choice reframing unlocks rl-based verification for open-ended tasks. *arXiv preprint arXiv:2511.02463*, 2025.

[63] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

[64] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[65] Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, et al. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025.

Appendix

Table of Contents

A Limitations and Future Directions **16**

B Impact Statement **16**

C Algorithm **17**

D Additional Notes on Derivations and Proofs **18**

D.1 Probability Upper Bound for Incorrect Rubric-based Judge Verdict (Eq.2) 18

D.2 Rubric Weighting in Whitened Space 19

E Discussion **22**

E.1 Proxy Design Under Latent Preferences 22

E.2 Computation Overhead 23

F Implementation Details **24**

F.1 Baselines 24

F.2 RFT Training Details 24

G Additional Results **26**

G.1 Additional LLM Judge Ablations 26

G.2 Ablation for Different Base Rubrics 26

G.3 RFT Training Results 26

G.4 Human Evaluation 27

H Prompts **29**

H.1 Rubric Generation 29

H.2 Filtering 31

H.3 Evaluation 32

I Qualitative Examples **34**

Appendix

A Limitations and Future Directions

While RRD effectively automates rubric-based evaluation, it introduces additional computational overhead due to iterative rubric refinement. Future work may explore budgeted refinement strategies and caching mechanisms to support more efficient large-scale deployment. In addition, while proxy-based filtering is necessary as open-ended generation lacks a unique canonical target and its evaluation rubrics are not directly observed, developing more sophisticated filtering procedures, such as multi-reference preference checks, is an important direction for further improving rubric quality.

B Impact Statement

This paper advances automated rubric construction and refinement for evaluating open-ended generation and for using rubrics as generative reward models in RFT. We propose RRD, a practical framework for *controlled rubric generation* that targets three desiderata: informativeness, comprehensiveness, and non-redundancy, and we demonstrate empirically that these properties translate into improved rubric-based LLM judging and stronger, more reliable reward signals for RFT. We discuss several dimensions of potential impact below.

Potential positive impacts. Our work strengthens the foundations of rubric-based evaluation by clarifying which rubric qualities are most predictive of judge accuracy, and by operationalizing them in a practical pipeline. By decomposing coarse criteria into discriminative rubrics, sanity checking for gross misalignment and duplication, and accounting for correlation during aggregation, RRD can yield more accurate and interpretable judgments than opaque or heuristic holistic scoring, supporting more consistent evaluation outcomes and actionable error analysis. Beyond evaluation, RRD bridges from judging to learning: it produces more auditable and fine-grained reward signals that can improve training on non-verifiable, open-ended objectives, where reward design is a central bottleneck. Finally, RRD is modular and easy to adopt in existing workflows: we demonstrate that it requires no new model architectures and can be integrated with standard "proposer + judge" stacks for data curation, preference evaluation, and reward modeling in RFT.

Responsible use and mitigation. More effective rubric-based rewards may also be misused to optimize for undesirable objectives, especially if an adversary can influence the rubric proposer, sampled responses, or evaluation context. Rubric generation is also inherently value-laden and may encode biases from the proposer model or the sampled responses. For high-stakes deployments, we recommend safeguards such as multi-reference or multi-judge checks and human-in-the-loop review for sensitive axes.

C Algorithm

Algorithm 1 RRD: Recursive Rubric Decomposition

Require: Let $\mathcal{P}$ be the prompt set; $\{R_i\}_{i=1}^{n} \in \mathcal{R}$ be sampled responses for each prompt $P \in \mathcal{P}$; Ψ be the LLM-based rubric proposer.
Procedure:
for each $P \in \mathcal{P}$: **do**
  Let Ψ generate an initial rubric list $\mathcal{G}_0 = \Psi(P, \{R_i\}_{i=1}^{n})$,
  Initialize $\mathcal{G}_{\text{final}} \leftarrow \mathcal{G}_0$, iteration counter $t \leftarrow 0$, $|\mathcal{G}_{\text{filtered}}| \leftarrow 0$
  while $|\mathcal{G}_{\text{filtered}}| < N$ **do**
    $t \leftarrow t + 1$
    for each rubric $g_m \in \mathcal{G}_{t-1}$ **do** *Step 1: Rubric Evaluation*
      $\mathcal{R}_m \leftarrow \{ R_i \in \{R_i\}_{i=1}^{n} : g_m(P, R_i) = 1 \}$
      if $|\mathcal{R}_m| \geq n$ **then** *Step 2: Rubric Decomposition*
        $\mathcal{G}_m^{\text{new}} \leftarrow \Psi(g_m, \mathcal{R}_m)$
        $\mathcal{G}_{t-1} \leftarrow \mathcal{G}_{t-1} \cup \mathcal{G}_m^{\text{new}}$
      end if
    end for
    $\mathcal{G}_t \leftarrow \{ m \in \mathcal{G}_{t-1} \mid \neg\text{misalign}(m; P) \wedge \neg\text{conflict}(m; \mathcal{G}_{t-1}) \wedge \nexists\, m' \in \mathcal{G}_{t-1},\, m' \neq m : \text{duplication}(m, m') \}$ *Step 3: Rubric Filter*
    $\mathcal{G}_{\text{final}} \leftarrow \mathcal{G}_t$
  end while
end for
return $\mathcal{G}_{\text{final}}$

D Additional Notes on Derivations and Proofs

D.1 Probability Upper Bound for Incorrect Rubric-based Judge Verdict (Eq.2)

Derivation. Condition on the class $Y = y$, an error is $\{\widehat{Y} \neq Y\} \iff \{\Gamma \leq 0\}$. Let $W_y = \mathbf{w}^\top Z$,

$$\Pr(\Gamma \leq 0 \mid Y = y) = \Pr(W_y \leq -\Delta_m \mid Y = y).$$

For a sub-Gaussian, it is trivial to show that

$$\Pr(X \leq -a) = \Pr(e^{-\lambda X} \geq e^{\lambda a}) \leq \mathbb{E}[e^{-\lambda X}]e^{-\lambda a} \leq \exp(\tfrac{\lambda^2\sigma^2}{2} - \lambda a).$$

For $\lambda > 0$,

$$\Pr(X \leq -a) \;\leq\; \exp\Big(-\frac{a^2}{2\sigma^2}\Big).$$

Substitute a with Δ_m and σ^2 with $= V_m(y)$,

$$\Pr(\Gamma \leq 0 \mid Y = y) = \Pr(W_y \leq -\Delta_m \mid Y = y) \leq \exp\Big(-\frac{\Delta_m^2}{2V_m(y)}\Big),$$

Therefore, by the law of total probability,

$$\Pr(\widehat{Y} \neq Y) = \pi_+ \Pr(\Gamma \leq 0 \mid Y = +1) + \pi_- \Pr(\Gamma \geq 0 \mid Y = -1),$$

Hence,

$$\begin{aligned}
\Pr(\widehat{Y} \neq Y) \;&\leq\; \pi_+ \, e^{-\Delta_m^2/(2V_m(+1))} \;+\; \pi_- \, e^{-\Delta_m^2/(2V_m(-1))} \\
&\leq\; \max\Big\{e^{-\Delta_m^2/(2V_m(+1))},\; e^{-\Delta_m^2/(2V_m(-1))}\Big\} \\
&\leq\; \exp\big(-\min\{\Delta_m^2/(2V_m(+1)),\, \Delta_m^2/(2V_m(-1))\}\big),
\end{aligned}$$

Now, we complete the derivation for Eq.2.

Note. As a side note following Eq.2, the motivation for expanding the set of rubrics with positive information edge is clearer under the *idealized case* where the rubrics are *equicorrelated, equal-variance, equal-weight*.

Recall $V_m(Y) = \mathrm{Var}(\sum_{k=1}^m w_k Z_k \,|\, Y)$. With $w_k = 1$,

$$\begin{aligned}
V_m(Y) &= \mathrm{Var}\Big(\sum_{k=1}^m Z_k \,\Big|\, Y\Big) \\
&= \sum_{k=1}^m \mathrm{Var}(Z_k \mid Y) \;+\; 2 \sum_{1 \leq i < j \leq m} \mathrm{Cov}(Z_i, Z_j \mid Y).
\end{aligned}$$

Plugging in the equicorrelation/equal-variance values:

$$\begin{aligned}
&\sum_{k=1}^m \mathrm{Var}(Z_k \mid Y) = m\,\sigma^2, \\
&\mathrm{Cov}(Z_i, Z_j \mid Y) = \rho\,\sigma^2 \;\; (i \neq j).
\end{aligned}$$

There are $\binom{m}{2} = \frac{m(m-1)}{2}$ distinct pairs, so

$$\begin{aligned}
V_m(Y) &= m\,\sigma^2 \;+\; 2 \cdot \frac{m(m-1)}{2}\,\rho\,\sigma^2 \\
&= \sigma^2\big[m + (m^2 - m)\rho\big].
\end{aligned}$$

Under the symmetric assumptions of the corollary, this does not depend on $Y = +1$ or $Y = -1$, so we simply write

$$V_m \;=\; \sigma^2\big[m + (m^2 - m)\rho\big].$$

As such, with the idealized assumptions of $w_k \equiv 1$, $\mu_k \equiv \mu > 0$, $\mathrm{Var}(S_k \mid Y) = \sigma^2$, and $\mathrm{Corr}(S_i, S_j \mid Y) = \rho\ (i \neq j)$,

$$\mathbb{P}(\widehat{Y} \neq Y) \;\leq\; \exp\!\left(-\frac{m\,\mu^2}{2\,\sigma^2\,[1 + (m-1)\rho]}\right).$$

Following this, it becomes obvious that when rubrics are uncorrelated (i.e., ρ = 0), the prediction error can be consistently reduced by consistently expanding the rubric pool, provided that each added rubric contributes a positive information edge (i.e., its prediction direction aligns with human preference). **This supports the recursive process to encourage more extensive rubric search for broader coverage**. In practice, however, rubrics are neither independent nor equicorrelated. When the exact information edge μ is *unobservable* and *individual evaluation dimensions carry comparable importance* (i.e., no single criterion dominating the overall judgment as often the case for open-ended tasks), **the aggregation of extended rubric set must explicitly account for the rubric correlation structure**. This motivates incorporating correlation-aware normalization via an effective weighting scheme, which we introduce below.

D.2 Rubric Weighting in Whitened Space

Define the ratio to be optimized as:

$$\Xi = (\boldsymbol{w}^\top \boldsymbol{\mu})^2 / (\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w})$$

Also define whitened coordinates:

$$u \;:=\; \frac{\Sigma^{1/2} w}{\big\|\Sigma^{1/2} w\big\|_2} \in \mathbb{R}^m, \qquad z \;:=\; \frac{\Sigma^{-1/2}\mu}{\big\|\Sigma^{-1/2}\mu\big\|_2} \in \mathbb{R}^m, \qquad \kappa := \Big\|\Sigma^{-1/2}\mu\Big\|_2^2.$$

Lemma D.1. *For any $w \neq 0$,*

$$\Xi(w) \;=\; \kappa\,\big\langle u,\, z\big\rangle^2 \qquad \textit{and} \qquad \frac{\Xi(w)}{\Xi(w^\star)} \;=\; \cos^2\!\angle\big(\Sigma^{1/2} w,\; \Sigma^{-1/2}\mu\big) \in [0,1].$$

Proof. By direct algebra: $(w^\top \mu)^2 = \big\langle \Sigma^{1/2} w,\, \Sigma^{-1/2}\mu \big\rangle^2 = \big\|\Sigma^{1/2} w\big\|_2^2 \,\big\|\Sigma^{-1/2}\mu\big\|_2^2 \,\langle u,\, z\rangle^2$, and $w^\top \Sigma w = \big\|\Sigma^{1/2} w\big\|_2^2$. Thus, $\Xi(w^\star) = \kappa$, where $w^\star$ is the optimized weight. □

Lemma D.2. *Let $X \in \mathbb{R}^m$ be a zero-mean rubric-score vector with covariance $\Sigma = \mathbb{E}[XX^\top] \succ 0$ (Note: X is sub-Gaussian given bounded rubric scores in $[0,1]$ and X_i for $i \in [1, N]$ are i.i.d. copies of X).*

Define the sample covariance

$$\widehat{\Sigma} := \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\top, \qquad r_{\mathrm{eff}}(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}} \in [1, m].$$

Then there exist constants $c,\ C > 0$ such that, for all $t \geq 0$,

$$\big\|\widehat{\Sigma} - \Sigma\big\|_{\mathrm{op}} \;\leq\; C\,\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\frac{r_{\mathrm{eff}}(\Sigma)}{N}} + \frac{r_{\mathrm{eff}}(\Sigma)}{N} + \frac{t}{\sqrt{N}}\right) \quad \textit{with probability at least } 1 - 2e^{-ct^2}.$$

In particular,

$$\mathbb{E}\,\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \;\leq\; C\,\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\frac{r_{\mathrm{eff}}(\Sigma)}{N}} + \frac{r_{\mathrm{eff}}(\Sigma)}{N}\right).$$

Moreover, if $\lambda_{\min}(\Sigma) > 0$ and $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \leq \frac{1}{2}\lambda_{\min}(\Sigma)$, then

$$\left\|\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2}\right\|_{\mathrm{op}} \;\le\; \tfrac{1}{2}\,\lambda_{\min}(\Sigma)^{-3/2}\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}},$$

Thus whitened uniform weights computed from $\widehat{\Sigma}$ are consistent and numerically stable.

Proof. Let $Y_i := X_i X_i^\top - \Sigma$, estimation error for or covariance Σ is $\widehat{\Sigma} - \Sigma = \frac{1}{N}\sum_{i=1}^N X_i^\top X_i - \Sigma = \frac{1}{N}\sum_{i=1}^N Y_i$ with $\mathbb{E}Y_i = 0$. To control the random fluctuation $\widehat{\Sigma} - \Sigma$ in operator norm, we will apply a matrix Bernstein inequality to the sum $\sum_{i=1}^N Y_i$. This requires bounding the matrix variance proxy $V := \left\| \sum_{i=1}^N \mathbb{E}\, Y_i^2 \right\|_{\mathrm{op}}$, which captures the second-order size of the fluctuations.

Since $Y_i^2 = (X_i X_i^\top - \Sigma)^2 = X_i X_i^\top X_i X_i^\top - X_i X_i^\top \Sigma - \Sigma X_i X_i^\top + \Sigma^2$, we have $\mathbb{E}\, Y_i^2 = \mathbb{E}[XX^\top XX^\top] - \Sigma^2$.

Thus, for any unit u, $u^\top \mathbb{E}[XX^\top XX^\top]\, u = \mathbb{E}[\|X\|_2^2\,(u^\top X)^2] \le \sqrt{\mathbb{E}\|X\|_2^4}\,\sqrt{\mathbb{E}(u^\top X)^4}$. Sub-Gaussian moment bounds give $\mathbb{E}\|X\|_2^4 \le C(\operatorname{tr}\Sigma)^2$ and $\mathbb{E}(u^\top X)^4 \le C(u^\top \Sigma u)^2 \le C\|\Sigma\|_{\mathrm{op}}^2$, where C is universal constant. Plugging this back we get

$$u^\top \mathbb{E}[XX^\top XX^\top]\, u \le C\,\|\Sigma\|_{\mathrm{op}}\,\operatorname{tr}(\Sigma), \qquad \Rightarrow \quad \left\|\mathbb{E}[XX^\top XX^\top]\right\|_{\mathrm{op}} \le C\,\|\Sigma\|_{\mathrm{op}}\,\operatorname{tr}(\Sigma).$$

Therefore,

$$V = \left\| \sum_{i=1}^N \mathbb{E} Y_i^2 \right\|_{\mathrm{op}} \le N\,C\,\|\Sigma\|_{\mathrm{op}}\,\operatorname{tr}(\Sigma) = C\,N\,\|\Sigma\|_{\mathrm{op}}^2\, r_{\mathrm{eff}}(\Sigma),$$

where $r_{\mathrm{eff}}(\Sigma) = \operatorname{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$.

To apply the non-commutative Bernstein inequality, we also require a uniform sub-exponential bound on the size of the summands, encoded by a parameter $L > 0$ such that $\|Y_i\|_{\mathrm{op}}$ is sub-exponential with $\|Y_i\|_{\psi_1} \le L$. Since

$$\|Y_i\|_{\mathrm{op}} = \|X_i X_i^\top - \Sigma\|_{\mathrm{op}} \le \|X_i X_i^\top\|_{\mathrm{op}} + \|\Sigma\|_{\mathrm{op}} = \|X_i\|_2^2 + \|\Sigma\|_{\mathrm{op}},$$

and X is sub-Gaussian, it follows that $\|Y_i\|_{\mathrm{op}}$ is sub-exponential with $\|Y_i\|_{\psi_1} \le C\|\Sigma\|_{\mathrm{op}}$ for some universal C. With V and $L = C\|\Sigma\|_{\mathrm{op}}$, we now invoke the intrinsic-dimension version of the matrix Bernstein inequality for the sum $\sum_{i=1}^N Y_i$ and have:

$$\mathbb{P}\left(\left\| \frac{1}{N}\sum_{i=1}^N Y_i \right\|_{\mathrm{op}} \ge s \right) \le 2\exp\left\{ -c\,\min\left(\frac{N^2 s^2}{V},\, \frac{Ns}{L} \right) \right\},$$

where V and L are as above and $c > 0$ is a universal constant.

Substituting $V \le CN\|\Sigma\|_{\mathrm{op}}^2 r_{\mathrm{eff}}(\Sigma)$ yields

$$\mathbb{P}\Big(\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}} \ge s \Big) \le 2\exp\left\{ -cN\,\min\left(\frac{s^2}{C\|\Sigma\|_{\mathrm{op}}^2 r_{\mathrm{eff}}(\Sigma)},\, \frac{s}{C\|\Sigma\|_{\mathrm{op}}} \right) \right\}.$$

Choosing $s = C\,\|\Sigma\|_{\mathrm{op}}\Big(\sqrt{r_{\mathrm{eff}}/N} + r_{\mathrm{eff}}/N + t/\sqrt{N} \Big)$ gives, after rescaling constants,

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}} \;\le\; C\,\|\Sigma\|_{\mathrm{op}} \left(\sqrt{\frac{r_{\mathrm{eff}}(\Sigma)}{N}} + \frac{r_{\mathrm{eff}}(\Sigma)}{N} + \frac{t}{\sqrt{N}} \right),$$

with probability at least $1 - 2e^{-ct^2}$. Integrating this tail yields the stated expectation bound.

For symmetric positive definite $A,\ B$ with $\lambda_{\min}(A), \lambda_{\min}(B) \ge \gamma > 0$ and any C^1 scalar function f on $[\gamma, \infty)$, the spectral calculus gives

$$\|f(A) - f(B)\|_{\mathrm{op}} \le \sup_{t \in [\gamma, \infty)} |f'(t)|\, \|A - B\|_{\mathrm{op}}.$$

Applying this with $f(t) = t^{-1/2}$ yields $\sup_{t \ge \gamma} |f'(t)| = \frac{1}{2}\gamma^{-3/2}$. Taking $A = \widehat{\Sigma}$, $B = \Sigma$, and $\gamma = \frac{1}{2}\lambda_{\min}(\Sigma)$ (valid when $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \le \frac{1}{2}\lambda_{\min}(\Sigma)$) gives

$$\left\|\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2}\right\|_{\mathrm{op}} \le \tfrac{1}{2}\lambda_{\min}(\Sigma)^{-3/2}\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}},$$

□

Theorem D.3. *Let the weight space be* $\mathcal{W} := \left\{ w \in \mathbb{R}^m : \left\|\Sigma^{1/2} w\right\|_2 = 1 \right\}$. *Then*

$$\sup_{w \in \mathcal{W}} \inf_{\mu \in \mathcal{U}} \mathrm{SNR}(w) \quad \textit{is attained by} \quad w_{wu} \propto \Sigma^{-1/2}\mathbf{1}$$

Proof. Assume whitened edge $\Sigma^{-1/2}\mu$ is nonnegative and exchangeable in distribution across coordinates. We fix its magnitude $\left\|\Sigma^{-1/2}\mu\right\|_2^2 = c > 0$ but leave its direction unknown.

By Lemma D.1, $\Xi(w) = c \langle u, z\rangle^2$ with $u = \Sigma^{1/2} w / \left\|\Sigma^{1/2} w\right\|_2$ and $z = \Sigma^{-1/2}\mu / \left\|\Sigma^{-1/2}\mu\right\|_2$. Under the assumptions, z is supported on the positive orthant and is exchangeable with fixed norm; the *least favorable* z against a given u is the one minimizing $\langle u, z\rangle^2$. By symmetry and convexity of the feasible set of z, the u that maximizes the worst-case inner product is the barycenter of the positive orthant, i.e., $u \propto \mathbf{1}$. Translating back gives $w \propto \Sigma^{-1/2}\mathbf{1}$. □

E Discussion

E.1 Proxy Design Under Latent Preferences

Optimizing open-ended generation remains challenging because the target supervision is inherently latent. In particular, the human preference label is unobserved, and therefore the rubric edge μ_k is not identifiable from unlabeled data alone without additional assumptions [9, 32]. Consequently, in the absence of a closed-form analytical solution to this optimization problem, the use of proxy-based filtering criteria and practical stopping rules is not an ad hoc shortcut, but a necessary and standard design choice in this setting. Similar forms of proxy supervision, filtering, and empirically motivated selection have been widely adopted in prior work on open-ended instruction tuning, preference modeling, and evaluation [11, 7, 35].

Importantly, our argument is not that the absence of an analytical solution validates any heuristic choice. Rather, it motivates the need for transparent proxy criteria whose utility should be assessed empirically through ablations and independent downstream evaluations. This observation also clarifies the appropriate standard for evaluating rubric-generation methods. Since the latent preference objective cannot be directly optimized in closed form, the relevant question is not whether every local step admits a strict optimality guarantee, but whether the overall procedure produces reliable end-to-end improvements across diverse evaluations.

Our experiments support this criterion for both LLM judging and reward modeling. As an example in LLM judging, RRD substantially improves JudgeBench accuracy from 55.6 to 73.3. Similarly, for reward modeling, RRD yields stronger and more stable reward improvements throughout RFT training, and these gains transfer to external benchmarks with independently constructed evaluation criteria, including HealthBench and BigGenBench. Moreover, our ablations show that the improvements arise from the full RRD pipeline rather than from rubric count alone or any single component, and that the gains are not tied to a particular model. This mitigates concerns that the proposed procedure merely exploits a single evaluator, a single model family, or a superficial increase in the number of rubrics.

We emphasize this point in the context of the misalignment filter. The filter is intended as a conservative sanity check on rubric directionality, not as a claim of optimal rubric identification. It is not meant to certify optimality of the latent preference objective, nor does it assume that the stronger reference model is universally preferable on every individual criterion. Instead, it acts as a guardrail against clear directionality errors by flagging rubrics whose induced preferences systematically favor a weaker reference model over a stronger one. As shown in Table 1, even when the misalignment filter is disabled, RRD still outperforms most baselines. Nevertheless, this simple directional check contributes meaningful additional improvement to the overall result. This ablation suggests that the filter is helpful, but not solely responsible for the effectiveness of the method. In other words, the empirical gains do not rely on the misalignment filter alone, while the filter still provides a useful safeguard against obviously reversed rubric directions.

The use of such a proxy is also consistent with prior empirical and methodological evidence. Strong LLM judges have been shown to closely track human pairwise preferences on open-ended prompts on average, with GPT-4 achieving human-comparable agreement in MT-Bench and Chatbot Arena [64]. Structured LLM-based evaluation frameworks can also improve correlation with human judgments in NLG evaluation [30]. More broadly, using a more capable model or system as a source of supervision is a standard alignment pattern, including in AI-feedback settings such as Constitutional AI and related RLAIF approaches [3, 27]. Our use of the stronger reference model is narrower than these full supervision settings: it is used only to detect systematic directionality inversions during rubric construction, while final effectiveness is evaluated through downstream judging accuracy, reward-modeling behavior, and transfer to external benchmarks with independently constructed criteria.

Finally, the misalignment filter is not meant to be the only possible proxy. It could be replaced or strengthened by more sophisticated preference-proxy procedures, such as checks against multiple strong reference models, uncertainty-aware aggregation, or axis-specific calibration for criteria that are known to be inversion-prone. In this work, our goal is to demonstrate the effectiveness of RRD as a general rubric-construction framework. We therefore use a deliberately simple and transparent filtering mechanism, while leaving the exploration of more advanced filtering strategies as an interesting direction for future work.

E.2 Computation Overhead

We quantify the computational overhead of RRD using JudgeBench as a representative setting. We report token usage rather than wall-clock time, since wall-clock measurements are heavily confounded by provider-side latency, caching, batching, and API implementation details, making them difficult to compare fairly across methods. The table below reports the average input and output tokens per prompt per iteration, as well as the default number of rounds used by each method. For RRD, the number of rounds is governed by its default early-stopping rule: a rejection threshold of 15. With this setting, 52% of prompts terminate by iteration 3, and 95% terminate by iteration 4.

Overall, RRD has a token cost comparable to Chasing the Tail and moderately higher than one-shot LLM Rubrics. In exchange, RRD delivers substantially stronger LLM judging and RFT performance.

Table 2: Average token usage and default number of rounds for rubric-generation methods.

Method	**Avg. input tokens**	**Avg. output tokens**	**# Rounds**
LLM Rubrics (w. resp.)	53,758	579	1
Chasing the Tail	66,331	806	4
RRD	83,445	757	3–4

F Implementation Details

F.1 Baselines

We compare RRD against the base model (preference labeling without explicit rubrics) and several rubric-based judge baselines that differ in their rubric generation strategies. Specifically, we include: (1) LLM Rubrics (w/o resp.): rubrics are proposed based solely on the prompt; (2) LLM Rubrics (w. resp.): rubrics are generated with access to sample responses for better grounding; (3) Chasing the Tail [61], a state-of-the-art iterative method that optimizes rubrics to differentiate high-quality response pairs; (4) Auto-Rubric [57], a training-free framework for explicit reward parameterization that learns natural-language rubrics by locally inducing discriminative criteria through verification-driven refinement and globally compressing them into a compact hierarchical set of evaluation dimensions and verification checks; (5) InfiMed-ORBIT [53], a rubric-based incremental-training framework that uses dynamically constructed rubric feedback for open-ended complex tasks; and (6) RRD variants: we evaluate three weighting schemes – $\text{RRD}_{\text{uniform}}$ (uniform weights), RRD_{LLM} (LLM-assigned weights), and RRD_{UW} (whitened weights).

For Auto-Rubric [57], we use the authors' original GitHub repository and follow the implementation protocol reported in their paper. InfiMed-ORBIT [53] was originally proposed for rubric-based reward modeling and incremental training, rather than direct pairwise judge-accuracy evaluation. To evaluate it in our setting, we adapt its rubric-construction procedure to the same pairwise preference-judgment protocol used for other rubric-based baselines. A practical requirement of InfiMed-ORBIT is access to seed rubrics. However, such seed rubrics are not available for either JudgeBench or PPE. We therefore construct seed rubrics by prompting an LLM to generate rubrics for the set of seed questions. For JudgeBench, we construct seed rubrics using questions with Claude-generated responses and evaluate on the GPT-4o-generated response pairs. For PPE, we randomly sample 10% of the questions as the seed set and evaluate on the remaining 90%. This adaptation keeps the final evaluation protocol fixed across methods while enabling InfiMed-ORBIT to be applied to datasets without human-authored seed rubrics.

We employ GPT-4o [21] and Llama-3.1-405B [10] as rubric generator and final judge, and consistently use GPT-4o and Gemini 2.5-pro [6] as sample response generator (4 each).

F.2 RFT Training Details

Reinforcement Fine-Tuning (RFT) aligns a policy model π_θ by maximizing expected return under prompts $q \sim p_Q$ and model-generated responses $o \sim \pi_\theta(\cdot \mid q)$:

$$J(\pi_\theta) = \mathbb{E}_{q\sim p_Q}\Big[\mathbb{E}_{o\sim\pi_\theta(\cdot|q)}[R(q,o)] - \beta\, D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid q)\,\|\,\pi_{\mathrm{ref}}(\cdot \mid q)\big)\Big],$$

where $R(q,o) = \sum_{t=1}^{|o|} r(s_t, o_t)$ is the trajectory return and π_{ref} is an optional reference policy.

In practice, policy-gradient methods such as PPO [43] optimize a clipped surrogate objective using an importance ratio against the pre-update policy $\pi_{\theta_{\text{old}}}$:

$$J_{\mathrm{PPO}}(\pi_\theta) = \mathbb{E}_{q\sim p_Q,\, o\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \sum_{t=1}^{|o|} \min\Big(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big),$$

with $r_t(\theta) = \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|q,o_{<t})}$. GRPO [44] removes the need for a learned value function by sampling a group of G responses $\{o_i\}_{i=1}^G$ per question and using a group-relative advantage for all tokens in response o_i:

$$\hat{A}_i = \frac{R(q,o_i) - \mathrm{mean}(\{R(q,o_j)\}_{j=1}^G)}{\mathrm{std}(\{R(q,o_j)\}_{j=1}^G)}.$$

However, the standard GRPO objective additionally aggregates token losses with a per-response length normalization factor $1/|o_i|$, and together with per-question $\mathrm{std}(\cdot)$ normalization this induces an optimization bias. Dr.GRPO [31] mitigates this bias by removing the $1/|o_i|$ token-aggregation term and removing the per-question $\mathrm{std}(\cdot)$ normalization, yielding an unbiased policy-gradient estimator and improved token efficiency.

In our experiments, we consistently apply Dr.GRPO algorithm in RFT training across baselines and RRD variants. We provide RFT training parameters as below in Table 3.

For the BigGen Bench and HealthBench-Hard, we report the *macro-average score*, computed as the mean of per-example rubric scores $s_i = \frac{\texttt{total_awarded}_i}{\texttt{total_possible}_i}$ (converted to a percentage), so that *each dialogue/prompt contributes equally* regardless of how many rubric criteria it contains. For HealthBench, we additionally *clip the macro-average to* $[0, 100]$ after averaging, following the benchmark's recommended presentation to prevent rare negative-penalty cases from overly skewing the headline score.

Table 3: Shared RFT training hyperparameters. We use the `verl` library [45] for RFT.

Parameter	**Value**
Algorithm	Dr,GRPO
Rollouts per prompt	8
Learning rate	1×10^{-6}
Learning rate schedule	Constant with no warmup
Global batch size	128
Reward-level KL coefficient	1×10^{-3}
Max. training steps	1000
Max. gen. tokens	2048
Training GPUs	8$\times$ NVIDIA H100s
π_J	GPT-OSS-120B [1]
Optimizer	AdamW [33]
Parallelism strategy	FSDP [39]

Note: We use the PyTorch FSDP implementation; see `https://pytorch.org/docs/stable/fsdp.html`.

Table 4: Ablation studies for RRD on JudgeBench with GPT-4o as rubric proposer. (a) Ablation of the termination threshold (i.e., the number of rubric proposals rejected due to redundancy or misalignment) for the RRD process. (b) Ablation of the sample response generation strategy, comparing (i) a combination of strong frontier models, (ii) a single strong model, and (iii) a mixture of strong and weaker models as inputs to the rubric proposer.

	Termination Threshold			
Method	**5**	**10**	**15**	**20**
RRD	68.6	67.4	70.0	66.3
RRD_{LLM}	69.7	69.1	72.4	67.1
RRD_{WU}	71.4	72.0	73.3	73.1

(a)

	Sample Responses Generator			
Method	**GPT+Gemini**	**Gemini Only**	**GPT Only**	**GPT+Llama-8B**
RRD	70.0	66.6	57.4	54.3
RRD_{LLM}	72.4	71.0	61.4	60.0
RRD_{WU}	73.3	72.4	62.3	60.3

(b)

G Additional Results

G.1 Additional LLM Judge Ablations

Table 4 reports additional JudgeBench ablations over two implementation choices: the recursion termination threshold and the sample-response generator used to condition rubric proposal.

Table 4(a) reports RRD variants under different termination thresholds, where the recursion stops after a fixed number of rubric proposals are rejected as redundant or misaligned. We observe that lower thresholds (e.g., 5 or 10) underperform, likely due to insufficient exploration of the rubric space. Raising the threshold to 20 yields a plateau or slight drop for basic RRD and RRD_{LLM}, consistent with diminishing returns and increased noise from correlated criteria. In contrast, RRD_{WU} is notably robust: accuracy remains high even at threshold 20, suggesting that whitening-uniform weighting effectively counteracts correlation-induced noise and stabilizes performance under deeper rubric exploration.

In Table 4(b), we ablate the sample response generator by comparing multi-model ensembles with single-model and mixed-capability setups. Using two strong frontier models (GPT-4o + Gemini 2.5-Pro) performs best, consistently outperforming single-model baselines. Notably, when GPT-4o is the rubric proposer, it benefits more from conditioning on high-quality samples produced by a different frontier model (Gemini) than on its own outputs, suggesting that exposure to diverse reasoning styles and perspectives improves rubric generation. Meanwhile, replacing one frontier model with a smaller model (e.g., Llama3.1-8B) tends to reduce performance in RRD setting, likely because decomposition benefits from high-quality samples that reveal when a rubric is overly coarse (i.e., satisfied by multiple strong responses) and should be refined.

G.2 Ablation for Different Base Rubrics

We further conducted the experiment using different base rubric generation method. Specifically, we used RubricHub-generated base rubrics [29] and followed the same JudgeBench setup as in Fig. 3 (a).

Method Uniform LLM WU RRD 70.0 72.4 73.3 RubricHub-base + RRD decomposition 68.6 71.7 72.9 RRD remains effective when applied to RubricHub-generated base rubrics, achieving 72.9 with WU and continuing to outperform the prior baselines. These results support the generalisability of RRD across different automatic base rubric generation methods.

G.3 RFT Training Results

Details of RFT training results can be found in Table 5 and 6.

Table 5: Performance (in %) on BiGGen Bench. RRD_{WU} consistently improves performance across both Qwen3-4B and Llama3.1-8B backbones, yielding gains across evaluation axes as well as the overall score (standard deviations in the overall RRD_{WU} score across three runs for both base models are below ± 0.48).

Model	Method	IF	Ground.	Refine.	Plann.	ToM	Reason.	Tool	Safety	Overall
Qwen3	Base Model	76.4	85.6	79.7	84.0	84.4	86.0	60.8	75.9	77.9
	LLM Judge as Reward	79.0	87.0	78.7	82.2	85.4	85.2	65.7	75.4	78.5
	LLM Rubrics (w. resp.)	79.6	82.6	83.1	83.5	86.0	86.6	73.2	71.1	79.2
	Chasing the Tail	80.2	86.0	80.2	82.5	84.8	86.0	63.1	74.6	78.6
	Auto Rubrics	78.8	87.4	83.4	83.2	82.0	86.2	74.8	74.4	79.1
	InfiMed-ORBIT	79.4	87.0	81.4	81.8	87.6	86.2	72.6	76.8	80.7
	$\text{RRD}_{\text{uniform}}$	79.2	87.8	83.6	82.8	88.2	87.4	74.2	79.4	81.3
	RRD_{LLM}	80.2	87.2	83.1	84.3	85.7	86.0	74.2	76.7	80.6
	RRD_{WU}	**82.6**	**87.6**	**84.3**	**85.4**	**88.6**	**88.2**	**74.9**	**79.5**	**82.8**
Llama3.1	Base Model	69.4	76.2	67.3	69.7	71.6	64.0	61.1	70.1	67.0
	LLM Judge as Reward	73.1	74.8	67.1	69.1	70.0	61.4	55.1	70.8	66.3
	LLM Rubrics (w. resp.)	72.6	76.6	69.2	72.7	72.6	66.2	61.4	61.0	66.9
	Chasing the Tail	72.6	78.4	66.0	72.0	73.8	63.0	62.0	60.8	67.3
	Auto Rubrics	73.0	77.2	70.5	72.9	72.8	65.8	64.0	70.1	69.2
	InfiMed-ORBIT	68.6	78.0	67.6	65.7	71.4	66.0	63.1	65.6	66.8
	$\text{RRD}_{\text{uniform}}$	72.4	73.0	72.6	70.2	72.6	65.8	64.8	70.9	68.6
	RRD_{LLM}	73.4	78.6	65.3	72.9	74.8	64.2	63.1	67.1	68.0
	RRD_{WU}	**74.6**	**79.1**	**73.3**	**73.1**	**76.4**	**67.4**	**64.9**	**70.9**	**71.1**

Table 6: Performance (in %) on HealthBench-Hard. RRD_{WU} outperforms baselines across key axes – including instruction following (IF), accuracy, completeness, context, and the overall score (Qwen3-4B as backbone). It also achieves comparable performance on the communication axis relative to other RRD-variants. Similar performance gains are observed for Llama3.1-8B, demonstrating the robustness of RRD_{WU} across architectures (standard deviations in the overall RRD_{WU} score across three runs for both base models are below ± 0.57).

Model	Method	IF	Accuracy	Completeness	Communication	Context	Overall
Qwen3	Base Model	48.3	59.7	59.2	63.3	42.4	55.9
	LLM Judge as Reward	51.0	57.0	57.9	64.8	34.9	57.0
	LLM Rubrics (w. resp.)	49.0	60.8	63.3	65.7	40.6	59.3
	Chasing the Tail	48.8	60.9	55.5	64.8	39.7	58.7
	Auto Rubrics	51.2	42.1	44.8	51.7	23.1	39.2
	InfiMed-ORBIT	42.4	50.2	47.9	56.7	32.0	44.9
	$\text{RRD}_{\text{uniform}}$	49.5	63.1	54.9	68.6	37.7	60.1
	RRD_{LLM}	50.7	64.2	66.6	69.0	47.2	60.9
	RRD_{WU}	**64.3**	**65.2**	**71.7**	**67.7**	**47.4**	**61.7**
Llama3.1	Base Model	33.0	40.7	34.3	57.3	23.1	37.1
	LLM Judge as Reward	39.0	46.7	45.1	59.4	27.5	44.1
	LLM Rubrics (w. resp.)	36.2	49.6	55.2	53.6	35.1	47.2
	Chasing the Tail	36.2	48.9	55.3	56.0	38.6	47.5
	Auto Rubrics	39.0	21.8	18.7	55.4	28.7	20.7
	InfiMed-ORBIT	36.7	28.2	22.1	39.6	28.9	23.9
	$\text{RRD}_{\text{uniform}}$	36.3	49.3	56.5	55.7	37.7	47.6
	RRD_{LLM}	42.2	48.0	58.0	62.5	37.6	48.5
	RRD_{WU}	**41.8**	**52.6**	**58.2**	**62.3**	**39.9**	**49.3**

G.4 Human Evaluation

We conducted an additional blind human evaluation to test whether rubric-based judge outcomes align with human preference on 50 held-out WildChat prompts. For each prompt, we generated two responses from different model families, Llama and Qwen, and asked three independent annotators to select the better response. Annotators were not shown the rubrics, model identities, or judge outputs. To make the evaluation more informative, we included both 25 easier cases, where the two responses had a larger judge-score gap, and 25 harder cases, where the score gap was smaller. Using the human

majority vote as the reference, RRD agrees with human preferences on $44/50$ examples (88%) overall, compared with $33/50$ (66%) for the base judge. The improvement is consistent across both difficulty levels: on the easy subset, RRD achieves $24/25$ agreement (96%), compared with $19/25$ (76%) for the base judge; on the hard subset, RRD achieves $20/25$ agreement (80%), compared with $14/25$ (56%) for the base judge.

These results provide supplementary evidence that the gains from RRD are not merely due to optimization toward rubric artifacts. On fresh held-out prompts with blind human preferences, RRD tracks human choices substantially better than the base judge, especially on the harder subset where the response differences are more subtle. This conclusion is further supported by the transferability of RFT gains in our earlier experiments to external benchmarks with distinct evaluation schemes, including HealthBench, which uses physician-written criteria, and BiGGen Bench, which uses independently authored criteria. Together, these results suggest that RRD improves preference alignment in a way that transfers beyond the specific rubrics used during construction.

H Prompts

H.1 Rubric Generation

Prompt for Initial Rubric Generation.

Role: You are a rubric designer for an LLM-as-judge system.

Inputs you will receive:

- Prompt: the task/question the response must answer.
- *Responses: a set of responses to be evaluated against rubrics.*

Goal: Design a comprehensive set of rubrics for evaluating responses to the given prompt. Only write rubrics you are confident about. Only propose the best new rubrics.

Requirements:

- Propose rubrics that collectively cover the most important dimensions needed to judge whether a response correctly and helpfully satisfies the prompt.
- Each rubric must be consistently judgeable across many responses (avoid vague wording like "good", "nice", "high-quality").
- Each rubric must be prompt-specific (tied to what the user asked), not generic writing advice.
- Each rubric should be written as a single criterion with clear, binary pass/fail boundaries. Prefer objective checks.
- New rubric MUST NOT answer the question directly.
- *New rubric MUST NOT repeat any of the responses provided.*

Tips for writing good rubrics:
i. MECE:

- Mutually Exclusive, Collectively Exhaustive

ii. Completeness:

- Consider all the elements you would want to include to create a perfect response and put them into the rubric. This means including not only the facts and statements directly requested by the prompt, but also the supporting details that provide justification, reasoning, and logic for your response. Each of these elements should have a criterion because each criterion helps to develop the answer to the question from a slightly different angle.

iii. No overlapping:

- the same error from a model shouldn't be punished multiple times.

iv. Diversity:

- The rubric items should include variable types of information.
- If all criteria are like "the response mentions A", "the response mentions B", then this is not a good rubric.

v: How many rubric items for each prompt

- There is no golden standard, and the desired number of rubrics varies by accounts and task types.
- Write rubrics that cover all aspects of an ideal response.

vi: How many rubric items to fail

- A good rule of thumb is that the model fails on 50 per cent of rubrics items

vii: Atomicity / Non-stacked

- Each rubric criterion should evaluate exactly one distinct aspect. Avoid bundling multiple criteria into a single rubric. Most stacked criteria with the word "and" can be broken up into multiple pieces.
- BAD: Response identifies George Washington as the first U.S. president and mentions he served two terms.
- GOOD: Response identifies George Washington as the first U.S. president.
- GOOD: Response mentions that George Washington served two terms.

viii: Specificity

- Criteria should be binary (true or false) and objective.
- Avoid vague descriptions (e.g., "the response must be accurate" is vague).
- Example: "The response should list exactly three examples."

ix: Self-contained

- Each criterion should contain all the information needed to evaluate a response. For example: "Mentions the capital city of Canada" $\rightarrow$ BAD; "Mentions the capital city of Canada is Ottawa" $\rightarrow$ GOOD.

x: Criterion should be verifiable without requiring external search.

- BAD: Response names any of the Nobel Prize winners in Physics in 2023
- GOOD: Response names any of the following Nobel Prize winners in Physics in 2023: Pierre Agostini, Ferenc Krausz, or Anne L'Huillier.

Below are the inputs:

- `Here is the user prompt for which we want to generate a rubric:`
  `PROMPT: {prompt}`
- *`Here are a list of reference responses:`*
  `RESPONSES: {responses}`

Output STRICTLY in below format. No other text is allowed:

- `<RUBRIC> Rubric 1 </RUBRIC>`
- `<RUBRIC> Rubric 2 </RUBRIC>`
- `...`

Note: *Italic* parts are only included when generating rubrics with sampled responses.

Prompt for Rubric Decomposition.

Role: You are a rubric designer for an LLM-as-judge system.

Inputs you will receive:

- Prompt: the task/question the response must answer.
- Responses: a set of responses to be evaluated against rubrics.
- Current rubric: criterion currently used by a judge. This rubric has already been satisfied by multiple responses, so it is too coarse and fail to distinguish response quality.
- Other rubrics: Other rubrics that the new rubric must NOT overlap with.

Goal: Propose exactly TWO new rubrics that are more granular than the existing ones and can better differentiate candidate responses. Only write rubrics you are confident about. Only propose the best new rubrics.

What "more granular" means (requirements):

- Each new rubric must target a specific, discriminative dimension of quality that is not sufficiently captured by the existing rubrics (e.g., completeness of key sub-steps, correctness of constraints, justification quality, handling of edge cases, faithfulness to prompt format, etc.).
- New rubrics should NOT miss critical information contained in the existing rubric.
- Each rubric must be consistently judgeable across many responses (avoid vague wording like "good", "nice", "high-quality").
- Each rubric must be prompt-specific (tied to what the user asked), not generic writing advice.
- Each rubric should be written as a single criterion with clear, binary pass/fail boundaries. Prefer objective checks.
- New rubric MUST NOT repeat any of the responses provided.
- New rubric MUST NOT answer the question directly.

Tips for writing good rubrics:
The same as those in the prompt for initial rubric generation. These have been omitted here for brevity.

Below are the inputs:

- `Here is the user prompt for which we want to generate a rubric:]`
  `PROMPT: {prompt}`
- `Here are a list of reference responses:`
  `RESPONSES: {responses}`
- `Here is existing rubric to improve:`
  `RUBRIC: {rubric}`
- `Here is other rubric that the new rubric should NOT overlap with:`
  `OTHER RUBRIC: {other_rubric}`

Output STRICTLY in below format. No other text is allowed:

- `<RUBRIC> New rubric 1 </RUBRIC>`
- `<RUBRIC> New rubric 2 </RUBRIC>`

H.2 Filtering

Prompt for Duplication Filter.

You are checking whether a new rubric substantially overlaps with ANY of the existing rubrics. If ANY overlap is found, output YES; otherwise output NO.

Definition of substantial overlapping:

- The new rubric has the same intent as an existing rubric, or is a strict subset/superset of it, or $\geq 70\%$ of its meaning is covered by the existing rubric so that applying both would not materially change scoring outcomes.
- Match on meaning, not wording. Treat synonyms, paraphrases, and inverses with the same effect as overlapping (e.g., "be concise" $\approx$ "avoid unnecessary verbosity").
- Ignore trivial phrasing, tone, and example differences unless they change the requirement.

Edge cases:

- If scopes are disjoint (different capability/goal) → NO.
- If the new rubric adds only minor qualifiers (e.g., "clearly"/"appropriately") without changing evaluation → YES.
- If the new rubric merely narrows the context while keeping the same criterion (subset) or broadens it (superset) → YES.

Input format:

- `EXISTING_RUBRICS: {existing_rubrics}`
- `NEW_RUBRIC: {new_rubric}`

Output STRICTLY in below format. No other text is allowed:

- `<EVALUATION> YES/NO </EVALUATION>`

Prompt for Conflict Filter.

You are checking whether a new rubric expresses opposite meaning of ANY of the existing rubrics. If ANY opposition is found, output YES; otherwise output NO.

Definition of opposition:

- "Opposite" means the new rubric asserts the negation or reverse polarity of the same requirement, property, or direction as an existing rubric.
- Examples:
  - require X ↔ forbid/avoid X
  - must include X ↔ must NOT include X
  - prefer more of X ↔ prefer less of X (same X, opposite direction)
  - answer should be optimistic ↔ answer should be pessimistic
- Do NOT flag different targets or contexts.
- Do NOT flag orthogonal dimensions (e.g., tone vs citations or "be clear" vs "be concise")
- Do NOT flag mere differences in emphasis, strength, scope, or style.
- Do NOT flag stricter/looser thresholds unless they clearly reverse direction on the same axis (e.g., "maximize brevity" vs "maximize elaboration" = opposite; "$\leq$ 120 words" vs "$\leq$ 150 words" = NOT opposite).

Input format:

- `EXISTING_RUBRICS: {existing_rubrics}`
- `NEW_RUBRIC: {new_rubric}`

Output STRICTLY in below format. No other text is allowed:

- `<EVALUATION> YES/NO </EVALUATION>`

H.3 Evaluation

Judge prompt for rubric-based response evaluation.

You are a judge, evaluating whether a response satisfies the given rubric. If the response satisfies the criterion of the rubric, output YES; otherwise output NO.

Requirement:

- You must follow the rubric strictly, and only consider the criteria listed in the rubric.
- You must NOT consider any other factors, such as your own opinions or external knowledge.

Below between <RESPONSE> and </RESPONSE> is the response to evaluate on

- <RESPONSE> {response} </RESPONSE>

Below between <RUBRIC> and </RUBRIC> is the rubric to evaluate on

- <RUBRIC> {rubric} </RUBRIC>

Output STRICTLY in below format. No other text is allowed:

- <EVALUATION> YES/NO </EVALUATION>

I Qualitative Examples

In this section, we present representative rubric outputs from multiple datasets. For each example, we show both the initially proposed rubrics and the corresponding refined rubrics produced post RRD. We then present covariance-spectrum and effective-rank analyses on representative rubric sets, which support the need to downweight highly correlated rubrics for final aggregation.

When reading below examples, it is also worth noting that the duplication filter is designed to remove genuinely redundant criteria, but not to delete rubrics that appear similar while capturing distinct evaluative subcriteria. For example, in the website-speed-optimization email-writing example, rubrics 21 and 22 may appear related, but they capture different aspects of the sales argument: one concerns indirect growth through improved visibility, while the other concerns direct traffic and sales growth. Similarly, rubrics 24 and 25 distinguish contextual understanding and domain expertise from the clarity of the value proposition. Retaining such rubrics is therefore not a failure of duplicate removal, but an intended feature of finer-grained rubric construction: it allows the evaluator to distinguish among responses along more specific dimensions. This behavior is also supported by more constrained tasks. In the math-reasoning example below, RRD produces only 9 rubrics, substantially fewer than in highly subjective open-ended settings. This suggests that rubric complexity adapts to the structure and openness of the task, rather than expanding uniformly or indiscriminately across all prompts.

Example from JudgeBench Dataset

Prompt:
A college student initiated a criminal case against his former roommate, claiming the former roommate snuck in and removed his state-of-the-art home theater equipment from his house. The college student took the complaint papers that were approved by the prosecution and filed them with a magistrate court. An ongoing police investigation later determined, however, that the college student was a vengeful person not averse to creating stories, and that he and the former roommate had been arguing over several related issues. They also determined that he never had any quality sound or video equipment. The police dropped the prosecution, and the criminal case was dismissed at the preliminary hearing. When the former roommate filed a civil case against the college student alleging the tort of malicious prosecution, will evidence of the student's reputation for dishonesty be admissible and why?

- (A) Yes, because the student's reputation for dishonesty directly relates to the tort of malicious prosecution.
- (B) No, because reputation evidence is considered too prejudicial to be admissible.
- (C) Yes, because dishonesty at any time in the past is always relevant to a malicious prosecution complaint.
- (D) Yes, because the student's dishonesty directly affects the credibility of his initial criminal complaint.
- (E) Yes, because reputation evidence, just like most character evidence, is virtually always admissible.
- (F) Yes, because even though character evidence is too collateral to be admitted for circumstantial evidence, it is admissible if it is directly at issue in the case.
- (G) No, because character evidence is only admissible if it is directly related to the act in question.
- (H) No, because the past reputation of a person is not considered as evidence in court.
- (I) Yes, because while most character evidence is not admissible, some is, and reputation evidence is always admissible.
- (J) No, because character evidence is never admissible in any case.

If you cannot determine the correct multiple-choice answer, take your best guess. Once you have your answer, please duplicate that letter five times in a single string. For example, if the answer is K, then write KKKKK. Let's think step by step.

Initial rubrics:

1. Correctly identifies that the admissibility of reputation evidence depends on its relevance to an element of the tort of malicious prosecution.
2. Distinguishes between general character evidence and reputation evidence admissible when character is directly at issue.
3. Correctly considers the legal elements required to prove malicious prosecution.
4. Accurately addresses whether the student's reputation for dishonesty is being used for propensity reasoning or as direct evidence of an element in the case.
5. Avoids overgeneralization regarding the admissibility of reputation or character evidence.
6. Acknowledges that rules of evidence may treat civil and criminal cases differently regarding character evidence.

Final rubrics:

1. Clearly identifies the key elements required to prove malicious prosecution: institution of criminal proceedings, termination in plaintiff's favor, absence of probable cause, and malice.
2. Correctly articulates the general rule that character evidence is inadmissible to prove a person acted in conformity with that character on a particular occasion, known as the propensity rule.
3. Accurately identifies the exception that character evidence is admissible when character is an essential element of a claim or defense, known as character being directly at issue.
4. Explains how the student's reputation for dishonesty directly relates to proving lack of probable cause and malice in a malicious prosecution case.
5. Explicitly mentions the role of the student's reputation for dishonesty in determining whether the student had an honest belief that a crime occurred.
6. Assesses the relevance of character evidence by determining whether it significantly impacts a central factual question or element in the case, beyond mere propensity.
7. Assesses the impact of admitting reputation evidence on the balance of prejudice versus probative value, specifically in relation to a malicious prosecution claim.
8. Assesses the contextual relevance and potential prejudicial impact of introducing reputation evidence of dishonesty, weighing its probative value against its potential to unfairly bias the court.
9. Explores the connection between the student's reputation and the historical pattern of behavior, evaluating whether this pattern offers insight into a motive or strategy crucial for the malicious prosecution claim, without relying on the propensity argument.
10. Clearly distinguishes between the legal concept of reputation and character as they apply to the admissibility of evidence, and explains any differences in their impact on the case.

Example from PPE Preference Dataset

Prompt:
Write an email for selling my website speed optimization services to e-commerce business owners. Make it at short as possible. Focus in benefits (revenue or money) that i make for them.

Initial rubrics:

1. Mentions website speed optimization as the core service offered.

2. Specifically targets e-commerce business owners as the intended audience.
3. Includes at least one concrete business benefit tied to money or revenue.
4. Maintains a short and concise format (no more than 100 words).
5. Avoids technical jargon or implementation details unrelated to benefits.
6. Includes a clear call to action (e.g., reply, book a call, request audit).
7. Avoids generic claims unrelated to website performance or monetary benefit.

Final rubrics:

1. Mentions that faster loading times lead to higher sales or increased revenue.
2. Advises that every second of delay can reduce sales or mentions specific loss metrics.
3. Offers a specific service aimed at optimizing website speed for e-commerce.
4. Includes a call to action inviting the recipient for further engagement, such as a chat or a video.
5. Uses an analogy or relatable scenario to illustrate the impact of website speed on sales or customer retention.
6. Provides an option to receive additional information or a report, implying a no-obligation offer.
7. References an impact of speed on conversion rates, either generally or with specific percentages.
8. Utilizes a subject line that directly mentions website speed and its impact on sales/revenue.
9. Highlights specific benefits for customer satisfaction or user experience beyond direct revenue increase, such as improved browsing experience or reduced bounce rates.
10. Craft a customizable template by including placeholders for recipient's name and company name.
11. Establishes credibility by referencing a well-known company or study related to website optimization and its impact on revenue.
12. Utilizes persuasive language that emotionally appeals to the business owner's pain points, like frustration from slow checkout or fear of losing customers, to emphasize urgency and drive interest in services offered.
13. Incorporates a sense of urgency or a compelling reason to act quickly within the email message.
14. Emphasizes the competitive advantage gained by having a faster-loading website compared to other e-commerce businesses.
15. Initiates engagement by offering a free initial assessment or audit of the recipient's current website performance to demonstrate potential improvements.
16. Explicitly quantifies potential revenue increase with specific dollar amounts, contrasting current losses due to slow website speed with projected gains post-optimization.
17. Incorporates industry-specific examples or case studies to demonstrate proven results of website speed optimization in increasing e-commerce sales.
18. Clearly differentiates the email by employing a visually unique format or structure that stands out, such as bullet points, bold headings, or highlighting key revenue benefits, to enhance readability and impact.
19. Measure the concise and efficient use of language in the email, aiming to keep the message succinct yet informative, meeting the prompt's requirement for brevity.
20. Mentions strategies or tools used in the website speed optimization, detailing specific technologies or methodologies applied to achieve faster loading times, such as caching, image optimization, or content delivery networks.

21. Specifies how increased website speed can amplify the business's online visibility or improve SEO rankings, thus contributing indirectly to revenue growth.
22. Informs the recipient that an optimized website speed can lead to improved search engine rankings, thereby attracting more organic traffic and increasing potential sales.
23. Demonstrates insight into user behavior changes stemming from improved website speed, such as increased time spent on site or higher interaction rates.
24. Incorporates a personalization element that leverages the recipient's specific industry or e-commerce niche, showcasing tailored expertise in optimizing website speed for their unique business needs.
25. Evaluates whether the email offers a personalized benefit or advantage uniquely tailored to the recipient's business or e-commerce niche, emphasizing customization over generic benefits.
26. Specifies how optimized website speed can enhance mobile user experiences, which is crucial for e-commerce businesses to capture a significant portion of their audience.
27. Presents a scalable solution that can easily integrate into the recipient's existing e-commerce infrastructure, emphasizing its convenience and low disruption to current operations.
28. Clearly outlines the process or specific steps involved in the website speed optimization service to give the reader insight into how the transformation will be achieved.
29. The email creatively uses a metaphor or storytelling that evokes a vivid image to emphasize the importance of website speed optimization, enhancing the overall persuasive impact.
30. Ensures the email utilizes statistical evidence or factual data, such as industry benchmarks, to validate the benefits of speed optimization comprehensively, enhancing the persuasive power through quantifiable credibility.

Example of a simple task from PPE Preference Dataset

Prompt:
There are 10 apples today. 1 was eaten yesterday. How many are left now? *(Original prompt in Chinese – translated to English for illustration)*

Initial rubrics:

1. States that there are 10 apples remaining today, regardless of yesterday's consumption.
2. Explains that the apples consumed yesterday do not affect today's apple count of 10.
3. Acknowledges that the count of apples as of today is based on the statement "There are 10 apples today" without reference to past events.
4. Provides a rationale or explanation for considering the past apple consumption as irrelevant to the current count.
5. Avoids making quantitative calculations that incorrectly deduct apples from today's count based on past consumption.
6. Does not ambiguously imply there could be fewer or more than 10 apples today without any additional information provided in the prompt.

Final rubrics:

1. (Same as the initial rubrics – omit for brevity.)
2. Provides a creative interpretation of the scenario, incorporating narrative or imaginative elements that explain the unchanged apple count.

3. Uses narrative or imaginative storytelling elements without deviating from the stated fact that today's count is 10 apples, enhancing engagement while maintaining accuracy.

Example of WildChat Dataset

Example 1

Prompt:
What influence did Greek thought have on the emergence and development of science?

Initial rubrics:

1. Identifies Greek thought as a historical antecedent rather than a modern scientific contributor.
2. Describes at least one conceptual contribution of Greek thought to the development of scientific reasoning.
3. Mentions at least one historical figure or school of Greek thought in a scientific or proto-scientific context.
4. Acknowledges the role of Greek thought in shaping methodological or epistemological foundations of science.
5. Avoids attributing anachronistic concepts of modern science (e.g., peer review, experimental control) directly to the Greeks.
6. Maintains focus on Greek influence specifically, not general ancient civilizations.
7. Frames the influence of Greek thought in terms of its impact on later scientific development (e.g., during the Islamic Golden Age, Renaissance, Enlightenment).

Final rubrics:

1. The response mentions Thales of Miletus predicting an eclipse as an example of early scientific reasoning by the Greeks.
2. The response includes an explanation of the Greek concept of 'Kosmos' as an ordered, harmonious system significant to scientific thought.
3. The response describes the influence of Pythagorean mathematics on the development of scientific principles.
4. The response discusses Socratic questioning or the Socratic method as essential for scientific inquiry.
5. The response acknowledges errors in Greek science but emphasizes their foundational contributions to scientific development.
6. The response cites Euclid's "Elements" and its impact on geometry as a part of scientific advancement.
7. The response mentions the Greek transition from mythological explanations to logical reasoning in understanding the universe.
8. The response includes reference to Greek thinkers' belief in a rational, intelligible universe governed by laws.
9. The response discusses Aristotle's role in the systematization and classification of the natural world.
10. The response explores Greek contributions to observational and empirical methods, such as Hippocrates' emphasis on documenting symptoms, in the development of scientific methodology.

11. The response addresses the synthesis of empirical observation and logical reasoning as a Greek methodological advancement in the pursuit of understanding the natural world.
12. The response identifies the impact of Greek dialectic methods on shaping the peer review process used in modern science.
13. The response explains the role of Greek mathematics, particularly its use in describing physical phenomena, in influencing later scientific approaches to the natural world.
14. The response illustrates how Greek thought contributed to the formation of a foundational ethical framework for science, highlighting examples like the Hippocratic Oath.
15. The response explores the role of Archimedes in advancing principles of physics and engineering, illustrating the practical application of Greek thought to material reality.
16. The response illustrates how Greek philosophical discourse in public spaces, such as the Agora, contributed to the communal and social nature of scientific inquiry.
17. The response explores the role of Greek philosophical schools, such as the Atomists or Stoics, in shaping fundamental principles that influenced later scientific exploration.
18. The response discusses the role of Greek thought in shaping the dual approach of both theoretical abstraction and practical experimentation which became fundamental to later scientific advancements, despite the Greeks not fully implementing experimentation themselves.
19. The response showcases how Greek thought laid the groundwork for interdisciplinary thinking, mixing philosophy, mathematics, and early natural sciences to broaden approaches in scientific exploration.
20. The response highlights the symbolic aspect of Greek thought as initiating a "cathedral of knowledge," metaphorically illustrating the foundational impact on science across millennia.
21. The response highlights the impact of Greek thought in establishing a systematic methodology for distinguishing between subjective beliefs and objective facts, facilitating a structured framework for scientific inquiry.
22. The response explores the lasting cultural and intellectual impact of Greek thought on the Renaissance and subsequent scientific revolution, highlighting the transfer and transformation of Greek ideas across different historical periods.
23. The response explains how Greek thought established an early conceptual framework for distinguishing between qualitative and quantitative analysis in scientific inquiry.
24. The response explores how Greek advancements in abstract thought, particularly through Plato's ideal forms, contributed to the long-term development of scientific concepts by encouraging the pursuit of universal principles and truths.
25. The response highlights the contribution of Greek exploration of metaphysics and ontology—concepts of being and existence—and their impact on shaping scientific inquiry into the nature of reality.
26. The response outlines the Greek philosophical commitment to deductive reasoning as a critical factor in establishing structured scientific theories.

Example 2

Prompt:
What skill can I learn in six months and generate a lot of money?

Initial rubrics:

1. Proposes a skill that is realistically learnable within approximately six months.
2. Proposes a skill that has clear, established paths to monetization.

3. Clearly identifies the skill being recommended.
4. Avoids unrealistic earnings claims (e.g., “become a millionaire instantly”) or unverifiable hype.
5. Connects the proposed skill to a plausible use case or real-world application.
6. Maintains relevance to both time constraint (six months) and financial goal (generating a lot of money).

Final rubrics:

1. Identifies and recommends a specific skill that can be learned in six months and is capable of generating substantial income.
2. Provides justification for why the skill chosen is lucrative, referring to current demand and market trends.
3. Lists actionable steps or a clear plan on how to learn the skill within the six-month time frame.
4. Includes examples of tools or platforms necessary to master the skill.
5. Discusses potential income or earning opportunities related to the skill.
6. Addresses the need for the skill to solve specific business problems or provide substantial value to clients.
7. Suggests ways to monetize the skill, such as freelance opportunities, consulting, or creating digital products.
8. Emphasizes the importance of differentiating oneself with complementary skills or additional strategies beyond the primary skill.
9. Mentions the importance of building a portfolio or case studies to showcase expertise in the chosen skill.
10. Encourages pursuing online courses, certification, or self-directed learning as part of the skill acquisition process.
11. Evaluates the long-term sustainability and adaptability of the skill in relation to evolving market conditions and potential technological advancements.
12. Evaluates the transformation of the learned skill into a clearly defined business model, detailing how the skill is integrated into a sustainable, repeatable, and scalable business process.
13. Evaluates the practicality of transitioning from learning the skill to implementing it in real-world scenarios within the six-month time frame.
14. Evaluates the potential scalability or expansion of the skill into a broader business or service offering in the future, beyond just immediate income generation.
15. Evaluates the effectiveness of the skill in addressing industry-specific challenges, detailing how it creates a competitive advantage or fills a gap in the current market.
16. Evaluates the creativity and uniqueness of the proposed learning path by assessing whether the response provides a vision or framework for integrating the skill into the user’s broader personal or professional development roadmap.
17. Evaluates potential challenges or pitfalls in learning the chosen skill and suggests methods to overcome them.
18. Evaluates the clarity and realism of the proposed six-month learning timeline in the context of the user’s current knowledge and experience level, assessing whether the response tailors the learning schedule to various starting points of potential learners.
19. Assesses whether the response includes a strategy for gaining the first practical experience or opportunities (e.g., internships, volunteer projects, or initial clients) in the chosen skill to build credibility and real-world expertise.

20. Examines the feasibility and validity of the skill learning timeline by analyzing the skill's complexity, required foundational knowledge, and intensity of the proposed learning schedule.
21. Assesses the completeness and coherence of the narrative or framework describing how the proposed skill will lead to financial success, ensuring it logically connects individual learning steps to the eventual income objectives.
22. Assesses the ability of the response to convey the transformative impact of the skill on both personal growth and career trajectory, highlighting the broader implications of mastering the skill beyond immediate financial gain.
23. Evaluates the response's inclusion of strategies to enhance personal branding related to the skill, including creating an online presence or sharing work publicly to attract potential clients or employers.
24. Assesses whether the response effectively illustrates a unique marketing or networking strategy to connect with potential clients or build a professional reputation in the first stages of the skill application.
25. Emphasizes the need to practice and iterate on real-world projects or scenarios relevant to the chosen skill, demonstrating iterative improvement and application of feedback.
26. Assesses the response for providing specific examples of successful individuals or case studies who have effectively learned and monetized the skill within a similar timeframe, showcasing real-world applicability and success stories.
27. Assesses the depth of detail and specificity in outlining how the proposed skill integrates with existing competencies or knowledge, illustrating how this integration enhances the potential for success and financial gain.
28. Assesses the degree to which the response incorporates practical examples or case studies demonstrating the real-world application and impact of the skill learned, highlighting any transformations in business operations or financial performance.
29. Assesses the alignment of the skill with the user's current abilities and interests, ensuring that the skill is not only compatible with but also builds upon their existing strengths, knowledge, or passions to maximize successful implementation and retention.
30. Assesses the ability to effectively communicate and articulate the value proposition of the learned skill to prospective clients or employers, ensuring that the pitch is tailored to their specific needs and objectives.
31. Assesses the response's ability to identify opportunities for networking or collaboration within the industry relevant to the chosen skill, and how this could accelerate learning and increase potential income.

Example 3

Prompt:
Write me an essay about monkeys.

Initial rubrics:

1. The response is structured in essay format with an introduction, body, and conclusion.
2. The essay remains focused on the topic of monkeys throughout the response.
3. The response refers specifically to monkeys and does not conflate them with apes or other primates.
4. Includes at least one factual detail about monkeys (e.g., diet, habitat, species, behavior).
5. Avoids listing-only or bullet-point format; presented as continuous prose.
6. Does not consist entirely of generic filler or vague commentary.

7. Maintains a neutral or informative tone appropriate for an essay.

Final rubrics:

1. Includes an exploration of monkeys' physical abilities, such as their agility and acrobatics.
2. Mentions the social structures and behaviors observed in monkey societies.
3. Describes the intelligence and problem-solving abilities of monkeys.
4. Examines the evolutionary connection between monkeys and humans.
5. Discusses the ecological role of monkeys, such as seed dispersers.
6. Provides factual information about the diversity of monkey species.
7. Reflects on the emotional experiences shared between humans and monkeys.
8. Utilizes symbolic language to enhance understanding of monkeys beyond biological terms.
9. Considers the environmental habitat diversity in which different monkey species thrive, including their geographical distribution across continents.
10. Compares and contrasts different monkey species to highlight their unique characteristics and ecological niches.
11. Highlights the role of monkeys in scientific research and their contributions to advancements in our understanding of human cognition and disease.
12. Explains the impact of human activity, such as the illegal wildlife trade, on monkey populations.
13. Considers the historical depiction of monkeys in literature and folklore across different civilizations.
14. Discusses the use of metaphorical elements to draw parallels between monkey behavior and broader existential themes or philosophical questions.
15. Discusses the sensory experiences and communication methods of monkeys, such as vocalizations, facial expressions, and tactile behaviors, and their significance in their daily life and interactions.
16. Analyzes the role of monkeys in fostering cross-cultural connections and exchanges, highlighting examples where different cultures have interacted with or been influenced by monkey-related myths, symbols, or depictions.
17. Evaluates the symbolic representation of monkeys as tricksters or wise figures in folklore and what this reveals about human self-perception and cultural values.
18. Analyzes the role of monkeys in maintaining ecological balance and biodiversity through their interactions with other species.
19. Critically analyzes the ethical considerations and responsibilities associated with human interactions and interventions with monkey species, including the implications of research and conservation efforts.
20. Analyzes the role of monkeys in facilitating biodiversity through their interactions with other species in their ecosystems.
21. Discusses the ethical considerations in scientific research involving monkeys, including the balance between scientific advancement and animal welfare.
22. Analyzes the ethical considerations and moral obligations involved in the conservation of monkey habitats and populations.
23. Presents a nuanced exploration of monkeys' interactions with their environment, focusing on their role as adaptors within ecosystems and their ability to thrive across varying environmental conditions.
24. Assesses the depiction of unique communication methods and the role they play in both cooperative interactions and competitive scenarios within monkey societies.
25. Assesses the aesthetic portrayal of monkeys in the essay, considering the use of descriptive imagery to evoke sensory experiences and emotional responses in the reader.

26. Examines the role of monkeys in inspiring artistic and creative expressions, such as in visual arts, music, or dance, and how these reflect human interpretations of monkey characteristics and significance.
27. Explores the challenges related to the social dynamics within monkey populations when faced with environmental pressures or changes, such as altered food availability or increased competition for resources.
28. Discusses the role of monkeys in their local food webs and their interactions with other species as both prey and predators.
29. Assesses the anthropomorphism of monkey behaviors and characteristics, examining how human-like traits are attributed to monkeys in the narrative.
30. Evaluates the physiological adaptations of monkeys that enable them to thrive in varied environmental conditions, such as skin type variations and dietary specializations.
31. Analyzes the adaptation mechanisms monkeys have developed to thrive in varying environmental pressures and challenges, such as climate variations and human-induced habitat changes.
32. Discusses the role of monkeys in influencing art forms, such as visual art, music, and dance, and how these representations impact cultural perception and creative expression.
33. Explores the role of monkeys in allegorical storytelling, examining how they serve as vehicles for moral lessons and social commentary in diverse cultural narratives.

Covariance-spectrum and effective-rank analyses. Additionally, we provide a direct quantitative diagnostic on a representative example, we report the covariance spectrum and effective rank of the rubric set. For the mild brainteaser example from the PPE preference dataset, which asks about 10 apples, the covariance spectrum is $[0.227, 0.165, 0.079, 0.069, 0.041, 0.041, 0.041, 0.041]$, yielding an effective rank of 6.2 out of 8. This indicates that even a small rubric set exhibits meaningful correlation structure that should be accounted for during aggregation. Under uniform weighting, each rubric receives weight 0.125; after WU, the weights are redistributed as shown in Table 7.

Table 7: Whitened weights assigned to each rubric dimension.

	R1	R2	R3	R4	R5	R6	R7	R8
WU weight	0.159	0.159	0.095	0.087	0.159	0.064	0.159	0.117

This provides direct intrinsic evidence that WU behaves as intended. Rather than weighting all rubrics equally, it shifts weight away from correlated rubric directions: several overlapping rubrics, mainly those evaluating nuanced aspects of explanation quality, receive smaller weights than under uniform weighting, while more distinct checks, such as stating the correct number, avoiding the core error, and providing a creative interpretation, receive relatively larger weights. At the same time, no eigenvalue in the covariance spectrum is zero, indicating that the rubric set is not exactly redundant: no rubric direction is a perfect duplicate of another under the observed scores.

We observe the same pattern in the PPE email-writing example. Its covariance spectrum is also non-degenerate and yields an effective rank of 20.7 out of 30, again indicating meaningful but incomplete redundancy among rubrics and motivating correlation-aware aggregation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The key claims of the paper mentioned int he abstract and the introduction are accurately reflected in the rest of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation is discussed in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Theoretical assumption and proofs can be found in Section 2.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setups can be found in Section 3, Section 4 and Appendix F.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and pretrained models used in the experiments are publicly available. We have provided complete citations for the public datasets and also describe the model fine-tuning and experiments hyperparameters settings in detail in Appendix F.2.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental setups can be found in the first part of Section 3 and Section 4, and the additional training hyperparameters can be found in Appendix F.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We have run all experiments repeatably for three times. Due to the limited space, we didn't include the error bar in the table.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: We have describe the resource in Table 2 and Table 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Discussed in Appendix B.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer:[Yes]

    Justification: Discussed in Appendix B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not intend to release any of our fine-tuned models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the dataset and pre-trained models used in the paper have been clearly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.