FROM STRUCTURE TO VALIDITY: MECHANISM-GUIDED GENERATION AND REFINEMENT OF LLM RUBRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rubrics have been widely used to evaluate and optimize language models on open-ended tasks. However, existing work primarily focuses on generating better rubrics, with limited understanding of what makes rubric design effective. In this work, we conduct a systematic mechanistic analysis of rubric design across criterion organization, criterion operationalization, and score aggregation. Our analysis shows that structural organization enables more comprehensive criterion construction, semantic specification alone does not ensure effective criterion-level judgments, and sophisticated score aggregation provides limited additional gains. Guided by these findings, we introduce Adaptive Rubric Tree Search with Hybrid refinement (ARTS-Hybrid), which combines adaptive tree construction with semantic-first, response-calibrated refinement. Experiments across benchmarks show that ARTS-Hybrid improves both end-to-end evaluation performance and intrinsic rubric quality. Further experiments suggest that, beyond evaluation, rubrics generated by ARTS-Hybrid provide effective signals for model optimization.

1 INTRODUCTION

Large language models (LLMs) have made remarkable progress in recent years and are increasingly applied to open-ended tasks (Shi et al., 2025). Unlike tasks with verifiable outcomes, such as mathematical reasoning and code generation, open-ended tasks often admit multiple valid responses and lack objective ground-truth signals. Consequently, reliable evaluation has become a critical bottleneck in the development of LLMs. LLM-as-a-Judge (Gu et al., 2026; Zheng et al., 2023) has emerged as a scalable alternative to costly human evaluation, with many existing methods relying on a single holistic score to assess response quality. However, this scalar evaluation collapses multiple dimensions of response quality, such as correctness, completeness, and relevance, into a coarse-grained judgment, making it difficult to provide fine-grained feedback for model improvement.

Inspired by the long-standing use of rubrics in educational assessment (Brookhart, 2018), rubric-based evaluation offers a structured alternative by decomposing overall response quality into explicit task-specific criteria. In the context of LLMs, rubrics are increasingly becoming an important tool for model evaluation (Hashemi et al., 2024) and optimization (Gunjal et al., 2026). Early rubric construction methods primarily rely on language models to generate evaluation criteria directly from task instructions (Viswanathan et al., 2026; Lin et al., 2025), often resulting in incomplete coverage and inconsistent levels of granularity. Subsequent contrastive methods derive more discriminative evaluation criteria by identifying differences between preferred and rejected responses (Liu et al., 2026b; Tan et al., 2025; Xie et al., 2025). More recent work has introduced structural decomposition (Gao et al., 2026; Shen et al., 2026) and iterative refinement (Zhang et al., 2026; Li et al., 2026) to further improve the coverage and reliability of the rubric. In parallel, online approaches dynamically update rubrics based on policy model rollouts (Sheng et al., 2026; Shao et al., 2025), allowing evaluation criteria to evolve with the capabilities of the policy model. However, whether rubrics are constructed offline or dynamically evolved with the policy model, existing methods typically rely on aggregate outcomes, such as preference accuracy or downstream rewards, to guide their optimization. Although these metrics quantify end-to-end performance on a given dataset, the mechanisms underlying rubric effectiveness remain largely unexplained. The effectiveness of a rubric is influenced by multiple interdependent factors, such as how the criteria are constructed and organized.

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

Consequently, the observed gains cannot be reliably attributed to specific design choices, providing limited systematic guidance on how rubrics should be structured or refined. This highlights the need for a systematic mechanistic analysis to disentangle the effects of individual design choices.

In this paper, we move beyond end-to-end comparisons and dissect rubric effectiveness along three stages, including how evaluation criteria are structurally organized, how individual criteria are operationalized, and how criterion-level scores are aggregated. Through systematic mechanistic analyses, we uncover three main findings. First, the structural decomposition improves criterion coverage and supports more targeted diagnosis. Second, the semantically sound evaluation criteria do not necessarily yield discriminative and stable judgments, motivating the use of response-based calibration evidence. Third, with consistent criterion content and criterion-level judgments across settings, the alternative score aggregation strategies provide only limited additional gains.

Guided by these findings, we introduce **A**daptive **R**ubric **T**ree **S**earch with **Hybrid** refinement (**ARTS-Hybrid**), a mechanism-guided framework that formulates rubric construction as a constrained search over hierarchical criterion structures. Specifically, ARTS-Hybrid constructs rubrics through structured decomposition, where each expansion follows a semantic-first, response-calibrated selection process to identify the child criteria. We evaluate ARTS-Hybrid in terms of end-to-end evaluation performance and intrinsic rubric quality. In addition, we examine whether its benefits can be transferred to a standalone rubric generator and whether the generated rubrics provide effective reward signals for policy optimization. Our contributions are summarized as follows:

- **A systematic mechanistic analysis of rubric effectiveness.** We conduct a systematic analysis of rubric design by disentangling three interdependent factors, including criterion organization, criterion operationalization, and score aggregation.
- **A mechanism-guided generation and refinement framework.** We translate these findings into ARTS-Hybrid, which integrates structured decomposition with semantic-first, response-calibrated refinement. Experimental results show that ARTS-Hybrid achieves competitive performance across benchmarks while producing high-quality rubrics.
- **Validation beyond rubric-based evaluation.** We further show that the benefits of ARTS-Hybrid can be transferred to a standalone rubric generator and the generated rubrics provide effective reward signals for policy optimization.

2 Related Work

Automatic rubric generation and refinement. Existing approaches to automatic rubric construction can be broadly categorized into three paradigms according to how rubrics are updated: one-shot generation, offline iterative refinement, and online co-evolution. One-shot methods generate evaluation criteria directly from each task instruction (Viswanathan et al., 2026; Lin et al., 2025), with some methods additionally conditioning on candidate responses (Gunjal et al., 2026). Contrastive variants (Liu et al., 2026b; Tan et al., 2025; Xie et al., 2025) further compare preferred and rejected responses to identify the dimensions that distinguish stronger outputs from weaker ones, yielding more discriminative criteria. Despite leveraging richer response evidence, one-shot methods still construct rubrics in a single pass, with limited opportunities to correct missing or redundant criteria. This has motivated the development of offline iterative refinement, which progressively validates and revises the generated rubrics. Among them, RRD (Shen et al., 2026) and RubricHub (Li et al., 2026) adopt coarse-to-fine refinement strategies, decomposing broad criteria into finer-grained, actionable evaluation items to improve coverage and applicability. Q-World (Gao et al., 2026) combines hierarchical and lateral expansion to explore a broader evaluation space across different levels of abstraction, task scenarios, and evaluation perspectives. These methods demonstrate the empirical benefits of hierarchical decomposition in rubric construction. Concurrent work (Yang et al., 2026) has also explored tree-structured rubric construction, further extending hierarchical decomposition into an explicit structural representation. Beyond offline refinement, more recent online approaches update rubrics throughout policy optimization using newly generated rollouts (Sheng et al., 2026; Shao et al., 2025; Chen et al., 2026), allowing the evaluation criteria to evolve with the policy.

Rubrics as rewards for LLM optimization. Rubrics can serve not only as evaluation instruments but also as reward functions for open-ended tasks. Rubrics as Rewards (RaR) (Gunjal et al., 2026) constructs instance-specific criteria for each task and uses an LLM to judge whether a policy re-

sponse satisfies each criterion. The resulting criterion-level scores are then aggregated into a scalar reward to guide policy optimization, extending reinforcement learning to open-ended tasks without programmatically verifiable outcomes. Subsequent work further improves the adaptability of rubric-based rewards. InfiMed-ORBIT (Wang et al., 2025) dynamically generates case-conditioned criteria for medical dialogues and employs them in incremental reinforcement learning, whereas DR-Tulu (Shao et al., 2025) continually updates rubrics using rollouts from the current policy, allowing the reward specification to co-evolve with model capabilities and emerging failure modes. Meanwhile, OptimSyn (Fan et al., 2026) optimizes rubrics from the perspective of training utility, using influence estimates of synthetic examples on the target model as feedback to align rubric optimization more directly with downstream performance.

Overall, existing studies have substantially improved rubric generation and refinement. However, they primarily focus on how to construct or update better rubrics, while the mechanisms underlying their effectiveness remain underexplored. In particular, it remains unclear how structural decomposition, criterion-level validity, and score aggregation each contribute to overall rubric effectiveness.

3 What Makes an Effective Rubric? A Mechanistic Analysis

To address this gap, we conduct a systematic mechanistic analysis of rubric effectiveness centered on three research questions. **RQ1:** How does structural organization affect rubric quality? **RQ2:** What makes an individual criterion effective in practice? **RQ3:** Does score aggregation affect rubric evaluation? We construct a controlled analysis set of 1,500 instances through stratified sampling from HealthBench (Arora et al., 2025) and RubricBench (Zhou et al., 2026). Both datasets provide expert-authored atomic evaluation criteria, which we treat as gold rubric leaves. For each query, we manually construct and validate a **gold rubric tree** by recursively grouping semantically related leaves into higher-level evaluation dimensions. Additional details on the datasets and sampling procedure are provided in Appendix A.1.

3.1 RQ1: How Does Structural Organization Affect Rubric Quality?

To answer RQ1, we examine two potential roles of structural organization in rubric-based evaluation. We investigate whether it helps the generator discover evaluation criteria more systematically and whether the resulting hierarchy provides useful context for judging individual criteria.

Structural organization for criterion generator. We first examine whether structural organization enables more systematic exploration of the evaluation space, rather than merely increasing the number of generated criteria. We compare tree-structured and flat rubric generation using the same generator. The hierarchical approach recursively constructs a rubric tree through structured decomposition, whereas the flat approach directly generates a set of criteria without an explicit hierarchy. To control for the additional generator calls introduced by recursive decomposition, we further include Flat-Multi, which independently runs flat generation the same number of times as the corresponding tree construction for each instance and reports the average performance. We use GLM-4.7-Flash as the generator and align the generated criteria with expert-annotated gold rubrics to measure coverage and novelty, with metric definitions provided in Appendix A.2. We use DeepSeek-V4-Flash-GA and Llama3.1-405B-Instruct as two independent judges.

Table 1: Comparison between tree-structured and flat rubric generation.

Judge Model	Coverage (%) ↑			Novelty (%) ↑		
	Flat	Flat-Multi	Tree	Flat	Flat-Multi	Tree
DeepSeek-V4-Flash-GA	52.17	56.78	**57.39**	9.69	9.16	**11.46**
Llama3.1-405B-Instruct	42.11	41.13	**43.90**	22.37	25.09	**27.49**

As shown in Table 1, tree-structured rubric generation outperforms both Flat and Flat-Multi in criterion coverage and valid novelty across judge models. Notably, the gains persist when the number of generator calls is matched, indicating that they cannot be explained by additional generation alone. These results suggest that hierarchical decomposition provides a more systematic scaffold for exploring the evaluation space and recovering fine-grained criteria that flat generation may overlook.

Hierarchical context for criterion judgment. To examine whether hierarchical structure provides useful semantic context for criterion-level judgments, we construct controlled counterfactual responses based on the gold rubric trees. We first verify whether the reference response satisfies all gold criteria and revise it when necessary to obtain a canonical correct response. Starting from this response, we apply minimal local edits to construct three types of counterfactuals with increasing structural scope: (1) single-leaf, violating one gold leaf criterion; (2) within-dimension multi-leaf, jointly violating multiple leaves under the same semantic dimension; and (3) cross-dimension multi-leaf, introducing violations across multiple semantic dimensions. Detailed editing operations are provided in Appendix A.3. We formulate this evaluation as a controlled criterion-level violation detection task. Given a query, a counterfactual response, and its gold rubric, the judge model determines whether each rubric leaf is satisfied. The predefined intervention targets serve as ground-truth labels. We evaluate the performance using exact-match accuracy, i.e., a prediction is correct only if the judge model identifies all and only the violated leaves. We compare three ways of using rubric structure during judgment. FLAT evaluates each leaf independently without hierarchical context. TREE uses the rubric hierarchy for top-down routing, selecting potentially violated branches before making leaf-level judgments. PATH also draws on the rubric tree but removes intermediate routing decisions, instead providing the full root-to-leaf semantic path when evaluating every criterion.

Table 2: Criterion-level violation detection under different uses of rubric hierarchy. All results are reported as percentages, and colored values indicate absolute changes from Flat in percentage points.

Judge Model	Counterfactual Type	Exact-Match Accuracy (%)↑		
		FLAT	TREE	PATH
DeepSeek-V4-Flash-GA	Single-leaf	26.84	26.78 (-0.06)	**30.77** (+3.93)
DeepSeek-V4-Flash-GA	Within-dimension	11.67	10.73 (-0.94)	**12.21** (+0.54)
DeepSeek-V4-Flash-GA	Cross-dimension	23.92	16.44 (-7.48)	**28.18** (+4.26)
Llama3.1-405B-Instruct	Single-leaf	33.91	32.35 (-1.56)	**37.25** (+3.34)
Llama3.1-405B-Instruct	Within-dimension	14.17	13.73 (-0.44)	**16.46** (+2.29)
Llama3.1-405B-Instruct	Cross-dimension	40.60	38.45 (-2.15)	**41.94** (+1.34)

As shown in Table 2, the top-down TREE routing does not outperform FLAT under either judge model. This does not imply that hierarchical structure is inherently ineffective. Instead, the parent-level routing introduces a decision bottleneck, i.e., if the judge fails to identify a violation at an internal node, the corresponding descendant leaves are excluded from further inspection, allowing a single routing error to obscure all violations within that branch. By contrast, PATH consistently outperforms FLAT across all six settings, suggesting that higher-level evaluation dimensions provide useful semantic context for interpreting the scope of individual criteria. Overall, hierarchical structure is more effective as a semantic scaffold for criterion judgment than as a direct basis for top-down pruning without explicit recall safeguards.

To verify that the gains from PATH arise from hierarchical semantics rather than merely increased context length, we introduce a SHUFFLED-PATH control. Specifically, we replace the intermediate dimensions along each root-to-leaf path with semantically mismatched dimensions at the same depth, while keeping all other inputs unchanged. Detailed results are reported in Appendix A.4. Having established the role of structural organization, we next shift from how criteria are organized to the effectiveness of the individual criteria themselves.

3.2 RQ2: WHAT MAKES AN INDIVIDUAL CRITERION EFFECTIVE IN PRACTICE?

We characterize an effective criterion by two operational properties: **discriminative power** and **judge stability**. The former captures whether a criterion can distinguish responses of different quality, while the latter reflects whether its judgments remain consistent across judge models. We investigate how these properties can be realized in practice through two complementary mechanisms, **semantic specification** and **response-based calibration**. We conduct this analysis on 750 preference instances from RubricBench (Zhou et al., 2026). For each instance, we assess discriminative power and judge stability using the two responses provided by the dataset and two independent judge models. We measure discriminative power by the variance of criterion scores across two responses under a fixed judge, and judge stability by the variance in scores across judges for the same re-

Table 3: Effects of semantic specification and response-based calibration on criterion-level properties and pairwise evaluation accuracy, evaluated using Llama-3.3-70B-Instruct and Qwen2.5-72B-Instruct (Qwen et al., 2025) as judge models. *Invalid* denotes the percentage of selected candidate revisions that fail the semantic-validity constraint and is not applicable to DIRECT.

Method	Criterion-Level Properties			Pairwise Accuracy (%)		
	Disc. ↑	Judge Var. ↓	Invalid (%) ↓	Llama	Qwen	Avg. ↑
Direct	0.5105	0.2749	–	48.27	47.87	48.07
Semantic	0.5207	0.2726	**0.00**	**49.87**	47.00	48.44
Calibration	**0.5450**	**0.2510**	17.84	48.27	49.73	49.00
Semantic + Calibration	0.5295	0.2626	**0.00**	49.33	**50.53**	**49.93**

sponse and criterion. In addition, we construct a calibration panel of six responses spanning strong, medium, and weak quality levels for each query. These responses serve only as signals for candidate criterion selection and are not used in the final evaluation. Detailed construction and scoring procedures are provided in Appendix B.2.1.

Can semantic specification induce operational effectiveness? We first examine whether specifying the desired properties during criterion generation is sufficient to realize them in practice. Specifically, we compare two settings, DIRECT and SEMANTIC. DIRECT generates criteria without additional constraints, whereas SEMANTIC explicitly encourages criteria that are semantically valid, discriminative, and judge-stable. We assess the resulting criteria in terms of discriminative power, cross-judge variance, and pairwise evaluation accuracy. As shown in Table 3, semantic specification shifts both operational properties **in the desired direction, but only marginally**. This reveals a gap between specifying desirable criterion properties and realizing them in actual evaluation behavior.

Grounding criterion effectiveness through response-based calibration. The limited gains from semantic specification motivate us to introduce empirical evidence from actual model responses. Rather than assessing criterion quality solely from its semantic description, response-based calibration evaluates how candidate criteria behave on a panel of calibration responses, providing direct evidence of their discriminative power and judge stability. We compare two settings, CALIBRATION and SEMANTIC+CALIBRATION, both operating on the same fixed pool of local rubric revisions. The former selects candidates based on their behavioral characteristics on calibration responses, whereas the latter applies the same utility only after semantic filtering. Detailed definitions are provided in Appendix A.5. As shown in Table 3, CALIBRATION achieves the highest discriminative power and the lowest cross-judge variance. However, 17.84% of the selected criteria are semantically invalid, indicating that operational effectiveness alone does not ensure alignment with the underlying task requirements. In contrast, applying semantic constraints before calibration eliminates invalid selections while retaining most of the gains in discrimination and stability. Consequently, SEMANTIC+CALIBRATION achieves the highest average pairwise evaluation accuracy. These results show that semantic constraints are necessary to prevent task-misaligned criteria, while response-based calibration is needed to identify criteria that are discriminative and stable in practice.

Table 4: Pairwise evaluation accuracy under different score aggregation schemes.

Aggregation Scheme	DeepSeek-V4-Flash-GA (%) ↑	Llama3.1-405B-Instruct (%) ↑
FLAT	**73.20**	73.00
TREE	72.60 (–0.60)	71.33 (–1.67)
FLAT + WU	72.20 (–1.00)	**73.27** (+0.27)
TREE + WU	72.53 (–0.67)	71.33 (–1.67)

3.3 RQ3: DOES SCORE AGGREGATION AFFECT RUBRIC EVALUATION?

RQ1 and RQ2 examine the roles of criterion organization and criterion-level effectiveness in rubric-based evaluation respectively. We now turn to the final stage, where criterion-level judgments are aggregated into an overall score. By fixing the criterion content and varying only the aggregation

strategy, we assess whether weighting and hierarchical aggregation contribute additional gains. For the FLAT setting, the rubric generator independently assigns each leaf l an importance score $a_l \in \{1, ..., 5\}$, which is normalized across all leaves as $w_l = a_l / \sum_{k \in \mathcal{L}} a_k$, where $\mathcal{L}$ denotes the set of leaf criteria. FLAT+WU applies the global weighting scheme introduced in RRD (Shen et al., 2026) to leaf-level scores before aggregation, while TREE+WU applies the same weighting locally within the hierarchy. All methods share identical rubric leaves and criterion-level judgments, differing only in how leaf scores are weighted and aggregated.

As shown in Table 4, neither hierarchical aggregation nor WU weighting yields consistent improvements. With DeepSeek-V4-Flash-GA as the judge, FLAT achieves the highest accuracy, while FLAT+WU provides only a marginal gain of 0.27 percentage points under Llama3.1-405B-Instruct. Overall, once criterion-level evaluation is established, the final aggregation strategy provides limited additional gains, suggesting that rubric effectiveness depends more on how criteria are designed.

4 ARTS-HYBRID: MECHANISM-GUIDED RUBRIC GENERATION AND REFINEMENT

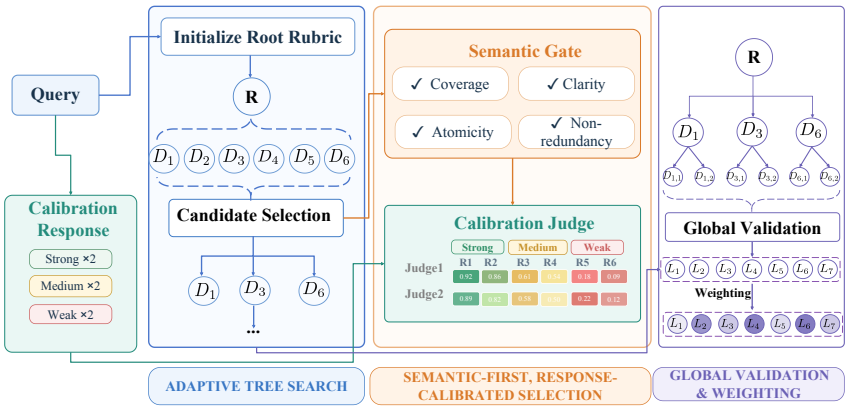


Figure 1: Overview of the ARTS-Hybrid framework.

Building on the above mechanistic findings, we introduce **ARTS-Hybrid**, a rubric generation and refinement framework that translates these insights into a principled construction process. Given a query, ARTS-Hybrid progressively constructs hierarchical evaluation criteria through adaptive tree search. At each node expansion, semantic constraints first identify feasible candidate decompositions, while calibration responses provide empirical evidence to guide the final selection. Once the search terminates, ARTS-Hybrid performs global validation and refinement of the final leaf criteria, followed by normalized weight assignment to produce the final rubric. We next describe the three key components of ARTS-Hybrid in detail, including adaptive tree construction, constrained candidate selection, and global validation with weight assignment.

4.1 ADAPTIVE RUBRIC TREE CONSTRUCTION

Motivated by the findings from RQ1, we adopt a hierarchical tree structure for adaptive rubric construction. Given a query, ARTS-Hybrid first initializes a root rubric representing overall response quality. The root serves only as a virtual structural node for organizing subsequent decomposition and does not directly contribute to the final score. Rather than expanding the tree in a predetermined order, ARTS-Hybrid maintains unprocessed nodes in a priority queue and performs best-first search according to their importance and depth. For each node v at depth d_v, we independently estimate an importance score $I(v) \in \{1, \ldots, 5\}$ according to its relevance to the current task. The expansion priority is defined as

$$P(v) = I(v)\gamma^{d_v}, \tag{1}$$

where $\gamma \in (0, 1]$ denotes the depth discount factor. At each iteration, ARTS-Hybrid selects the highest-priority unexpanded node from the queue and determines whether it is sufficiently atomic

and independently assessable to serve as a leaf criterion. Nodes that satisfy these conditions are finalized as leaves. Otherwise, the rubric generator produces multiple finer-grained child criteria, with the selected children added to the rubric tree and priority queue for further expansion. This process continues until no expandable nodes remain or a predefined stopping condition is reached.

4.2 Semantic-First Response-Calibrated Selection

Once an expandable node is selected, ARTS-Hybrid selects the most suitable decomposition from the generated candidate set. Following the findings from RQ2, semantic constraints and response-based calibration play complementary roles. Accordingly, ARTS-Hybrid adopts a semantic-first, response-calibrated selection strategy.

Semantic feasibility. For each expandable node v, the rubric generator first produces a pool of candidate child criteria. These criteria are then combined into multiple candidate decompositions $\mathcal{C}_v$, where each $C \in \mathcal{C}_v$ represents a possible set of children for node v. ARTS-Hybrid applies a multi-stage semantic validation procedure to determine which decompositions are feasible with respect to the query q and the current rubric tree T. The validation considers several aspects of rubric quality, including criterion clarity, task relevance, independent assessability, coverage of parent requirements, and semantic non-redundancy among child criteria. Only candidate decompositions that pass all semantic checks are retained in the feasible set

$$\mathcal{F}_v = \{C \in \mathcal{C}_v \mid \mathrm{SemValid}(C, q, T) = 1\}, \tag{2}$$

where $\mathrm{SemValid}(C, q, T)$ indicates whether candidate decomposition C satisfies all semantic constraints with respect to the query q and the current rubric tree T. Semantically invalid candidates are filtered out at this stage and do not proceed to subsequent response-based calibration.

Response-based calibration. As demonstrated in RQ2, semantic information alone is insufficient to determine how effectively a criterion operates on actual model responses. We therefore use response-based calibration to provide empirical evidence for selecting among semantically valid decompositions. For each query, we construct a panel of six calibration responses spanning strong, medium, and weak quality levels and reuse it throughout the tree search. For every semantically feasible candidate, two independent calibration judges evaluate each criterion across six calibration responses, yielding a 6×2 response-judge score matrix. This matrix captures criterion behavior along two complementary dimensions. The variation across responses reflects the criterion's discriminative power, whereas variation across judges captures the consistency of its scoring behavior. We further compare response-score profiles across child criteria to identify behavioral redundancy. Based on these signals, we define the following calibration objective for each feasible candidate decomposition C as

$$J(C) = \alpha_D D(C) + \alpha_A A(C) + \alpha_S S(C) - \lambda_{\mathrm{red}} R(C), \tag{3}$$

where $D(C)$ measures response discrimination, $A(C)$ and $S(C)$ capture complementary aspects of cross-judge consistency, and $R(C)$ penalizes redundant scoring behavior among child criteria. Detailed definitions of these terms are provided in Appendix B.1. The final decomposition for node v is then selected from the semantically feasible set by

$$C_v^* = \arg\max_{C \in \mathcal{F}_v} J(C). \tag{4}$$

In this way, semantic validation determines which decompositions are admissible, while response-based calibration determines which admissible decomposition is most effective in practice.

4.3 Global Validation and Weight Assignment

While the above local search ensures semantic validity at each expansion step, locally valid decompositions do not necessarily yield a coherent and complete rubric. ARTS-Hybrid therefore performs global validation once the tree search is complete, ensuring full coverage of the evaluation requirements, leaf-level atomicity, and semantic non-redundancy. When global validation identifies remaining issues, ARTS-Hybrid refines the leaf set by recovering missing evaluation requirements

Table 5: Pairwise accuracy (%) across benchmarks. The best result is highlighted in bold.

Benchmark	Direct Judge	Query-Only	Response-Aware	Q-World	ARTS-Hybrid (ours)
JudgeBench	63.14	58.57	61.43	63.71	**64.29**
RM-Bench Chat	61.24	60.47	59.69	62.79	**64.34**
RB Chat	87.15	82.40	86.87	81.28	**89.39**
Chat-Hard	58.11	64.47	59.87	65.79	**68.20**
PreciseIF	52.92	**56.25**	55.00	56.04	55.63
Focus	79.80	80.27	84.44	83.03	**85.59**

and restructuring existing criteria when necessary. With the leaf set established, ARTS-Hybrid assigns a normalized weight to each criterion. Motivated by the findings from RQ3, we avoid propagating weights through the rubric hierarchy and instead estimate the importance of each leaf directly. Specifically, for each leaf l, the rubric generator assigns an importance score $a_l \in \{1, \ldots, 5\}$ based on the query and the criterion itself.

5 Evaluation of ARTS-Hybrid

We evaluate ARTS-Hybrid in terms of end-to-end evaluation performance and intrinsic rubric quality. In addition, we examine whether its benefits can be transferred to a standalone rubric generator and extended to policy optimization through rubric-based rewards.

5.1 Experimental Setup

Benchmarks and baselines. We evaluate on preference benchmarks across diverse domains and difficulty levels, including JudgeBench (Tan et al., 2025), RM-Bench Chat (Liu et al., 2026a), the Chat and Chat-Hard subsets of RewardBench (Lambert et al., 2025), the Precise-IF and Focus subsets of RewardBench 2 (Malik et al., 2026), and RubricBench (Zhou et al., 2026). Each instance contains a query, a pair of candidate responses, and a preference label. We report pairwise accuracy against human preferences as the primary metric. We compare ARTS-Hybrid with Direct Judge, Query-Only, Response-Aware, and Q-World (Gao et al., 2026). These baselines represent direct evaluation, query-only rubric generation, response-aware rubric generation, and structured rubric refinement, respectively. We reproduce all baselines under the same evaluation protocol using the same generator and judge models. Detailed implementation settings are provided in Appendix B.2.

5.2 Main Results

End-to-end evaluation performance. Table 5 reports pairwise evaluation accuracy across six benchmark settings. ARTS-Hybrid achieves the highest accuracy on five of the six settings and remains competitive on PreciseIF. In particular, ARTS-Hybrid improves over both direct judgment and one-shot rubric generation on most benchmarks, while also outperforming Q-World on five settings. These results suggest that the mechanism-guided construction process provides a robust basis for rubric-based evaluation across diverse preference tasks.

Intrinsic rubric quality. We further evaluate the intrinsic quality of the generated rubrics using DeepSeek-V4-Flash as an independent evaluator, with results from an additional evaluator provided in Appendix C. As shown in Table 6, ARTS-Hybrid achieves the highest scores across all four quality dimensions, including coverage, concreteness, non-redundancy, and actionability. This result complements the end-to-end evaluation results by showing that the gains are accompanied by improvements in the quality of the generated criteria themselves.

Rubric generator distillation. The ARTS-Hybrid incurs additional computation primarily during offline rubric construction. We therefore examine whether the benefits of mechanism-guided construction can be distilled into a standalone rubric generator, avoiding repeated execution of the full ARTS-Hybrid pipeline. Specifically, we fine-tune the same Qwen3-4B rubric generator using three sources of supervision, including direct-prompt rubrics, expert-provided rubrics from the original dataset, and the rubrics generated by our framework, with the untrained Qwen3-4B generator as

Table 6: Comparison of intrinsic rubric quality. Higher values indicate better performance.

Method	Coverage	Concreteness	Non-red.	Actionability	Overall
Query-Only	1.926	2.720	2.349	2.871	2.049
Response-Aware	1.966	2.677	2.329	2.734	2.029
Q-World	2.082	2.434	1.041	2.222	1.997
ARTS-Hybrid (ours)	**2.163**	**2.871**	**2.851**	**2.929**	**2.206**

the baseline. As shown in Table 7, training on ARTS-Hybrid rubrics achieves the highest accuracy on JudgeBench. These results suggest that the benefits obtained through offline mechanism-guided construction can be transferred to a standalone rubric generator, providing a potential way to reduce reliance on the full ARTS-Hybrid pipeline for subsequent rubric generation.

Table 7: The performance of Qwen3-4B rubric generators under different SFT supervision sources.

SFT Supervision	JudgeBench Accuracy (%)↑
Qwen3-4B	69.14
Direct-Prompt Rubrics	66.86
OpenRubrics Expert Rubrics	66.57
Qworld Rubrics	64.86
ARTS-Hybrid Rubrics	**69.43**

Rubric-as-Reward for policy optimization. Beyond evaluation, we examine whether ARTS-Hybrid can provide effective reward signals for policy optimization. We stratifiedly sample 3,000 training instances from OpenRubrics (Liu et al., 2026b) and optimize Qwen3-4B using Dr.GRPO (Liu et al., 2025). We compare reward signals derived from the same three rubric sources described above and additionally report the untrained Qwen3-4B model as the base model. Table 8 reports the performance under both thinking and non-thinking settings. The ARTS-Hybrid achieves the best results in both settings, while directly generated and expert-provided rubrics provide limited gains over the base model. These results suggest that the advantages of ARTS-Hybrid are not limited to evaluation, but can also translate into more effective reward signals for policy optimization.

Table 8: The score performance on HealthBench under different rubric-based rewards.

Reward Source	Thinking	Non-Thinking
Qwen3-4B	38.14	34.86
Direct-Prompt Rubric	38.08	34.76
OpenRubrics Expert Rubric	38.33	34.85
Qworld Rubric	37.64	34.95
ARTS-Hybrid Rubric (ours)	**38.93**	**35.18**

Additional experiments. We conduct two additional experiments in Appendix C. (i) **Cross-judge robustness.** We evaluate the generated rubrics with different judge models to assess whether their effectiveness generalizes across evaluators. (ii) **Ablation study.** We conduct an ablation study to examine the contributions of semantic validation and response-based calibration.

6 Conclusion

In this work, we systematically investigate what makes rubric design effective from three perspectives, including criterion organization, criterion-level operationalization, and score aggregation. Guided by the resulting mechanistic findings, we introduce ARTS-Hybrid, which combines adaptive tree construction with semantic-first, response-calibrated refinement. Experiments show that ARTS-Hybrid produces higher-quality rubrics and achieves strong evaluation performance across different judge models. Beyond evaluation, our reinforcement learning experiments further show that well-designed rubrics can provide effective reward signals for model optimization.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, including grammar checking and improving the clarity and fluency of the manuscript. We have not used generative AI tools for developing theoretical framework, formulate mathematical claims, propose hypotheses, design the research methodology or experiments, process datasets, conduct qualitative analysis. Additionally, we used generative AI tools to assist with code implementation. We have reviewed all AI-assisted work. AI-assisted code was reviewed and tested by the authors for correctness, and all AI-edited text was manually reviewed before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the generation and refinement of rubrics for LLM evaluation and optimization. Our experiments use public benchmarks and do not involve human-subject studies or the collection of personally identifiable information. As rubric-based evaluation relies on LLM judges, the resulting criteria and judgments may inherit biases from the underlying models. Our experiments are intended to study rubric construction and evaluation methodology rather than support high-stakes automated decision making. We encourage appropriate human oversight when applying rubric-based evaluation or optimization in sensitive domains.

REPRODUCIBILITY STATEMENT

To support reproducibility, we release the manually constructed counterfactual responses and gold rubric trees used in the mechanistic analysis, together with the core code for both the mechanistic analysis and ARTS-Hybrid, in our anonymous repository (https://anonymous.4open.science/r/Tree_of_Rubrics). The construction and refinement procedures of ARTS-Hybrid are described in the main text. Detailed implementation settings for the mechanistic analysis and ARTS-Hybrid are provided in the appendix. Appendix A.1 describes the data sources and sampling procedure, while Appendix A.3 details the construction of counterfactual responses. The response-based calibration setup, including the construction of calibration responses and the calibration judges, is provided in Appendix B.2.1. Appendix B.2.2 summarizes the key hyperparameters and their search ranges. Finally, Appendix D provides the core prompts used in the mechanistic analysis and ARTS-Hybrid.

REFERENCES

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.

Susan M Brookhart. Appropriate criteria: Key to effective rubrics. In *Frontiers in education*, volume 3, pp. 22. Frontiers Media SA, 2018.

Yifan Chen, Haitao Li, Qingyao Ai, Fengbin Zhu, Tat-Seng Chua, Min Zhang, and Yiqun Liu. Genrubric: Self-evolving rubric generation for scalable LLM evaluation. *CoRR*, abs/2608.29856, 2026. doi: 10.48550/ARXIV.2608.29856. URL https://doi.org/10.48550/arXiv.2608.29856.

Zhiting Fan, Ruizhe Chen, Tianxiang Hu, Ru Peng, Zenan Huang, Haokai Xu, Yixin Chen, Jian Wu, Junbo Zhao, and Zuozhu Liu. Optimsyn: Influence-guided rubrics optimization for synthetic data generation. *arXiv preprint arXiv:2604.00536*, 2026.

Shanghua Gao, Yuchang Su, Pengwei Sui, Curtis Ginder, and Marinka Zitnik. Qworld: Question-specific evaluation criteria for llms. *arXiv preprint arXiv:2603.23522*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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. *The Innovation*, 7(6), 2026.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *Proceedings of ICLR*, pp. 127924–127945, 2026.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of ACL*, pp. 13806–13834, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

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 ACL*, pp. 31320–31344, 2026.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *Proceedings of ICLR*, volume 2025, pp. 47852–47870, 2025.

Dengcan Liu, Fengkai Yang, Xiaohan Wang, Shurui Yan, Jiajun Chai, Jiahao Li, Yikun Ban, Zhendong Mao, Wei Lin, and Guojun Yin. Cdrrm: Contrast-driven rubric generation for reliable and interpretable reward modeling. In *Proceedings of ACM SIGKDD*, pp. 3152–3163, 2026a.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. In *Proceedings of ACL*, pp. 17417–17437, 2026b.

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. In *Proceedings of COLM*, 2025.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah Smith, Hanna Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. In *Proceedings of ICLR*, volume 2026, pp. 144839–144866, 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

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

Leheng Sheng, Wenchang Ma, Ruixin Hong, Xiang Wang, An Zhang, and Tat-Seng Chua. Reinforcing chain-of-thought reasoning with self-evolving rubrics. *arXiv preprint arXiv:2602.10885*, 2026.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, et al. Deep research: A systematic survey. *arXiv preprint arXiv:2512.02038*, 2025.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *Proceedings of ICLR*, volume 2025, pp. 63277–63303, 2025.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists are better than reward models for aligning language models. In *Proceedings of NeurIPS*, volume 38, pp. 114728–114754, 2026.

Pengkai Wang, Pengwei Liu, Qi Zuo, 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.

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

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.

Haocheng Yang, Licheng Pan, Xiaoxi Li, Zhichao Chen, Zhiheng Zhang, Yuan Lu, Haoxuan Li, and Hao Wang. Rubrics on trial: Evolving rubrics from a single query via synthetic pairwise evidence. *arXiv preprint arXiv:2607.15092*, 2026.

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. In *Proceedings of ICLR*, volume 2026, pp. 133430–133457, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the NeurIPS*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, et al. Rubricbench: Aligning model-generated rubrics with human standards. In *Proceedings of ACL*, pp. 31179–31200, 2026.

A Mechanistic Analysis: Experimental Details

A.1 Dataset

The HealthBench dataset (Arora et al., 2025) provides physician-authored criteria for evaluating open-ended medical responses, while RubricBench (Zhou et al., 2026) covers a broader range of domains, including safety, instruction following, dialogue, STEM, and coding, with expert-annotated atomic criteria, candidate response pairs, and human preference labels. We sample 750 instances from fixed versions of each dataset using stratified sampling. For HealthBench, we perform proportional sampling based on theme combinations and rubric cardinality, preserving the original distributions of task themes and rubric complexity. For RubricBench, we retain all Safety and Instruction Following instances and sample the remaining instances from the Chat, STEM, and Code categories according to their original task domains, data sources, and rubric-count intervals. The resulting mechanistic analysis set comprises 1,500 instances across six task categories. We further summarize the distributions of task categories and rubric counts in Figure 2.

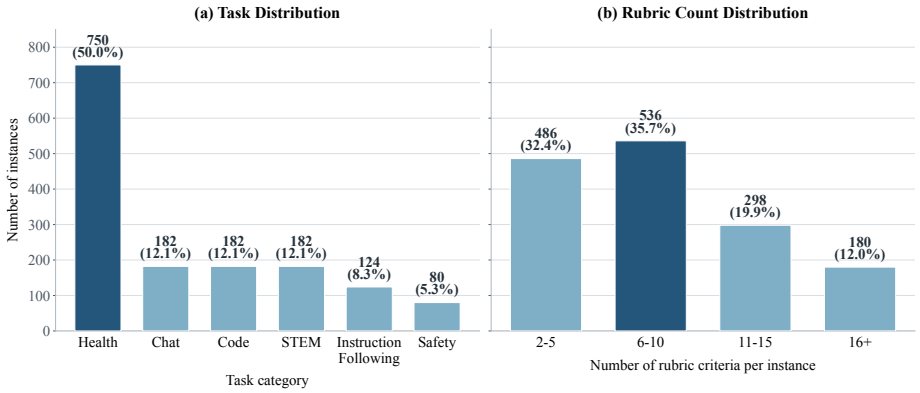


Figure 2: Distribution of task categories and rubric counts in the mechanism analysis set.

A.2 METRIC

Coverage. For the i-th instance, let G_i denote the set of expert-annotated gold leaves and R_i the set of generated rubric criteria. We define the coverage indicator as

$$c_{ig} = \begin{cases} 1, & \text{if the generated rubric fully covers gold leaf } g, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The coverage of instance i is computed as

$$\text{Coverage}_i = \frac{1}{|G_i|} \sum_{g \in G_i} c_{ig}. \tag{6}$$

Novelty. Similarly, the novelty indicator is defined as

$$v_{ir} = \begin{cases} 1, & \text{if } r \text{ is valid, query-relevant, and absent from } G_i, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The valid novelty of instance i is computed as

$$\text{Novelty}_i = \frac{1}{|R_i|} \sum_{r \in R_i} v_{ir}. \tag{8}$$

A.3 COUNTERFACTUAL RESPONSE

In this section, we detail the construction of the counterfactual responses. We follow a minimal-intervention principle, modifying only the content required to violate the targeted gold leaves while leaving unrelated content unchanged. The edits include removing required information, altering factual content, violating formatting constraints, omitting necessary safety warnings, modifying applicability conditions, and introducing behaviors specified by negative criteria. After each edit, we re-evaluate the response against all gold leaves and retain it only if the targeted leaves change state while all non-targeted leaves remain unchanged. If a sample fails this isolation check, it is further revised and discarded if the required isolation cannot be achieved. Across the 1,500 instances, this procedure yields 1,498 single-leaf, 1,482 within-dimension multi-leaf, and 1,409 cross-dimension multi-leaf counterfactual responses.

A.4 SHUFFLED PATH

For each counterfactual type, we randomly sample 50% of the instances for which such mismatched paths can be constructed. We repeat this procedure with two random seeds, denoted as Seed 1 and Seed 2, which correspond to 20260813 and 123456, respectively. For a fair comparison, CORRECT

Table 9: Comparison between correct and shuffled hierarchical paths. Values in parentheses denote changes from CORRECT to SHUFFLED.

	Seed 1		Seed 2	
Counterfactual Type	Correct	Shuffled	Correct	Shuffled
DeepSeek-V4-Flash-GA				
Single-leaf	**23.77**	23.63 (–0.13)	21.23	**21.50** (+0.27)
Within-dimension	**12.55**	11.61 (–0.94)	**12.28**	11.88 (–0.40)
Cross-dimension	**27.66**	25.67 (–1.99)	**27.09**	25.11 (–1.99)
Llama3.1-405B-Instruct				
Single-leaf	**35.91**	34.05 (–1.87)	**37.52**	36.98 (–0.53)
Within-dimension	**14.71**	13.09 (–1.62)	**15.65**	13.77 (–1.89)
Cross-dimension	**43.26**	42.41 (–0.85)	**44.26**	42.55 (–1.70)

PATH and SHUFFLED PATH are evaluated on the same random subset. Consequently, the CORRECT PATH scores vary across seeds and differ from the full-set results in Table 2.

The experimental results are reported in Table 9. The correct path generally outperform their shuffled counterparts across judge models, random seeds, and counterfactual types. This suggests that the gains from path conditioning cannot be attributed to additional context length alone. This indicates that the judge benefits specifically from the structured semantics encoded by the hierarchy.

A.5 CALIBRATION-BASED SELECTION FOR RQ2

For each local candidate C, we use two calibration judges to independently score every criterion on the same set of six calibration responses using a five-point scale. Based on these scores, we characterize each candidate set from four aspects, including response discrimination, cross-judge variance, empirical redundancy, and the number of criteria. Response discrimination captures how much the criterion scores vary across calibration responses and is measured by their average variance. Cross-judge variance reflects the extent to which the two judges disagree on the same criterion-response pair. Finally, the empirical redundancy captures similarity in scoring behavior across criteria and is defined as the proportion of criterion pairs whose response-score profiles have a Pearson correlation above 0.90. We combine these signals into a calibration utility that rewards response discrimination while penalizing cross-judge variance, empirical redundancy, and unnecessary growth in the number of criteria, as defined below.

$$U_{\mathrm{RQ2}}(C) = D_{\mathrm{RQ2}}(C) - V_{\mathrm{RQ2}}(C) - 0.10R_{\mathrm{RQ2}}(C) - 0.02\max(0, |C| - 1). \tag{9}$$

For SEMANTIC+CALIBRATION, this utility is applied only after semantic filtering. A candidate replaces the default semantic choice only when its utility improves by at least 0.05. The calibration responses are used exclusively for candidate selection, while the discrimination, judge variance, and pairwise accuracy reported in Table 3 are computed independently on the held-out response pairs.

B ARTS-HYBRID: EXPERIMENTAL DETAILS

B.1 CALIBRATION OBJECTIVE

For a semantically feasible decomposition C of the current node, let $c \in C$ denote a child criterion. Each criterion is independently evaluated by $n = 2$ calibration judges on the same set of $m = 6$ calibration responses. We define

$$p_{c,r,j} \in [0, 1], \tag{10}$$

where $p_{c,r,j}$ denotes the probability assigned by judge j that response r satisfies criterion c.

Candidate-level score aggregation. To characterize the behavior of a candidate decomposition C on each response, we first aggregate scores across its child criteria using a soft-min operator as

$$z_{r,j}(C) = \left(\frac{1}{|C|} \sum_{c \in C} \max(p_{c,r,j}, \epsilon)^{-\beta} \right)^{-1/\beta}, \tag{11}$$

where $\epsilon = 10^{-6}$ and $\beta = 8$. This aggregation approximates the minimum score across child criteria, such that low satisfaction on any criterion reduces the overall score of the candidate decomposition. When C contains a single criterion, we directly set $z_{r,j}(C) = p_{c,r,j}$.

Based on the aggregated scores $z_{r,j}(C)$, we next define the response discrimination $D(C)$, judge agreement $A(C)$, score stability $S(C)$, and behavioral redundancy $R(C)$ that constitute the calibration objective in Equation 3.

Response discrimination. The response discrimination measures whether a candidate decomposition produces sufficiently varied scores across the calibration responses. We compute the variance of the aggregated scores across responses for each judge and then average over judges as

$$D(C) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{Var}_r \left[z_{r,j}(C) \right]. \tag{12}$$

The larger $D(C)$ indicates greater variation across responses and thus stronger discriminative power.

Cross-judge agreement. The cross-judge agreement measures whether different calibration judges assign similar scores to the same criterion-response pair. For each criterion c, we first compute the score variance across judges for each response as

$$v_{c,r} = \mathrm{Var}_j \left[p_{c,r,j} \right]. \tag{13}$$

We then define the agreement of criterion c as

$$a(c) = \frac{1}{1 + \frac{1}{m} \sum_{r=1}^{m} v_{c,r}}. \tag{14}$$

The overall agreement of a candidate decomposition is obtained by averaging across its child criteria:

$$A(C) = \frac{1}{|C|} \sum_{c \in C} a(c). \tag{15}$$

The larger $A(C)$ indicates greater agreement between judges on the same responses and criteria.

Judge stability. Beyond agreement on individual responses, the judge stability measures whether different judges exhibit similar discrimination behavior across the calibration responses. For each judge j, we first compute the variance of the aggregated scores across responses as

$$u_j(C) = \mathrm{Var}_r \left[z_{r,j}(C) \right]. \tag{16}$$

We then define judge stability as

$$S(C) = \frac{1}{1 + \mathrm{Var}_j \left[u_j(C) \right]}. \tag{17}$$

While $A(C)$ captures agreement on individual responses, $S(C)$ captures the consistency of discrimination patterns across judges, providing a complementary measure of overall scoring stability.

Score redundancy. The score redundancy measures the similarity of scoring patterns across child criteria on the calibration responses. We first average each criterion's scores across judges as

$$\bar{p}_{c,r} = \frac{1}{n} \sum_{j=1}^{n} p_{c,r,j}. \tag{18}$$

For each criterion c, let

$$\bar{\mathbf{p}}_c = [\bar{p}_{c,1}, \ldots, \bar{p}_{c,m}] \tag{19}$$

denote the score vector across calibration responses. We use Spearman rank correlation (Spearman, 1987) to measure the similarity between the scoring profiles of each pair of child criteria, and define $R(C)$ as the maximum positive correlation

$$R(C) = \max_{c,c' \in C} \max\left(0,\ \rho_{\mathrm{S}}\left(\bar{\mathbf{p}}_c, \bar{\mathbf{p}}_{c'}\right)\right). \tag{20}$$

When $|C| = 1$, we set $R(C) = 0$. The larger $R(C)$ indicates stronger similarity in scoring behavior among child criteria, suggesting increased redundancy.

B.2 Implementation Details

We use GLM-4.7-Flash as the common rubric generator and judge model across all evaluation methods. For ARTS-Hybrid, we follow the response construction and calibration protocol introduced in RQ2, using the same calibration panel and judge models. The specific models used for calibration response generation and judging are listed in Table 10.

Table 10: Calibration response generators and judges used in ARTS-Hybrid. Each response generator produces two responses, and each calibration judge evaluates all six responses.

Role	Target Quality	Model
Response Generator	Strong	Qwen3-30B-A3B-Instruct-2507
Response Generator	Medium	Llama-3.1-70B-Instruct
Response Generator	Weak	Mistral-7B-Instruct-v0.2
Calibration Judge	–	Qwen2.5-72B-Instruct
Calibration Judge	–	Llama-3.3-70B-Instruct

B.2.1 Response-Based Calibration Setup

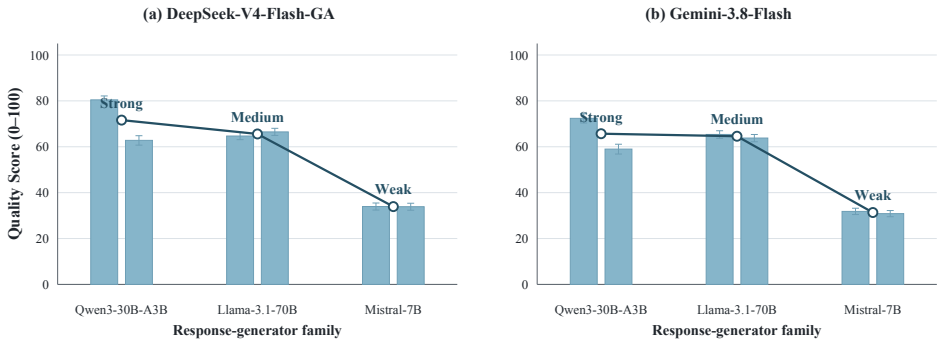


Figure 3: Independent validation of calibration response quality. DeepSeek-V4-Flash-GA and Gemini-3.8-Flash independently evaluate responses generated at the intended strong, medium, and weak quality levels.

Calibration responses. For each query, we pre-generate six calibration responses before rubric construction and keep them fixed throughout the tree search. **The same calibration panel is shared across all candidate criteria.** To cover diverse response quality and generation characteristics, we use three model families, with two responses generated by each model. Specifically, we use Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) to generate two strong responses, Llama-3.1-70B-Instruct to generate two medium-quality responses, and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) to generate two weak responses. We further control the intended quality level through group-specific generation instructions. For the strong group, the instructions emphasize accuracy, completeness and adherence to explicit requirements, with a variant encouraging greater conciseness. For the medium group, the instructions require the response to address the main task while allowing minor omissions or incomplete explanations, with a variant encouraging greater elaboration. For

the weak group, the instructions preserve overall plausibility while allowing incomplete, vague, or poorly organized content, with a variant introducing a minor factual error. Therefore, these responses provide a common calibration basis spanning different quality levels and response characteristics.

To verify that the calibration responses follow the intended quality levels, we use two strong independent evaluators, DeepSeek-V4-Flash-GA and Gemini-3.8-Flash, to assess their response quality. As shown in Figure 3, both evaluators exhibit a consistent quality gradient, with average scores decreasing from the strong to medium and weak response groups. This pattern aligns with the intended quality ordering and supports the use of these responses as a quality-diverse calibration panel.

Calibration judges. We use Qwen2.5-72B-Instruct and Llama-3.3-70B-Instruct as two independent calibration judges. Each judge evaluates the candidate criteria on the same calibration responses, providing the scoring signals used for response-based candidate selection. The calibration responses are used exclusively during rubric construction and are excluded from the final evaluation.

Table 11: The hyperparameter configurations used by ARTS-Hybrid.

Hyperparameter	Value
Leaf budget	16
Candidates per expansion	{6, 10}
Minimum children per expansion	{2, 3}
Maximum tree depth	3
Depth discount γ	0.85
Calibration responses per query	6
Calibration judges	2
Semantic score threshold	0.60
Calibration threshold τ_{scor}	{0.75, 0.80}
Coverage threshold τ_{cov}	{0.80, 1.00}
Maximum regeneration attempts	{2, 4}
Objective weights $(\alpha_D, \alpha_A, \alpha_S)$	(0.25, 0.30, 0.45)
Redundancy penalty λ_{red}	{0.05, 0.10}
Size penalty λ_{size}	{0.00, 0.01}
Maximum global repair rounds	2

B.2.2 Hyperparameters

In this section, we summarize the hyperparameter configurations used in ARTS-Hybrid, including the search settings for rubric tree construction, the thresholds and regeneration settings used during semantic and calibration-based selection, and the weights and regularization terms in the calibration objective. Table 11 reports the corresponding values used in our experiments.

B.2.3 Rubric Generator Distillation

We fine-tune Qwen3-4B on the same stratified set of 3,000 OpenRubrics instances using rubrics from different supervision sources. All variants use identical training examples and hyperparameters. We apply LoRA with rank 16, a scaling factor of 32, and dropout of 0.05 for one epoch, with a learning rate of 2×10^{-5}, a global batch size of 32, and a maximum sequence length of 4,096. Training is conducted on 8 NVIDIA H200 GPUs. For evaluation on JudgeBench, we use GLM-4.7-Flash as the common response judge and to resolve ties in rubric-aggregated scores.

B.2.4 Rubric-as-Reward Training Details

We initialize all policy optimization experiments from the same Qwen3-4B model and use a fixed stratified set of 3,000 OpenRubrics prompts. For each rollout, a frozen Qwen2.5-7B-Instruct model independently evaluates whether the response satisfies each rubric criterion. The resulting binary criterion-level judgments are aggregated using normalized criterion weights to produce a scalar reward. We optimize the policy with Dr.GRPO for three epochs, corresponding to 186 optimization

steps. Each step contains 48 prompts with eight rollouts per prompt. We use an actor learning rate of 1×10^{-6}, maximum prompt and response lengths of 4,096 tokens, and a low-variance KL penalty with a coefficient of 0.001. Following Dr.GRPO, we disable group-reward standard-deviation normalization and use sequence-level token-sum loss aggregation. All reward-source variants use identical policy initialization, training prompts, graders, and optimization settings. The thinking and non-thinking settings differ only in whether thinking mode is enabled. For evaluation on HealthBench (Arora et al., 2025), we use Qwen2.5-7B-Instruct as the judge model to evaluate each generated response against the physician-authored criteria and aggregate the criterion-level judgments into the final HealthBench score.

C Additional Results

Cross-Evaluator rubric quality. To assess robustness across evaluators, we repeat the intrinsic rubric quality evaluation using Llama3.1-405B-Instruct as an additional independent evaluator. As shown in Table 12, the ARTS-Hybrid achieves the highest overall score and performs best in concreteness, non-redundancy, and actionability.

Table 12: Comparison of intrinsic rubric quality. Higher values indicate better performance.

Method	Coverage ↑	Concreteness ↑	Non-red. ↑	Actionability ↑	Overall ↑
Query-Only	3.679	4.251	3.420	4.219	3.609
Response-Aware	3.711	4.262	3.248	4.216	3.600
Q-World	**4.478**	4.047	1.948	4.064	3.119
ARTS-Hybrid (ours)	4.009	**4.405**	**4.073**	**4.402**	**3.938**

Training dynamics of rubric-based rewards. To further examine how different rubric-based rewards affect policy optimization, we track three diagnostic metrics throughout Dr.GRPO training. Figure 4(a) reports the *Informative Group Rate*, defined as the fraction of prompt groups whose rollout rewards are not identical and therefore yield non-zero GRPO advantages. ARTS-Hybrid provides informative relative reward signals for more prompt groups than Direct-Prompt, while OpenRubrics Expert achieves the highest rate. Notably, ARTS-Hybrid still yields better downstream performance than OpenRubrics Expert, suggesting that reward effectiveness depends not only on how often rewards differentiate responses, but also on **whether these differences reflect task-relevant response quality**. Figure 4(b) shows the mean response length during training. The three curves closely overlap and remain stable at approximately 2,000-2,100 tokens, indicating that the gains from ARTS-Hybrid are not driven by a preference for longer responses. Finally, Figure 4(c) shows similar and stable KL loss across all three methods, suggesting that the advantage of ARTS-Hybrid does not arise from more aggressive policy updates.

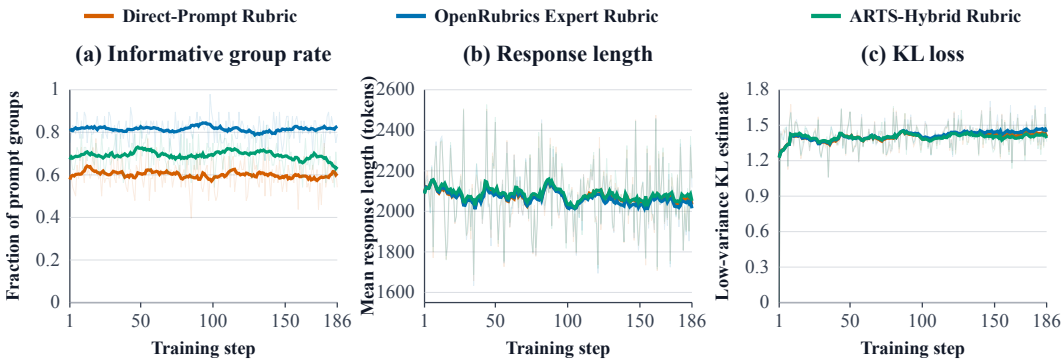


Figure 4: Training dynamics under different rubric-based reward signals, including informative group rate, response length, and KL loss.

Cross-judge robustness. In addition, we examine whether the gains of ARTS-Hybrid are sensitive to the choice of judge model. As shown in Table 13, the ARTS-Hybrid achieves the highest pairwise accuracy with both Qwen3.5-397B-A17B and Llama3.1-405B-Instruct (Grattafiori et al., 2024), with particularly clear gains under the latter. Notably, while the performance of one-shot rubric generation varies considerably across judges, the ARTS-Hybrid maintains consistent improvements over direct judgment. These results suggest that the effectiveness of the generated rubrics is not tied to a specific evaluator and can generalize across different judge models.

Table 13: Pairwise accuracy (%) across judge models. The best result is highlighted in bold.

Method	Qwen3.5-397B-A17B	Llama3.1-405B-Instruct
Direct Judge	77.14	57.40
Query-Only	80.57	41.50
Response-Aware	79.43	56.00
Q-World	80.00	63.86
ARTS-Hybrid (ours)	**85.67**	**69.71**

Ablation study. We conduct an ablation study to examine the contributions of semantic validation and response-based calibration. The results are reported in Table 14. When semantic validation is removed, the performance decreases across all four benchmarks, with the largest drop on the RM-Bench Chat dataset. When response-based calibration is removed, performance decreases on RM-Bench Chat, PreciseIF, and JudgeBench datasets, while remaining comparable on Focus. Overall, both components contribute to the performance of ARTS-Hybrid, with semantic validation showing a more consistent effect across benchmarks.

Table 14: Ablation study of semantic specification and response-based calibration. All results are pairwise evaluation accuracies (%). The best result in each column is highlighted in bold.

Method	RM-Bench Chat	PreciseIF	Focus	JudgeBench
ARTS-Hybrid	**64.34**	**55.63**	85.59	**64.29**
w/o Semantic	50.39	54.17	77.04	61.71
w/o Calibration	58.14	55.00	**86.13**	62.86

D PROMPTS

In this section, we provide the core prompts used in the mechanistic analysis and ARTS-Hybrid. We omit auxiliary prompts for formatting, retries, and other implementation-specific operations.

D.1 PROMPTS FOR MECHANISTIC ANALYSIS

We first provide the prompts used in the mechanistic analysis for RQ1. These include the flat and hierarchical rubric generation prompts used in Table 1, as well as the criterion judgment prompts used to evaluate counterfactual responses in Table 2.

Rubric generation. For flat rubric generation, we use the following prompt.

Flat Rubric Generation

Using ONLY the query, directly create a flat set of atomic evaluation rubrics.

QUERY
<QUERY>

Each item must be independently judgeable, relevant, non−overlapping, and describe what a good response must satisfy. Do not create a hierarchy, inspect or speculate about candidate responses, or reveal an answer.

Return the result in the following format.

{”rubrics”: [
{”criterion”: ”...”, ”category”: ”...”}
]}

Generate 3 to <MAX_LEAVES> items, stopping when the query is adequately covered.

For hierarchical rubric generation proceeds in two stages. We first construct the initial branches of the rubric tree and then recursively examine each node to determine whether further decomposition is needed. The prompts for these two stages are provided below.

Hierarchical Rubric Generation

Create the first level of an adaptive evaluation−rubric tree using ONLY the query below.

QUERY
<QUERY>

The root is ”Overall response quality” Split it into 2−8 distinct and useful semantic branches. Each branch should describe what a good response must satisfy without revealing the answer. Do not inspect or speculate about candidate responses.

Return the result in the following format.

{”children”: [
{”name”: ”...”, ”criterion”: ”...”}
]}

Recursive Node Expansion

Decide whether the current rubric node should remain atomic or be further decomposed, using only the query and its semantic path.

QUERY
<QUERY>

NODE PATH
<NODE_PATH>

CURRENT CRITERION
<CRITERION>

CURRENT DEPTH
<DEPTH>

MAXIMUM DEPTH
<MAX_DEPTH>

Keep the node atomic when it represents one independently judgeable requirement or failure mode. Split it only when it contains multiple non−overlapping requirements that would be more diagnostic if assessed separately. Avoid unary splits and cosmetic restatements.

Return either an atomic decision or a split containing 2−6 child criteria.

Counterfactual violation detection. We provide the core prompt used for counterfactual violation detection under the PATH condition. The root-to-leaf path provides semantic context for judging the final leaf criterion, while intermediate nodes are not used as routing gates.

Path-Conditioned Criterion Judgment

You are given ONE root−to−leaf semantic path from a larger Gold rubric tree. The final item in the path is the only leaf being judged.

Use the parent nodes to understand the leaf's intended semantic scope, terminology, and required evidence. Parent nodes are context rather than separately scored routing gates.

QUERY
<QUERY>

RESPONSE
<RESPONSE>

FULL SEMANTIC PATH
<SEMANTIC_PATH>

LEAF CRITERION
<CRITERION>

DESIRED STATE
<DESIRED_STATE>

Determine whether the response satisfies the final leaf criterion.

Return the result in the following format.

{"passes": true, "evidence": "..."}

We next provide the core prompts used in RQ2, where we examine how semantic assessment and response-based calibration provide complementary signals for criterion selection.

Semantic and calibration-based criterion selection. The Semantic Assessment Prompt determines whether a local revision is semantically better than, or at least no worse than, the original Direct Rubric. The Calibration Scoring Prompt evaluates how candidate criteria behave on actual responses. As described above, calibration-based candidate selection is performed programmatically using the utility defined in Equation 9, rather than through an additional model call. We therefore provide only the Calibration Scoring Prompt below.

Semantic Assessment

You are an independent semantic rubric comparator. Compare an immutable Direct rubric set with one local revision candidate.

QUERY
<QUERY>

DIRECT RUBRIC
<DIRECT_RUBRIC>

CANDIDATE REVISION
<CANDIDATE>

Compare the two alternatives along four dimensions.

1. Coverage of material requirements grounded in the query.
2. Atomicity of independently fail−able requirements.
3. Concrete and observable scoreability.
4. Absence of true semantic duplication or subsumption.

For each dimension, decide whether Direct is better, Candidate is better, or they are tied. Identify any requirement that is lost, weakened, broadened, or duplicated by the candidate.

Do not use Gold rubrics, responses, calibration scores, or preference labels.

Return the comparison for all four dimensions, any candidate regressions, and a brief summary.

Calibration Scoring

Score every calibration response against the ONE rubric below independently.

QUERY
<QUERY>

RUBRIC
<RUBRIC>

CALIBRATION RESPONSES
<CALIBRATION_RESPONSES>

For desired_state=criterion_met, a high score means that the response satisfies the criterion. For desired_state=criterion_not_met, a high score means that the undesirable behavior is absent .

Use the following scale.

1 means clearly violating the desired state.
2 means mostly violating it.
3 means mixed, partial, or genuinely borderline.
4 means mostly satisfying it with a minor defect.
5 means clearly and fully satisfying it.

Judge only observable response content. Do not use the response's declared quality target as a substitute for inspection.

Return one score for each calibration response.

We do not provide a separate prompt for RQ3, as the same judging protocol is used across all settings, with criterion-level judgments held fixed while only the score aggregation scheme is varied.

D.2 PROMPTS FOR ARTS-HYBRID

In this section, we provide the core prompts used in ARTS-Hybrid. We first present the node decomposition prompt used to generate candidate child criteria, followed by the semantic validation prompt used to identify semantically feasible candidates. Calibration-based selection does not require an additional prompt, as feasible candidates are selected programmatically using the calibration objective defined above. Once the tree search is complete, ARTS-Hybrid performs global validation over the resulting leaf set, for which we also provide the corresponding prompt below.

Candidate Decomposition

USER QUERY
<QUERY>

PARENT RUBRIC
<PARENT_RUBRIC>

Generate <N_CANDIDATES> sub−rubrics that jointly preserve and cover the meaning of the parent rubric for this query.

Each sub−rubric must be concrete, verifiable, independently meaningful, specific to the query, and sufficiently unambiguous for different judges to interpret consistently. Avoid redundant near−duplicates.

Use AND semantics. Every selected child should represent a necessary aspect of the parent. Do not assign weights.

Return the selected sub−rubrics.

Semantic Validation

Evaluate whether a rubric criterion is suitable for this exact task.

USER QUERY
<QUERY>

CRITERION
<CRITERION>

Assess the following properties on a 0−1 scale.

− concrete means specific, observable, and not vague.
− verifiable means answerable as YES/NO from the response.
− unambiguous means likely to be interpreted consistently across judges.
− task_specific means tied to the query rather than generic quality.
− distinctive means capable of distinguishing responses of different quality.
− leaf_quality means suitable as a final scoring criterion.
− atomic means containing exactly one independently fail−able requirement.

Also determine whether the criterion is substantive, format−only, and an explicit query requirement.

A criterion is semantically admissible only if it is concrete, verifiable, unambiguous, and task −specific. A valid but non−atomic criterion may be retained as an internal node but not as a final leaf.

Return the assessment for all properties together with the final decision and a brief reason.

Global Validation

Check whether the final leaf rubric set adequately covers the evaluation needs of this task.

USER QUERY
<QUERY>

ROOT RUBRIC
<ROOT_RUBRIC>

FINAL LEAF RUBRICS
<LEAF_RUBRICS>

The leaf set must satisfy the following requirements.

– Cover the principal task–specific requirements.
– Preserve every explicit query constraint.
– Express independently fail–able requirements as separate leaves.
– Contain concrete and directly scoreable criteria.
– Avoid semantic duplication and strict subsumption.
– Include substantive correctness, reasoning, or domain criteria when the task contains substantive content.

If validation fails, identify the missing aspects, atomicity violations, and redundant pairs.

Return the validation decision together with the corresponding diagnostics.

E Case Study of ARTS-Hybrid Rubric Construction

We present a representative example to illustrate how ARTS-Hybrid structures and refines evaluation criteria compared with direct rubric generation.

Query

What is the author's purpose for writing this article? What words or phrases indicate that purpose? How does this text connect back to Chapter One and the Prologue of *Hidden Figures*?

Direct-Generated Rubric

D1. Identify the author's primary purpose.
D2. Provide at least two words or phrases indicating that purpose.
D3. Explain how the text connects to both Chapter One and the Prologue.
D4. Write the response in complete sentences.

ARTS-Hybrid-Generated Rubric

A1. Identify the author's purpose for writing the article.
A2. Cite specific words or phrases that indicate this purpose.
A3. Explain how the article connects to Chapter One.
A4. Support the Chapter One connection with details from the article.
A5. Explain how the article connects to the Prologue.
A6. Use relevant evidence to support the cross-text connections.

Compared with direct generation, the ARTS-Hybrid produces finer-grained criteria that separate independently assessable requirements. The directly generated rubric combines the connections to Chapter One and the Prologue into a single criterion, while introducing additional constraints such as requiring at least two supporting phrases and complete sentences.