Review

Not peer-reviewed version

Seeing the Forest and the Trees: A Survey of Analytic Rubrics for Holistic Reward Modeling in LLMs

Mingqing Yuan †, Xiaobo Liang †, Qipeng Huang, Zixuan Cai, Wanfu Wang, Yang Qiao, Pu Lu, Caishuang Huang, Meng Zhou, Lijun Wu, Juntao Li *, Min Zhang

Posted Date: 25 May 2026

doi: 10.20944/preprints202605.1624.v1

Keywords: large language model; reinforcement learning; rubric reward models

 doi:10.20944/preprints202605.1624.v1

Article

Seeing the Forest and the Trees: A Survey of Analytic Rubrics for Holistic Reward Modeling in LLMs

Mingqing Yuan [1,†], Xiaobo Liang [1,†], Qipeng Huang [1], Zixuan Cai [1], Wanfu Wang [1], Yang Qiao [2], Pu Lu [2], Caishuang Huang [2], Meng Zhou [2], Lijun Wu [3], Juntao Li [1,*] and Min Zhang [1]

[1] Soochow University
[2] Tencent Financial Technology
[3] Shanghai AI Laboratory
* Correspondence: ljt@suda.edu.cn
† Equal contribution.

Abstract

AforestHolistic Reward Models (RMs) are a cornerstone that acts as a proxy for human preferences or rules, enabling the alignment of Large Language Models (LLMs) through reinforcement learning (RL). However, such methods often treat the complex alignment process as a "*black box*", which may introduce risks of both inner and outer misalignment. To address these limitations, treeAnalytical reward modeling, which is also known as **rubric reward modeling**, has recently gained significant attention in both academia and industry. This approach provides multi-dimensional, fine-grained feedback for model behavior by deriving judgments from explicit rubrics, serving as a complementary mechanism to holistic RMs. To this end, we present a comprehensive survey of rubric reward models, covering their foundations, learning paradigms, reasoning and aggregation mechanisms, and functions in the optimization loop. Furthermore, we critically discuss current technical bottlenecks and outline promising directions for future research, offering insights to guide the evolution from *holistic* modeling to *interpretable, analytical* modeling. The comprehensive list of surveyed papers and related resources is maintained at https://github.com/EADMO/Rubric-RMs-Survey.

Keywords: large language model; reinforcement learning; rubric reward models

1. Introduction

Reinforcement learning has become a mainstream paradigm for aligning AI models through interaction with humans and environments [1,2]. In this paradigm, reward models play a central role by translating human preferences into trainable supervision signals for policy optimization [3], while also serving as scalable proxies for continual adaptation [4]. Yet reward modeling is inherently vulnerable to misalignment at multiple levels [5]. *Outer alignment* concerns whether the specified reward faithfully captures the intended human objective. *Inner alignment* concerns whether the learned policy robustly optimizes the specified reward, rather than relying on internally learned proxy objectives (often referred to as *mesa-objectives*) [6] that achieve high reward only superficially or distributionally. This motivates the study of rubric reward models, which seek to make reward specification more explicit, decomposable, and controllable.

As a bridge between human intentions and model optimization, reward modeling plays an anchoring role in both representing human preferences and enabling models to learn from reward signals. As summarized in Figure 1, we analyze the major paradigms in terms of their trade-offs between scalability and alignment risks. *Generative RMs* [7] can be more expressive than *Scalar RMs* [8], as they support preference reasoning, yield more interpretable intermediate judgments, and better exploit inference-time scaling. In contrast, both *Scalar RMs* and *Rule RMs* [9] are constrained by human-provided supervision: *Sacalar RMs* depend on preference annotations, whereas *Rule RMs* rely on manually specified rules, which may limit their scalability. From the perspective of alignment, *Rule*

 doi:10.20944/preprints202605.1624.v1

RMs reduce some inner misalignment concerns because they do not themselves introduce learned internal objectives, but their manually specified criteria are often too narrow to faithfully capture human intentions [10]. In contrast, *Generative* and *Scalar RMs* are better suited to capturing the complexity of human objectives through preference supervision and richer reasoning traces, but they remain less controllable during policy optimization. Beyond these paradigms, *Rubric RMs* [11–13] decompose holistic judgments into multiple explicit and interpretable criteria, providing a more fine-grained interface for aligning reward specification with human objectives. Rubric reward modeling provides a promising way to balance expressiveness, controllability, and scalability by organizing evaluation dimensions into structured and inspectable criteria.

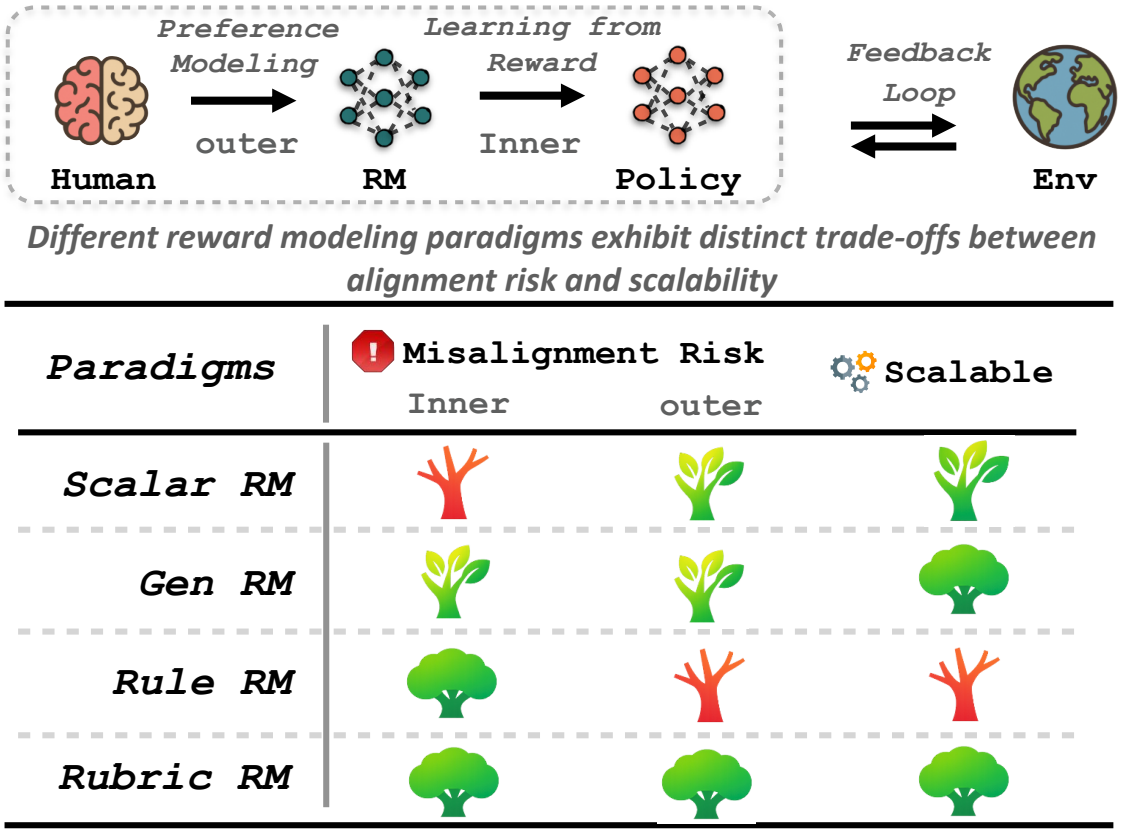

Figure 1. Trade-offs among current reward modeling paradigms in alignment risk and scalability.

treeRubrics originate from educational assessment and psychometrics, where they are primarily introduced to mitigate the reliance of forestholistic scoring on human raters' *overall impressions*, which are prone to cognitive biases such as the *Halo Effect* [14]. In contrast, treeanalytic scoring decomposes evaluation into multiple relatively independent dimensions, helping standardize the assessment process, improve inter-rater reliability, and provide more precise feedback[15]. Motivated by these insight, *Rubric RMs* have recently emerged as a new paradigm for reward modeling, extending earlier approaches based on human-written principles or evaluation criteria [10,16]. However, as this line of work is still evolving, there is not yet a detailed survey that provides a systematic account of its terminology, conceptual foundations, and technical development. To fill the existing gaps in the literature, this survey provides a systematic review of rubric reward modeling. Specifically, we discuss the formulation, modeling, and reasoning of *Rubric RMs*, as well as their role in policy optimization. The main contributions of this survey are as follows:

- To the best of our knowledge, this is the first systematic attempt to survey *Rubric RMs*, an emerging RM paradigm based on explicit and interpretable evaluation criteria.
- We introduce a unified framework for *Rubric RMs* from four perspectives: rubric foundations, modeling, reasoning, and their role in policy optimization, As shown in Figure 2.

 doi:10.20944/preprints202605.1624.v1

- Under this framework, we synthesize representative methods, applications, and evaluation practices, trace the technical evolution of the field, and highlight key open challenges and future directions.

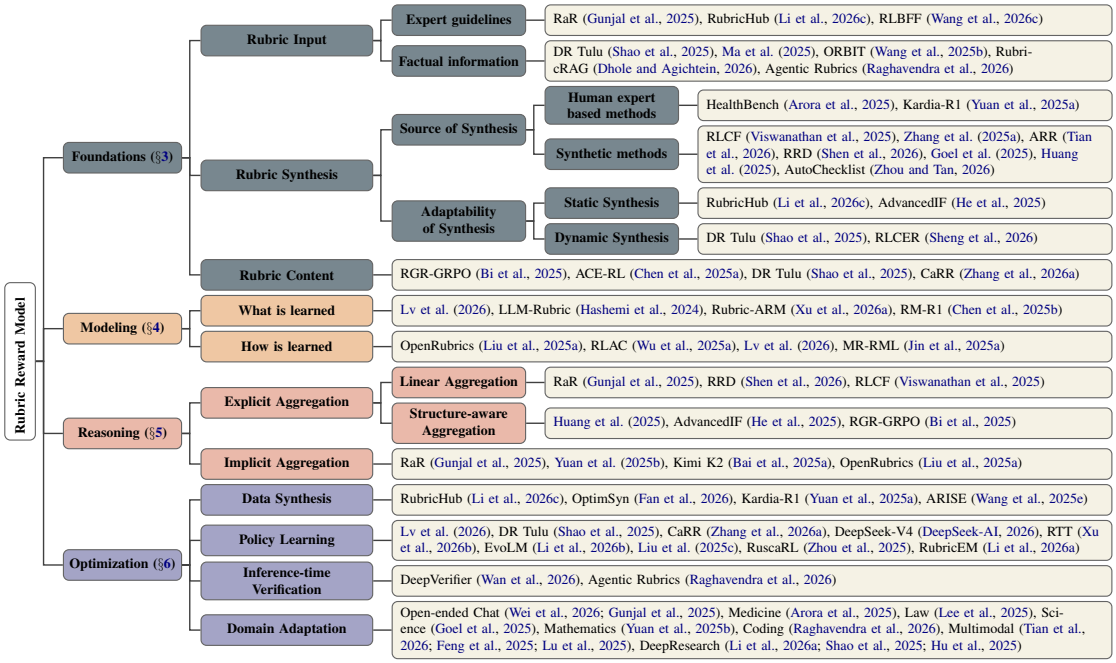

Figure 2. We organize the literature into four perspectives: Foundations, Modeling, Reasoning, and Optimization. Each branch further summarizes key subproblems and representative works. See Appendix A for more details.

2. Overview

In AI alignment [17,18], the RM serves as a bridge between humans and the model, capturing human preferences over states of the world and enabling the model to act consistently with these preferences. In practice, however, human preferences are not directly observable and must be inferred from data. Holistic RMs address this challenge by learning a parameterized reward function R_θ that maps inputs to scalar rewards, thereby implicitly approximating the underlying preference function. Typically, this objective can be effectively operationalized through the Bradley-Terry loss [19], which maximizes the likelihood of observed human preferences. Given a dataset of pairwise comparisons $\mathcal{D} = \{(x, y, y', l)\}$, the loss is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{BT}} = -\,\mathbb{E}\Big[& l \log \sigma(R_\theta(x, y') - R_\theta(x, y)) \\ & + (1 - l) \log \sigma(R_\theta(x, y) - R_\theta(x, y'))\Big],\end{aligned}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $l \in \{0, 1\}$ indicates the preference label, with $l = 1$ meaning y' is preferred over y, and $l = 0$ otherwise. However, the theoretical guarantees of Holistic RMs rely on the assumption of "infinite and consistent preference data," which cannot be satisfied in practice, making them inevitably exposed to the risk of misalignment.

Analytic RMs introduce an intermediate variable to guide preference learning. Drawing upon core concepts from Education and Psychometrics [15,20], and a large body of recently emerging works [11,21], we define a rubric $\mathcal{C}$ as a set of structured, explicit, and fine-grained natural language scoring criteria:

$$\mathcal{C} = \{c_1, c_2, \ldots, c_K\},$$

where each c_k denotes an explicit evaluation criterion. The core paradigm of *Rubric RMs* consists of two distinct phases: rubric construction and rubric-guided judgment. The rubric construction process

 doi:10.20944/preprints202605.1624.v1

is defined as follows: given an input $\mathcal{I}$, a generator $\mathcal{G}$ produces a corresponding set of rubric criteria $\mathcal{C}$. Formally, this is defined as:

$$\mathcal{C} \sim \mathcal{G}(\cdot|\mathcal{I}),$$

Conditioned on $\mathcal{C}$, the judge (RMs) evaluates the candidate responses along multi criteria and aggregates the judgments into an overall reward:

$$R_\theta(x, y \mid \mathcal{C}) = \mathcal{F}(J(x, y, c_1), \ldots, J(x, y, c_K)),$$

where J denotes a criterion-level judge and $\mathcal{F}$ is an aggregation function. In contrast, which infer preferences through a single implicit reward signal, Analytic RMs make the evaluation process more explicit by decomposing overall judgments into criteria before aggregation.

Accordingly, this survey is organized into four parts (in Figure 3): We first discuss the **foundations** (§3) of *Rubric RMs* , including what rubrics are and how they are constructed. We then review **rubric modeling** (§4), focusing on how rubrics are learned, synthesized, or parameterized. Next, we examine **rubric reasoning** (§5), namely how multiple criteria are interpreted and aggregated into reward signals. Finally, we cover **optimization** (§6), where *Rubric RMs* are used to support downstream policy learning and decision-making.

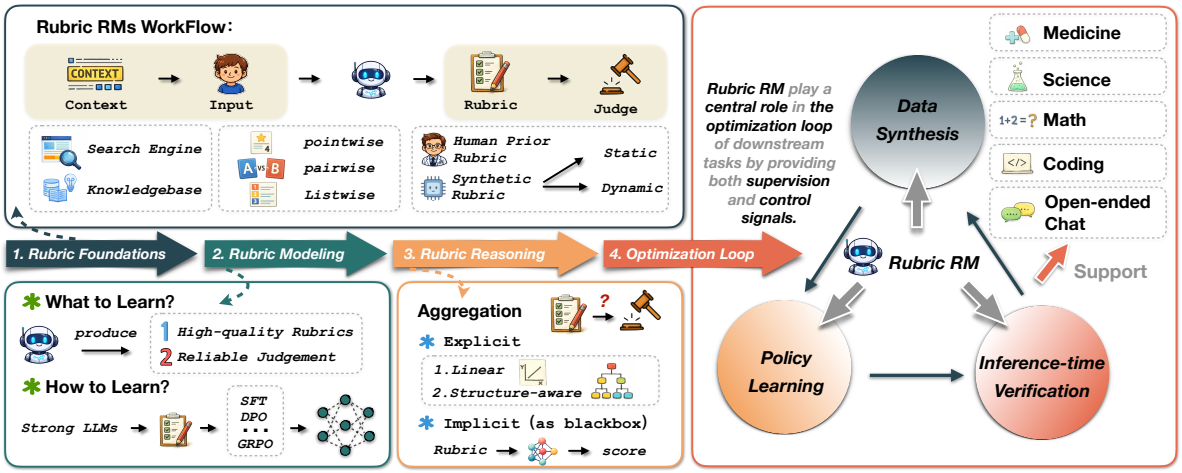

Figure 3. Overview of the Rubric Reward Models (*Rubric RMs*) framework.

3. Foundations

We establish the foundational concepts and methodologies underlying rubric construction. To provide a structured overview, we decompose the rubric formulation pipeline into three primary dimensions: inputs, synthesis, and contents.

3.1. Rubric Inputs

The rubric construction process can be categorized into three settings based on how many candidate responses are considered jointly to extract the criteria: ***Point-wise***: The generator synthesizes a rubric by examining the query [22] or a single response [23] at a time to formulate absolute criteria, optionally accompanied by reference examples [24]. This manner is simple, but often suffers from calibration issues due to the lack of explicit comparison. ***Pair-wise***: The generator considers two responses jointly, extracting criteria based on the contrasting features that determine relative preference [25,26]. This manner naturally produces discriminative criteria and aligns with preference learning frameworks. However, because it is based on local comparisons, it may not fully capture the relative ordering of a larger candidate set. ***List-wise***: The generator analyzes a broader set of responses simultaneously to distill criteria that govern their relative ordering and global quality [27–30]. This manner can capture richer ranking information than pairwise comparison, though it is typically more complex and costly to implement.

 doi:10.20944/preprints202605.1624.v1

The input to generator may include not only the response itself, but also context for rubric construction. Such context is particularly important in long-horizon decision-making tasks, where responses are often affected by hallucinations and cross-step inconsistencies. To improve rubric quality, prior work enriches the context with signals of different origins, ranging from model-internal knowledge to externally grounded feedback. These contextual signals can be broadly organized into the following categories:

Expert guidelines: A common way to enrich rubric context is to introduce normative signals designed by humans, such as predefined extraction principles [11,31,32] or annotations [33,34]. These guidelines provide explicit standards and can substantially improve quality and consistency of the generated rubrics, especially in tasks requiring structured judgment. However, such context often comes with high construction cost and limited scalability, since the quality of the rubric depends on the availability of domain expertise and human supervision.

Factual information: Another line of work augments rubric context with grounded factual signals, including retrieved knowledge from external environments [27,35], retrieving domain knowledge and cases from semantically similar queries as reference [24,36] or feedback from interactions with external systems such as agents or codebases [37]. Compared with purely intrinsic context, this paradigm can more effectively mitigate hallucinations in open-domain and complex tasks by anchoring rubric construction to verifiable evidence. Its main limitation is that rubric quality becomes highly dependent on the reliability, latency, and coverage of the external feedback source.

3.2. Rubric Synthesis.

Given a target task and input, the corresponding rubric can be constructed and incorporated into the alignment pipeline, which can be characterized along two dimensions: source and adaptability.

3.2.1. Source of Synthesis

Human expert based methods rely on manually designed criteria provided by human experts. Such rubrics often encode subjective yet valuable human judgments, domain-specific standards, and task-specific requirements that are difficult to capture automatically, especially in specialized or high-stakes domains. Early work often focused on universal rubrics designed for a dataset or domain [10,38], while more recent efforts increasingly emphasize instance-specific and fine-grained criteria [21]. Although expert-designed rubrics usually offer strong alignment quality and interpretability, they are inherently limited in scalability, since constructing and maintaining high-quality rubrics requires substantial human effort and expertise.

Synthetic methods construct rubrics automatically, typically using LLMs as the generator. A prevalent mechanism involves directly prompting LLMs to generate rubrics tailored to specific inputs [11,39]. Beyond simple one-shot prompting, recent methods also explore iterative refinement [13,28,40,41] and filtering mechanisms [33,42,43] to extract more discriminative, reliable, and task-aligned criteria. Taking a different approach, Huang et al. [44] introduce a unique Rubric-First strategy, which pre-constructs a high-quality rubric library and subsequently filters or synthesizes queries that align with these rubrics. Benefiting from their broad prior knowledge and generalization ability, LLMs can synthesize reasonably accurate rubrics across a wide range of tasks and domains, substantially improving scalability over purely human-designed approaches. However, the quality of synthetic rubrics still depends on the model's internal knowledge, prompt design, and external grounding signals when available. Specially, Orcust [34] takes a hybrid approach that integrates human-expert design with automatic synthesis. While human experts define explicit domain requirements and prohibited behaviors, LLMs are leveraged to synthesize articulate additional strategies. It is worth noting that recent work [45] has begun to consolidate these fragmented generation, refinement, and evaluation into a composable pipeline, which enables a more systematic and rigorous comparison of different methodologies.

 doi:10.20944/preprints202605.1624.v1

3.2.2. Adaptability of Synthesis

Beyond the source of synthesis, another important dimension is whether rubrics remain fixed during RL optimization; accordingly, existing approaches can be divided into static and dynamic paradigms.

Static synthesis generates a fixed set of rubrics prior to the RL process or relies exclusively on the query during the RL process, making them entirely independent of the evolving policy distribution or rollout behavior. This paradigm is simple, efficient, and easy to scale, and it remains the default setting in most works [31,33]. However, as the policy model π_ϕ evolves, fixed rubrics may gradually become incomplete or exploitable, failing to cover newly emerged failure modes. As a result, static criteria are especially vulnerable to Goodhart-like reward hacking, where the policy learns to satisfy superficial rubric signals without achieving the intended behavior. Some work attempts to alleviate this issue by manually summarizing observed failure patterns and injecting defensive criteria [39,46], but such patching remains labor-intensive and difficult to sustain under continuous policy shift.

Dynamic synthesis updates, extracts, or regenerates rubrics online according to the current policy distribution or rollout behavior. Compared with static rubrics, this paradigm is better suited to capturing evolving failure modes and maintaining evaluative relevance throughout optimization. Recent methods explore several directions, including online rubric extraction, which dynamically constructs rubrics based on rollout results during the reinforcement learning process [25,27,29,47], and joint optimization frameworks where the policy model co-evolves with a dynamic rubric generator [48, 49]. Although dynamic approaches are more promising for robust long-term optimization, they also introduce additional complexity in stability, computational cost, and coordination between rubric generation and policy learning.

3.3. Rubric Contents

Synthesized rubrics specify the criteria used to evaluate a response, action sequence, or decision trajectory. For complex tasks, rubrics are often designed with distinct evaluation foci to assess different aspects of performance separately, such as outcome and process [23], stylistic constraints [50], or positive and negative [27,36]. Furthermore, rubrics can also serve as fine-grained factual constraints on both the content and the reasoning trajectory of response [51,52]. Despite these explorations, there is still no unified consensus on what rubric design is universally optimal; in practice, the most effective formulation often depends on the task complexity, domain requirements, and downstream optimization objective.

4. Modeling

The modeling stage focuses on learning the key components of the rubric evaluation pipeline, including rubric construction, rubric-guided judgment, and, in some cases, their joint optimization.

4.1. What to Learn?

Rubric serves as an intermediate variable in the rubric reward modeling, bridging the designer's intended evaluation objectives and downstream supervision signals. Accordingly, modeling in *Rubric RMs* involves two closely related goals: *optimizing the generator to construct high-quality rubrics*, and *optimizing the judge to perform reliable rubric-guided judgment*. Some studies focus on the *generator*, which serves as the core component of *Rubric RMs* by constructing task-specific evaluation criteria. Lv et al. [22] train a generator aligned with human values using human preference data. In contrast, other studies target the *judge*, namely the reasoning and aggregation component, to improve its ability to interpret generated rubrics and derive reliable reward signals from them. LLM-RUBRIC [53] aligns the judge's evaluations with human values by training a personalized calibration network. Beyond such separate optimization, Xu et al. [54] simultaneously optimize the generator and the judge, while Chen et al. [55] and Liu et al. [43] treat rubric construction and rubric-guided judgment as a unified process, enabling end-to-end optimization.

 doi:10.20944/preprints202605.1624.v1

4.2. How to Learn?

In the absence of explicit intermediate supervision for rubric construction and judgment, many studies directly leverage a strong LLMs as the generator or the judge, relying on its pretrained capabilities for rubric synthesis and score aggregation [31]. While this training-free strategy is flexible, it often struggles in complex scenarios that demand hierarchical intent structures, faithful subtask decomposition, and task-specific reasoning. To mitigate these weaknesses, some works adopt SFT [26, 46,56], which uses large amounts of high-quality training data to teach the model to follow the desired rubric format and perform more reliable task planning and decomposition. More advanced approaches introduce reinforcement learning based methods, which alleviate the reliance on explicit supervision over intermediate reasoning processes and instead optimize model behavior directly from preference signals. Typical examples include Direct Preference Optimization (DPO) [49] and Group Relative Policy Optimization (GRPO) [22,36]. Specially, Jin et al. [57] and Srivastava et al. [58] inject rubric knowledge into the judge's parameters.

5. Reasoning

Rubric Reasoning focuses on how multiple criteria are jointly considered by the judge to produce the final reward or supervision signal. This process is inherently challenging because **the relationships among criteria can be complex**: they may be *complementary*, *partially overlapping*, *hierarchically dependent*, or even *conflicting*. Therefore, rubric reasoning is not merely a matter of simple aggregation, but of deciding how these criteria should interact, be prioritized, and be balanced in context.

5.1. Explicit Aggregation

Explicit aggregation maps criterion-based judgments over a selected rubric set $\mathcal{R}$ into a final scalar value through an explicit aggregation function. Concretely, the judge first assigns a score s_i, usually a binary score, to each criterion $c_i \in \mathcal{R}$, yielding a score vector $\mathbf{s} = [s_1, \ldots, s_K]^\top$, which is then converted into a scalar reward r by applying $\mathcal{F}(\mathbf{s})$. Depending on the relational assumptions encoded in $\mathcal{F}$, such aggregation functions can be broadly divided into two primary categories: *linear aggregation*, which assumes additive and largely independent contributions across criteria, and *structure-aware aggregation*, which captures richer interactions such as hierarchical dependencies.

Linear Aggregation.

Linear function provides a simple and interpretable way to combine multiple evaluation dimensions: $\mathcal{F}(\mathbf{s}) = \frac{\sum_i w_i s_i}{\sum_i w_i}$. Here, w_i reflects the relative importance of criterion c_i in the final evaluation. A larger weight indicates a stronger contribution of the corresponding dimension, whereas a smaller weight suggests lower importance. In many cases, such weights must be manually specified or calibrated with additional supervision in order to faithfully reflect the relative importance of different criteria [11,28]. Since such supervision is often unavailable in practice, many approaches adopt uniform weighting [59,60], e.g., $w_i = 1$, which implicitly assumes that all criteria are equally important. While this simplification improves usability, it also imposes a strong assumption and may limit the generality of linear aggregation in complex scenrios. Furthermore, some works [34,39,61] target scenarios where hard constraints (e.g., format) and soft constraints (e.g., logic) coexist, thereby achieving a synergy between Rule RMs and Rubric RMs. Overall, linear aggregation is interpretable and flexible, making it suitable for dynamically updating rubrics to mitigate reward hacking. However, it struggles to capture complex non-linear dependencies and its computational cost typically grows linearly with the number of criteria.

Structure-Aware Aggregation.

In complex settings, criteria may exhibit non-linear trade-offs or hierarchical dependencies (e.g., safety typically possesses veto power), meaning that the contribution of one dimension to the final reward depends on the values of the others. To address this, Huang et al. [44] attempt to employ

 doi:10.20944/preprints202605.1624.v1

hand-crafted non-linear rules to capture the complex relationships among various criteria. Similarly, He et al. [46] assign a reward signal if and only if all criteria are strictly satisfied. Furthermore, Bi et al. [23] implement a dynamic strategy that activates the evaluation of subsequent process criteria only upon the successful verification of factual outcomes. Despite these advancements, relying on such manually hard-coded logic still struggles to fully encompass complex interactions.

5.2. Implicit Aggregation

Implicit Aggregation leverages the powerful In-context Learning (ICL) capabilities of LLMs. This method foregoes the intermediate step of item-by-item scoring. Instead, it inputs the entire rubric $\mathcal{R}$ as a context prompt into the model. The model performs internal reasoning to weigh various criteria, directly outputting a final scalar signal. The judge J implicitly subsumes the roles of both the individual criterion scoring and the explicit aggregation function $\mathcal{F}$. Gunjal et al. [11] first coined the term Implicit Aggregation, demonstrating, and demonstrate that Implicit Aggregation outperforms Explicit Aggregation under equivalent conditions by experiment. Following this trajectory, Yuan et al. [32], Bai et al. [62] and Liu et al. [56] have also adopted this approach. Despite its empirical success in capturing complex interactions, Implicit Aggregation presents significant challenges. Its inherent "black-box" nature severely restricts interpretability and fine-grained controllability, as researchers can neither observe intermediate criteria scores nor directly adjust their weights. Furthermore, this approach is hindered by context length constraints, increased prefill costs for extensive rubrics, and susceptibility to position bias, where the ordering of criteria within the prompt can skew the model's judgment.

6. Rubric RMs as the Core of the Optimization Loop

Rubric RMs play a central role in the optimization loop of downstream tasks by providing both *supervision* and *control* signals. Beyond assigning scalar rewards, they improve data quality, guide policy optimization, and strengthen inference-time decision making by translating implicit and holistic task objectives into explicit and analyzable criteria. In the following, we discuss how *Rubric RMs* contributes to downstream optimization across different stages of the loop.

6.1. Data Synthesis

Compared with traditional data synthesis methods, rubrics enable more fine-grained quality control over both synthesized outputs and their intermediate construction processes via structured criteria. For result quality, RubricHub [31] develop automated coarse-to-fine rubric generation frameworks to construct comprehensive, fine-grained, and highly discriminative rubrics, and further use rubric-based rejection sampling to curate fine-tuning data. To further optimize data synthesis, Optimsyn [63] treat rubric generation as a learnable policy. It utilizes influence-guided optimization to generate rubrics and synthesize corresponding data, thereby maximizing the synthesized data's training utility on the target model. For process control, rubrics can additionally serve as iterative guidance signals in long-horizon generation. For example, ARISE [64] shows that rubrics not only evaluate intermediate outputs, but also drive step-by-step refinement throughout agentic optimization.

6.2. Policy Learning

In policy learning, rubrics can be transformed into structured supervision signals that improve credit assignment. Compared with traditional outcome feedback, rubric-based supervision provides denser and more structured learning signals, allowing models to more precisely identify which behaviors should be reinforced. At the supervision level, rubrics can guide both outcome-level [22, 27,47] and process-level [48,51,52] optimization by decomposing complex objectives into explicit dimensions. Pushing this granularity even further, recent work translates coarse response-level rubric evaluation into fine-grained token-level credit assignment [65]. Furthermore, some methods [66,67] introduce a self-rewarding mechanism, allowing the policy model to serve simultaneously as both the generator and the judge, thereby mitigating reward overoptimization while enabling the quality of

 doi:10.20944/preprints202605.1624.v1

the reward signal to evolve synchronously with the policy's own capabilities. To further enhance the quality and robustness of these supervision signals, Liu et al. [43] demonstrate inference-time scaling for generative reward models, by parallelly sampling multiple diverse criteria to evaluate responses, and employing a meta reward model to score and filter criteria. Specially, RuscaRL [68] demonstrates that rubrics can serve not only as verifiable rewards, but also as explicit scaffolding for exploration, effectively alleviating exploration bottlenecks and improving reinforcement learning on open-ended reasoning tasks.

6.3. Inference-Time Verification

At inference time, *Rubric RMs* can serve as a verifier to directly improve policy outputs, extending its role beyond training supervision into inference-time decision making. They can support ranking, verification, and revision of candidate outputs, for example in Best-of-N, reranking, self-refinement, and critique-and-revision pipelines, by providing explicit criteria for test-time evaluation. For instance, Wan et al. [69] propose DeepVerifier, a rubric-guided verification framework for Deep Research Agents that enables iterative self-correction through test-time feedback and retry. Beyond iterative refinement, Raghavendra et al. [37] show that context-grounded rubrics constructed by agents can serve as an effective and scalable execution-free verifier for SWE agents, improving test-time selection while providing more interpretable reward signals.

6.4. Domain Adaptation

Rubrics are naturally applicable across a wide range of domains because many downstream tasks rely on implicit expert knowledge that cannot be easily captured by preference signals. The core value of rubric modeling lies in making such domain knowledge explicit by recovering it from task requirements and data, providing a more scalable way to align models with domain-specific objectives. This is particularly important in domains such as medicine [21,24,57], law [52], science [42], mathematics [32], and coding [37], where outputs must satisfy fine-grained constraints over multiple criteria. More broadly, in complex settings such as DeepResearch [22,27,51], and multimodal reasoning [34,60], rubrics can further incorporate expert priors, tool feedback, and environmental interactions as structured constraints, enabling models to better coordinate decision making under real-world task requirements.

7. Evaluation and Benchmarking

In this section, we review existing open-source benchmarks (Table 1) and evaluation methods for *Rubric RMs* evaluation from two perspectives: ***Rubric quality*** evaluates the model's ability to generate accurate and useful rubrics, which directly measures the quality of rubrics. ***Rubric-guide judgment*** evaluates whether the generated rubric can support reliable reward modeling and downstream task judgment. We also provide a collection of related datasets in Appendix B.

 doi:10.20944/preprints202605.1624.v1

Table 1. Overview of reward model benchmarks, grouped into rubric-based and general reward benchmarks, and compared by evaluation paradigm, domain, modality, and scale.

Benchmark	Evaluation	Domain	Modality	Scale
Rubric Benchmarks				
RUBRICBENCH [70]	*Rubric & Pairwise*	*Multi-domain*	Language	1.1k
RUBRICEVAL [71]	*Rubric*	*Multi-domain*	Language	3.5k
HEALTHBENCH [21]	*Rubric*	*Medical*	Language	5k
PAPERBENCH [72]	*Rubric*	*Science*	Language	8.3k
PROFBENCH [73]	*Rubric*	*Science & Business*	Language	40
AUTO-J [74]	*Rubric*	*Multi-domain*	Language	1.3k
Reward Benchmarks				
REWARDBENCH [75]	*Pairwise*	*Multi-domain*	Language	3.0k
MT-BENCH [76]	*Pairwise*	*Multi-domain*	Language	80
RM-BENCH [77]	*Pairwise*	*Multi-domain*	Language	1.3k
JUDGEBENCH [78]	*Pairwise*	*Multi-domain*	Language	350
M-REWARDBENCH [79]	*Pairwise*	*Multi-Lingual*	Language	2.9k
RAG-REWARDBENCH [80]	*Pairwise*	*Multi-domain*	Language	1.5k
REWARDMATH [81]	*BoN*	*MATH*	Language	483
REWARDBENCH2 [82]	*BoN*	*Multi-domain*	Language	1.9k
RMB [83]	*Pairwise & BoN*	*Multi-domain*	Language	18.0k
PPE [84]	*Pairwise & BoN*	*Multi-domain*	Language	16.0k
VL-REWARDBENCH [85]	*Pairwise*	*Multi-domain*	Multi-modal	1.3k
MULTIMODALBENCH [86]	*Pairwise*	*Multi-domain*	Multi-modal	5.2k
MJ-BENCH [87]	*Pairwise*	*Multi-domain*	Multi-modal	3.0k

Rubric Benchmarks.

Rubric Benchmarks can be directly used to evaluate the quality of rubrics. RUBRICBENCH provides a curated collection of challenging, expert-annotated examples for systematically evaluating reward models under rubric-guided preference alignment. RUBRICEVAL is the first rubric-level meta-evaluation benchmark designed to assess the performance and reliability of RMs in instruction-following tasks. In specific domains, HEALTHBENCH introduces expert-defined rubrics for clinical tasks and employs LLM-based judges to assess whether model outputs adhere to medical standards. PAPERBENCH extends this paradigm to scientific research, using multi-step rubrics to evaluate reproducibility in machine learning. PROFBENCH further broadens the scope to professional domains such as science and business. As a result, this fine-grained evaluation is particularly suitable for high-stakes domains, such as medicine, scientific research, and other settings where evaluation criteria require expert specification.

Reward Benchmarks.

Reward Benchmarks can be used to evaluate the effectiveness of judgments guided by model-generated rubrics. Among existing benchmarks, REWARDBENCH has become one of the most widely used general-purpose testbeds, covering preference evaluation across a broad range of domains. REWARDBENCH2 further increases the difficulty of this setting by introducing harder examples and multi-candidate selection. Other benchmarks [76–78] refine evaluation from complementary perspectives, emphasizing robustness to superficial cues and faithfulness to objective reasoning quality. At the same time, holistic evaluation has an important limitation: correct preference outcomes do not necessarily imply correct reward reasoning. This motivates benchmarks such as RMB and PPE, which adopt best-of-N (BoN) evaluation to more stringently test whether reward models can rank multiple candidates, rather than simply distinguish between two responses. Beyond these general-purpose benchmarks, several specialized benchmarks have been introduced for specific scenarios. These include M-REWARDBENCH for multilingual evaluation, RAG-REWARDBENCH for retrieval-

 doi:10.20944/preprints202605.1624.v1

augmented generation, and REWARDMATH for mathematical reasoning. A similar trend is also visible in multimodal evaluation. Benchmarks such as VL-REWARDBENCH, MULTIMODALBENCHMARK, and MJ-BENCH extend preference-based reward evaluation to multimodal settings with cross-modal reasoning challenges, typically relying on strong multimodal foundation models to support the evaluation process.

Evaluation Metrics.

For ***process-level*** rubric quality evaluation, model-generated rubrics are typically compared against human-designed ones using metrics such as *F1*, *Recall*, *N-gram overlap*, *Hallucination*, *Missed*, and *Redundancy* [36,70], which capture different aspects of alignment with human rubrics. For ***outcome-level*** judgment evaluation, *Accuracy* is commonly used to measure the agreement between Rubric RMs and human preferences. In addition, some evaluation protocols assess positional bias by examining the consistency of model judgments across repeated evaluations or when the response order is swapped. Beyond direct preference agreement, the capability of *Rubric RMs* can also be reflected through downstream task performance. For example, PPE evaluates whether *Rubric RMs* can select the best response from a Best-of-*K* candidate set on downstream benchmarks such as MMLU-Pro and MATH.

8. Challenges and Future Directions

1. How can we evaluate rubric quality and the faithfulness of rubric-guided judgment under limited human annotation?

Evaluating rubric quality is itself an open problem for *Rubric RMs*, especially under limited human supervision [36,70]. In current practice, rubrics are often treated as intermediate content in reasoning, without explicit constraints or independent verification. This increases a typical *false-positive* risk: a model may reach the correct final judgment even when the rubric it relies on is incomplete, redundant, or otherwise unfaithful. Although prior work has explored various quality dimensions such as discriminability [13];coverage, non-redundancy, and criteria trade-offs [41]; validity [48] and reliability [31] (both of which are defined in education and psychometrics [15]); and value alignment [70,88], there is still no unified definition of rubric quality, nor is there a standardized evaluation framework. Developing a unified framework or designing meta-reward methods remains an important direction for future work. Furthermore, systematic evaluations comparing different rubric construction methodologies are still lacking, and foundational theoretical research regarding the underlying feasibility and mechanisms of rubrics remains extremely scarce [89]. Ultimately, bridging these empirical and theoretical gaps is essential for transitioning Rubric RMs from heuristic tools into rigorous, trustworthy foundations for model alignment.

2. How can common values be identified from human feedback and datasets?

In many practical secnrios, *Rubric RMs* rely on fine-grained values to evaluate and improve individual instances. However, instance-specific rubrics are often too local to support robust reuse or transfer. While they may be effective for a particular example, they often fail to generalize across out-of-distribution tasks or datasets, limiting their ability to provide stable supervision at scale. This motivates the need to move beyond instance-level rubrics and identify more persistent dataset-level values across human feedback and data distributions. Recent efforts in this direction, such as the automatic construction of dataset-level rubrics by Xie et al. [41] and Wang and Xiong [59], suggest that common values may be extractable and reusable beyond individual cases. Furthermore, evaluation will extend beyond the single-instance level to dataset analysis. Dataset-level extraction enables researchers to reverse-engineer the implicit human preference distribution from Dataset-level rubric, thereby allowing for a more scientific assessment of the quality and bias inherent in the dataset itself. Nevertheless, an important open question remains: how should such shared values be discovered, abstracted, and organized from human feedback and dataset-level supervision in a principled way?

 doi:10.20944/preprints202605.1624.v1

3. What is the appropriate structured representation of rubrics, and how can such structure be captured or learned effectively?

Current research has not yet sufficiently addressed how rubrics should be represented as structured objects in complex multi-criteria evaluation. In practice, rubric criteria are rarely fully independent: because it is difficult to construct a completely disentangled rubric. Nevertheless, many existing approaches still implicitly assume that each fine-grained criterion contributes independently to the final judgment. This suggests that an important research direction is to move beyond independent criterion formulations toward more structured representations of rubrics. Rather than treating rubrics as a simple collection of isolated dimensions, future work should explore how to represent and model the relational structure among criteria, including dependency, overlap, priority, and compositionality. Such structured representations are valuable not only for capturing non-linear criterion interactions, but also for improving generalization, interpretability, credit assignment, and context-sensitive evaluation.

4. How should reward models and policies be jointly optimized?

Current LLM reasoning faces a dual exploration dilemma: sparse solution spaces limit autonomous exploration in out-of-distribution tasks, while the lack of fine-grained guidance hinders capability acquisition in long-horizon reasoning. This suggests that the core challenge is not only exploration, but also guidance. Educational theories such as the Zone of Proximal Development (ZPD) [90] and scaffolding [91] suggest that effective learning requires intermediate support, and rubrics are valuable precisely because they can make such intermediate goals explicit [15,92]. From this perspective, *Rubric RMs* are better viewed not as static evaluators of response quality, but as structured capability scaffolds that support progressive policy improvement [23,29,68,93,94]. This shifts the focus from post hoc evaluation to joint optimization, where rubric-based rewards serve not only to assess outputs, but also to support capability acquisition during learning.

9. Conclusion

This paper has presented a comprehensive survey of *Rubric RMs* from four core perspectives: foundations, modeling, reasoning, and their integration into the policy optimization loop. By drawing on established principles from education and psychometrics, *Rubric RMs* mark an important shift from holistic reward modeling toward more analytic assessment. We have further highlighted key bottlenecks in current research and outlined several promising directions for future work. As the field advances, *Rubric RMs* may evolve from evaluation tools into foundational mechanisms for guiding reasoning, exploration, and learning in complex, open-ended domains.

Appendix A. Full Taxonomy

As shown in Figure A1, we provide a more detailed taxonomy of the existing literature. Compared with the taxonomy in the main text, this expanded version offers a more complete and systematic view of the field.

 doi:10.20944/preprints202605.1624.v1

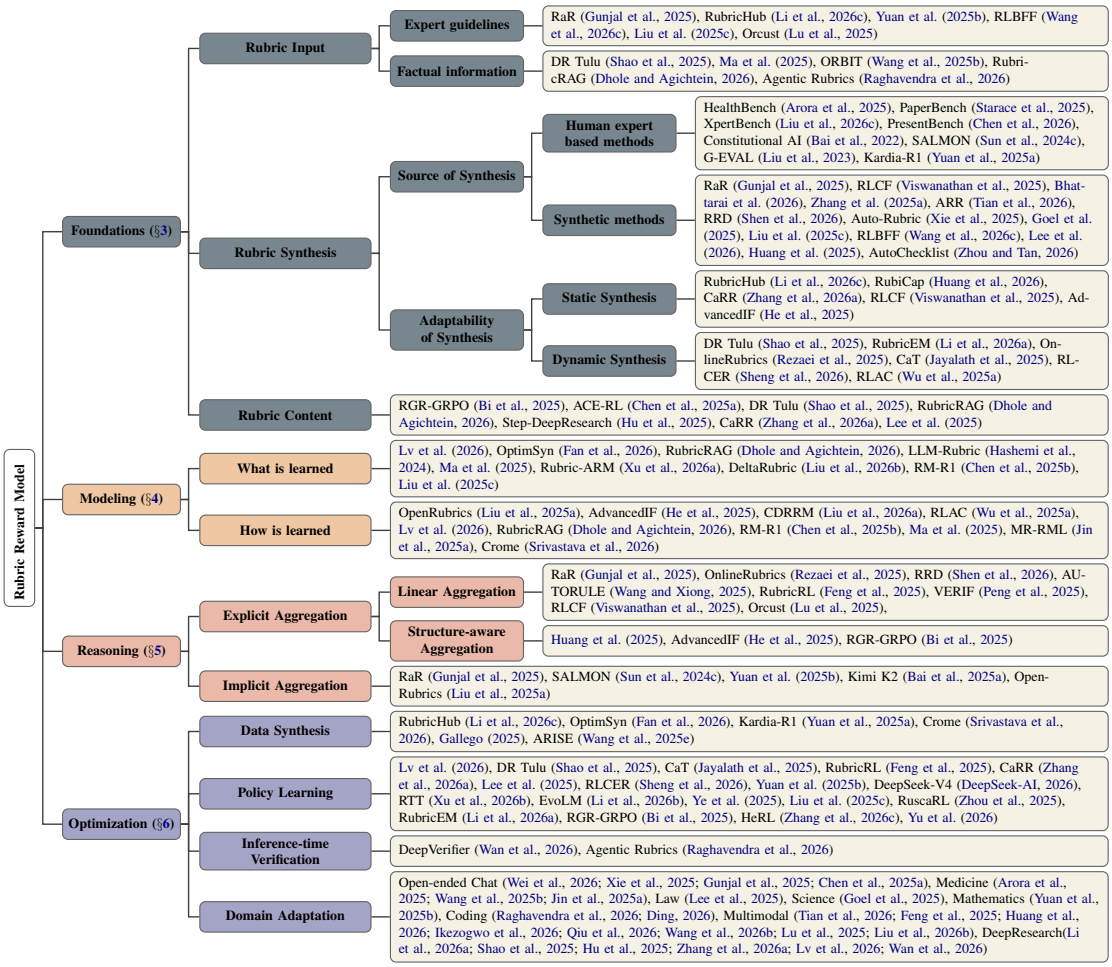

Figure A1. Full taxonomy of Rubric RMs.

 doi:10.20944/preprints202605.1624.v1

Table A1. Overview of Dataset.

Dataset	Domain	Modality	Scale
Rubric Dataset			
RubricHub [31]	*Multi-domain*	Language	110k
WildChecklists [39]	*Multi-domain*	Language	130K
OpenRubrics [56]	*Multi-domain*	Language	35.6k
RaR-Medicine [11]	*Medical*	Language	20k
RaR-Science [11]	*Science*	Language	20k
LEGIT [52]	*Legal*	Language	24K
Auto-Rubric [41]	*Multi-domain*	Language	70
Source Dataset			
WildChat [95]	*Multi-domain*	Language	1m
UltraFeedack [96]	*Multi-domain*	Language	64k
Skywork-Reward [97]	*Multi-domain*	Language	80k
HelpSteer3-Preference [98]	*Multi-domain*	Language	40k
Magpie [99]	*Multi-domain*	Language	300k
MegaScience [100]	*Multi-domain*	Language	1.25m
LongWriter [101]	*Writing*	Language	6k
LongWriter-Zero [102]	*Writing*	Language	8.6k
LongAlign [103]	*Writing*	Language	10k
LMsys [104]	*Chat*	Language	1m
medical-o1 [105]	*Medical*	Language	19.7k
Rlaif-v [106]	*Multi-domain*	Multi-modal	83.1k
LLaVA-Human-Preference-10K [107]	*Multi-domain*	Multi-modal	9.4k
llava-critic-113k [108]	*Multi-domain*	Multi-modal	113.0k
MM-RLHF [109]	*Multi-domain*	Multi-modal	16.3k
pixmo-cap [110]	*Multi-domain*	Multi-modal	717k
DenseFusion-1M [66]	*Multi-domain*	Multi-modal	1m
ViRL-39K [111]	*Multi-domain*	Multi-modal	39k

Appendix B. Training Dataset for Rubric RMs

In many complex scenarios, there is still a practical need for private Rubric RMs. As shown in Table A1, we summarize the high-quality data resources currently used for training, including datasets with explicit rubric and source datasets for extract criteria.

Rubric Datasets.

Recent works have focused on extracting or constructing reusable rubric datasets to facilitate fine-grained evaluation and alignment. RubricHub is a large-scale, multi-domain rubric dataset designed to evaluate and optimize open-ended generation by providing fine-grained, highly discriminative supervision. WildChecklists provides a large-scale collection of synthetically generated checklists to evaluate fine-grained constraint satisfaction using LLMs judges and verifier programs, facilitating language model alignment through reinforcement learning. OpenRubrics is a large-scale, diverse collection of synthetic prompt and rubric pairs derived through contrastive generation. In specialized domains, Gunjal et al. [11] introduce RaR-Medicine and RaR-Science consisting of LLM-generated, instance-specific evaluation criteria, which are utilized as structured, multi-dimensional reward signals for on-policy reinforcement learning to improve complex reasoning in expert domains. LEGIT introduces expert-level rubrics structured as hierarchical legal issue trees extracted from real-world court judgments, which are used to systematically evaluate LLM-generated reasoning traces for both issue coverage and logical correctness. Distinct from prior approaches, Auto-Rubric introduces open-source datasets of compact, query-agnostic "Theme-Tips" rubrics extracted from minimal preference data. Ultimately, these consolidated rubric datasets serve multiple critical functions: training

 doi:10.20944/preprints202605.1624.v1

automated rubric generators, providing dense reward signals for reinforcement learning, and filtering high-quality training data.

Source Datasets.

Existing rubric construction processes typically rely on adapting various source and preference datasets. For instance, WILDCHAT provides a large-scale, multilingual corpus of real-world user-chatbot interactions to capture natural user behaviors. To facilitate preference modeling, datasets like ULTRAFEEDACK and SKYWORK-REWARD provide large-scale, automated preference data; the former utilizes GPT-4 for fine-grained scores and critiques, while the latter offers lightweight, high-quality filtered pairwise comparisons from public sources. Notably, while these rely heavily on model-generated feedback or automated filtering, HELPSTEER3-PREFERENCE distinguishes itself by providing high-quality, human-annotated preference pairs. For a complete list of source datasets, please refer to Table A1.

References

1. Wang, S.; Zhang, S.; Zhang, J.; Hu, R.; Li, X.; Zhang, T.; Li, J.; Wu, F.; Wang, G.; Hovy, E. Reinforcement learning enhanced llms: A survey. arXiv preprint arXiv:2412.10400 **2024**.
2. Zhang, K.; Zuo, Y.; He, B.; Sun, Y.; Liu, R.; Jiang, C.; Fan, Y.; Tian, K.; Jia, G.; Li, P.; et al. A survey of reinforcement learning for large reasoning models. arXiv preprint arXiv:2509.08827 **2025**.
3. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds. Curran Associates, Inc., 2022, Vol. 35, pp. 27730–27744.
4. Silver, D.; Sutton, R.S. Welcome to the era of experience. Google AI **2025**, 1, 11.
5. Hubinger, E.; Van Merwijk, C.; Mikulik, V.; Skalse, J.; Garrabrant, S. Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820 **2019**.
6. Skalse, J.; Howe, N.; Krasheninnikov, D.; Krueger, D. Defining and Characterizing Reward Gaming. In Proceedings of the Advances in Neural Information Processing Systems; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds. Curran Associates, Inc., 2022, Vol. 35, pp. 9460–9471.
7. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Proceedings of the Advances in Neural Information Processing Systems; Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; Levine, S., Eds. Curran Associates, Inc., 2023, Vol. 36, pp. 46595–46623.
8. Sun, H.; Shen, Y.; Ton, J.F. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. arXiv preprint arXiv:2411.04991 **2024**.
9. Mu, T.; Helyar, A.; Heidecke, J.; Achiam, J.; Vallone, A.; Kivlichan, I.; Lin, M.; Beutel, A.; Schulman, J.; Weng, L. Rule Based Rewards for Language Model Safety. In Proceedings of the Advances in Neural Information Processing Systems; Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; Zhang, C., Eds. Curran Associates, Inc., 2024, Vol. 37, pp. 108877–108901. https://doi.org/10.52202/079017-3457.
10. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI Feedback. ArXiv **2022**, abs/2212.08073.
11. Gunjal, A.; Wang, A.; Lau, E.; Nath, V.; He, Y.; Liu, B.; Hendryx, S.M. Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains. In Proceedings of the NeurIPS 2025 Workshop on Efficient Reasoning, 2025.
12. Liang, X.; Zhang, H.; Li, J.; Chen, K.; Zhu, Q.; Zhang, M. Generative Reward Modeling via Synthetic Criteria Preference Learning. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 26755–26769. https://doi.org/10.18653/v1/2025.acl-long.1297.
13. Zhang, J.; Wang, Z.; Gui, L.; Sathyendra, S.M.; Jeong, J.; Veitch, V.; Wang, W.; He, Y.; Liu, B.; Jin, L. Chasing the Tail: Effective Rubric-based Reward Modeling for Large Language Model Post-Training. ArXiv **2025**, abs/2509.21500.
14. Nisbett, R.E.; Wilson, T.D. The halo effect: Evidence for unconscious alteration of judgments. Journal of personality and social psychology **1977**, 35, 250.

 doi:10.20944/preprints202605.1624.v1

15. Jonsson, A.; Svingby, G. The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review **2007**, 2, 130–144.
16. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153.
17. Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852 **2023**.
18. Sierra, C.; Osman, N.; Noriega, P.; Sabater-Mir, J.; Perelló, A. Value alignment: a formal approach. arXiv preprint arXiv:2110.09240 **2021**.
19. Bradley, R.A.; Terry, M.E. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **1952**, 39, 324–345.
20. Nitko, A.J.; Brookhart, S.M. Educational assessment of students; Prentice-Hall, Inc., 2006.
21. Arora, R.K.; Wei, J.; Hicks, R.S.; Bowman, P.; Candela, J.Q.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health. ArXiv **2025**, abs/2505.08775.
22. Lv, C.; Zhou, J.; Zhao, W.; Xu, J.; Huang, Z.; Tian, M.; Dou, S.; Gui, T.; Tian, L.; Zhou, X.; et al. Learning Query-Specific Rubrics from Human Preferences for DeepResearch Report Generation **2026**.
23. Bi, B.; Liu, S.; Wang, Y.; Tong, S.; Mei, L.; Ge, Y.; Xu, Y.; Guo, J.; Cheng, X. Reward and Guidance through Rubrics: Promoting Exploration to Improve Multi-Domain Reasoning. ArXiv **2025**, abs/2511.12344.
24. Wang, P.; Zuo, Q.; Liu, P.; Sang, Z.; Xie, C.; Yang, H. InfiMed-ORBIT: Aligning LLMs on Open-Ended Complex Tasks via Rubric-Based Incremental Training. ArXiv **2025**, abs/2510.15859.
25. Rezaei, M.; Vacareanu, R.; Wang, Z.; Wang, C.; Liu, B.; He, Y.; Akyürek, A.F. Online Rubrics Elicitation from Pairwise Comparisons. ArXiv **2025**, abs/2510.07284.
26. Liu, D.; Yang, F.; Wang, X.; Yan, S.; Chai, J.; Li, J.; Ban, Y.; Mao, Z.; Lin, W.; Yin, G. CDRRM: Contrast-Driven Rubric Generation for Reliable and Interpretable Reward Modeling. arXiv preprint arXiv:2603.08035 **2026**.
27. Shao, R.; Asai, A.; Shen, S.Z.; Ivison, H.; Kishore, V.; Zhuo, J.; Zhao, X.; Park, M.; Finlayson, S.G.; Sontag, D.; et al. DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ArXiv **2025**, abs/2511.19399.
28. Shen, W.F.; Qiu, X.; Whitehouse, C.; Alazraki, L.; Goel, S.; Barbieri, F.; Willi, T.; Mathur, A.; Leontiadis, I. Rethinking Rubric Generation for Improving LLM Judge and Reward Modeling for Open-ended Tasks. arXiv preprint arXiv:2602.05125 **2026**.
29. Li, G.; Mishra, B.D.; Wang, Z.; Yan, J.; Chen, Y.; Li, C.L.; Le, L.T.; Han, R.; Lee, G.; Tong, H.; et al. RubricEM: Meta-RL with Rubric-guided Policy Decomposition beyond Verifiable Rewards. arXiv preprint arXiv:2605.10899 **2026**.
30. Huang, T.H.; Salekin, S.; Movellan, J.; Sala, F.; Bilkhu, M. RubiCap: Rubric-Guided Reinforcement Learning for Dense Image Captioning. arXiv preprint arXiv:2603.09160 **2026**.
31. Li, S.; Zhao, J.; Wei, M.; Ren, H.; Zhou, Y.; Yang, J.; Liu, S.; Zhang, K.; Chen, W. RubricHub: A Comprehensive and Highly Discriminative Rubric Dataset via Automated Coarse-to-Fine Generation. arXiv preprint arXiv:2601.08430 **2026**.
32. Yuan, Y.; Mang, Q.; Chen, J.; Wan, H.; Liu, X.; Xu, J.; Huang, J.T.; Wang, W.; Jiao, W.; He, P. Curing Miracle Steps in LLM Mathematical Reasoning with Rubric Rewards. ArXiv **2025**, abs/2510.07774.
33. Wang, Z.; Zeng, J.; Delalleau, O.; Evans, E.; Egert, D.; Shin, H.C.; Soares, F.; Dong, Y.; Kuchaiev, O. RLBFF: Binary Flexible Feedback to bridge between Human Feedback & Verifiable Rewards. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
34. Lu, J.; Zhang, S.; Xie, Z.; Song, Z.; Zhang, J. Orcust: Stepwise-Feedback Reinforcement Learning for GUI Agent. ArXiv **2025**, abs/2509.17917.
35. Ma, L.; Xu, Y.; Long, X.; Zheng, Z. An Efficient Rubric-based Generative Verifier for Search-Augmented LLMs. ArXiv **2025**, abs/2510.14660.
36. Dhole, K.D.; Agichtein, E. RubricRAG: Towards Interpretable and Reliable LLM Evaluation via Domain Knowledge Retrieval for Rubric Generation. arXiv preprint arXiv:2603.20882 **2026**.
37. Raghavendra, M.; Gunjal, A.; Liu, B.; He, Y. Agentic Rubrics as Contextual Verifiers for SWE Agents. arXiv preprint arXiv:2601.04171 **2026**.

 doi:10.20944/preprints202605.1624.v1

38. Sun, Z.; Shen, Y.; Zhang, H.; Zhou, Q.; Chen, Z.; Cox, D.D.; Yang, Y.; Gan, C. SALMON: Self-Alignment with Instructable Reward Models. In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
39. Viswanathan, V.; Sun, Y.; Kong, X.; Cao, M.; Neubig, G.; Wu, T. Checklists Are Better Than Reward Models For Aligning Language Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
40. Tian, J.; Liu, F.; Han, J.; Jiang, Y.; Wu, Y.; Liu, Y.; Li, H.; Xu, F.; Li, W. Auto-Rubric as Reward: From Implicit Preferences to Explicit Multimodal Generative Criteria. arXiv preprint arXiv:2605.08354 **2026**.
41. Xie, L.; Huang, S.; Zhang, Z.; Zou, A.; Zhai, Y.; Ren, D.; Zhang, K.; Hu, H.; Liu, B.; Chen, H.; et al. Auto-Rubric: Learning to Extract Generalizable Criteria for Reward Modeling. ArXiv **2025**, abs/2510.17314.
42. Goel, S.; Hazra, R.; Jayalath, D.; Willi, T.; Jain, P.; Shen, W.F.; Leontiadis, I.; Barbieri, F.; Bachrach, Y.; Geiping, J.; et al. Training AI Co-Scientists Using Rubric Rewards. arXiv preprint arXiv:2512.23707 **2025**.
43. Liu, Z.; Wang, P.; Xu, R.; Ma, S.; Ruan, C.; Li, P.; Liu, Y.; Wu, Y. Inference-Time Scaling for Generalist Reward Modeling. ArXiv **2025**, abs/2504.02495.
44. Huang, Z.; Zhuang, Y.; Lu, G.; Qin, Z.; Xu, H.; Zhao, T.; Peng, R.; Hu, J.; Shen, Z.; Hu, X.; et al. Reinforcement Learning with Rubric Anchors. ArXiv **2025**, abs/2508.12790.
45. Zhou, K.; Tan, C. AutoChecklist: Composable Pipelines for Checklist Generation and Scoring with LLM-as-a-Judge. arXiv preprint arXiv:2603.07019 **2026**.
46. He, Y.; Li, W.; Zhang, H.; Li, S.; Mandyam, K.; Khosla, S.; Xiong, Y.; Wang, N.; Peng, S.; Li, B.; et al. AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following. ArXiv **2025**, abs/2511.10507.
47. Jayalath, D.H.; wat Goel, S..; Foster, T.; Jain, P.; Gururangan, S.; Zhang, C.; Goyal, A.; Schelten, A. Compute as Teacher: Turning Inference Compute Into Reference-Free Supervision. ArXiv **2025**, abs/2509.14234.
48. Sheng, L.; Ma, W.; Hong, R.; Wang, X.; Zhang, A.; Chua, T.S. Reinforcing Chain-of-Thought Reasoning with Self-Evolving Rubrics. arXiv preprint arXiv:2602.10885 **2026**.
49. Wu, M.; Zhang, G.; Min, S.; Levine, S.; Kumar, A. Rlac: Reinforcement learning with adversarial critic for free-form generation tasks. arXiv preprint arXiv:2511.01758 **2025**.
50. Chen, J.; Sun, W.; Yin, Q.; Tan, Z.; Zhang, J. ACE-RL: Adaptive Constraint-Enhanced Reward for Long-form Generation Reinforcement Learning. arXiv preprint arXiv:2509.04903 **2025**.
51. Zhang, J.; Lv, X.; Feng, L.; Hou, L.; Li, J. Chaining the Evidence: Robust Reinforcement Learning for Deep Search Agents with Citation-Aware Rubric Rewards. arXiv preprint arXiv:2601.06021 **2026**.
52. Lee, J.; On, K.W.; Han, S.; Cohan, A.; Hockenmaier, J. Evaluating Legal Reasoning Traces with Legal Issue Tree Rubrics. ArXiv **2025**, abs/2512.01020.
53. Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; Kedzie, C. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 13806–13834. https://doi.org/10.18653/v1/2024.acl-long.745.
54. Xu, R.; Liu, T.; Dong, Z.; You, T.; Hong, I.; Yang, C.; Zhang, L.; Zhao, T.; Wang, H. Alternating Reinforcement Learning for Rubric-Based Reward Modeling in Non-Verifiable LLM Post-Training. arXiv preprint arXiv:2602.01511 **2026**.
55. Chen, X.; Li, G.; Wang, Z.; Jin, B.; Qian, C.; Wang, Y.; Wang, H.; Zhang, Y.; Zhang, D.; Zhang, T.; et al. RM-R1: Reward Modeling as Reasoning. ArXiv **2025**, abs/2505.02387.
56. Liu, T.; Xu, R.; Yu, T.; Hong, I.; Yang, C.; Zhao, T.; Wang, H. OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment. ArXiv **2025**, abs/2510.07743.
57. Jin, Y.; Li, X.; Cao, F.; Gao, L.; Yao, J. Multidimensional Rubric-oriented Reward Model Learning via Geometric Projection Reference Constraints. ArXiv **2025**, abs/2511.16139.
58. Srivastava, P.; Singh, H.; Madhavan, R.; Patil, G.; Addepalli, S.; Suggala, A.; Aravamudhan, R.; Sharma, S.; Laha, A.; Raghuveer, A.; et al. Robust Reward Modeling via Causal Rubrics. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
59. Wang, T.; Xiong, C. AutoRule: Reasoning Chain-of-thought Extracted Rule-based Rewards Improve Preference Learning. ArXiv **2025**, abs/2506.15651.
60. Feng, X.; Li, Y.; Wan, Z.; Gao, Z.; Yuan, J.; Chen, D.; Qiao, C. RubricRL: Simple Generalizable Rewards for Text-to-Image Generation. ArXiv **2025**, abs/2511.20651.

 doi:10.20944/preprints202605.1624.v1

61. Peng, H.; Qi, Y.; Wang, X.; Xu, B.; Hou, L.; Li, J. VerIF: Verification Engineering for Reinforcement Learning in Instruction Following. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 30324–30339. https://doi.org/10.18653/v1/2025.emnlp-main.1542.
62. Bai, K.T.Y.; Bao, Y.; Chen, G.; Chen, J.; Chen, N.; Chen, R.; Chen, Y.; Chen, Y.; Chen, Y.; Chen, Z.; et al. Kimi K2: Open Agentic Intelligence. ArXiv **2025**, abs/2507.20534.
63. Fan, Z.; Chen, R.; Hu, T.; Peng, R.; Huang, Z.; Xu, H.; Chen, Y.; Wu, J.; Zhao, J.; Liu, Z. Optimsyn: Influence-Guided Rubrics Optimization for Synthetic Data Generation. arXiv preprint arXiv:2604.00536 **2026**.
64. Wang, Z.; Wang, X.; Lee, S.; Xu, X. ARISE: Agentic Rubric-Guided Iterative Survey Engine for Automated Scholarly Paper Generation. ArXiv **2025**, abs/2511.17689.
65. Xu, T.; Zheng, Y.; Lu, P.; Ye, L.; Wu, Y.; Zhang, Z.; Yu, Y.; Ma, C.; Zhu, J.; Liu, P.; et al. Rubrics to Tokens: Bridging Response-level Rubrics and Token-level Rewards in Instruction Following Tasks. arXiv preprint arXiv:2604.02795 **2026**.
66. Li, X.; Zhang, F.; Diao, H.; Wang, Y.; Wang, X.; DUAN, L. DenseFusion-1M: Merging Vision Experts for Comprehensive Multimodal Perception. In Proceedings of the The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024.
67. Ye, Z.; Yue, Y.; Wang, H.; Han, X.; Jiang, J.; Wei, C.; Fan, L.; Liang, J.; Zhang, S.; Li, J.; et al. Self-Rewarding Rubric-Based Reinforcement Learning for Open-Ended Reasoning. ArXiv **2025**, abs/2509.25534.
68. Zhou, Y.; Li, S.; Liu, S.; Fang, W.; Zhao, J.; Yang, J.; Lv, J.; Zhang, K.; Zhou, Y.; Lu, H.; et al. Breaking the Exploration Bottleneck: Rubric-Scaffolded Reinforcement Learning for General LLM Reasoning. ArXiv **2025**, abs/2508.16949.
69. Wan, Y.; Fang, T.; Li, Z.; Huo, Y.; Wang, W.; Mi, H.; Yu, D.; Lyu, M.R. Inference-Time Scaling of Verification: Self-Evolving Deep Research Agents via Test-Time Rubric-Guided Verification. arXiv preprint arXiv:2601.15808 **2026**.
70. Zhang, Q.; Zhou, J.; Wang, Y.; Lyu, F.; Ming, Y.; Xu, C.; Sun, Q.; Zheng, K.; Kang, P.; Liu, X.; et al. RubricBench: Aligning Model-Generated Rubrics with Human Standards. arXiv preprint arXiv:2603.01562 **2026**.
71. Pan, T.; Lin, X.; Yang, W.; He, Q.; Chen, S.; Qi, L.; Xu, W.; Feng, H.; Xu, B.; Xiao, Y. RubricEval: A Rubric-Level Meta-Evaluation Benchmark for LLM Judges in Instruction Following. arXiv preprint arXiv:2603.25133 **2026**.
72. Starace, G.; Jaffe, O.; Sherburn, D.; Aung, J.; Chan, J.S.; Maksin, L.; Dias, R.; Mays, E.; Kinsella, B.; Thompson, W.; et al. PaperBench: Evaluating AI's Ability to Replicate AI Research. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
73. Wang, Z.; Jung, J.; Lu, X.; Diao, S.; Evans, E.; Zeng, J.; Molchanov, P.; Choi, Y.; Kautz, J.; Dong, Y. ProfBench: Multi-Domain Rubrics requiring Professional Knowledge to Answer and Judge. arXiv preprint arXiv:2510.18941 **2025**.
74. Li, J.; Sun, S.; Yuan, W.; Fan, R.Z.; hai zhao.; Liu, P. Generative Judge for Evaluating Alignment. In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
75. Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L.; Lin, B.Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; et al. RewardBench: Evaluating Reward Models for Language Modeling. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025; Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 1755–1797. https://doi.org/10.18653/v1/2025.findings-naacl.96.
76. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Proceedings of the Advances in Neural Information Processing Systems; Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; Levine, S., Eds. Curran Associates, Inc., 2023, Vol. 36, pp. 46595–46623.
77. Liu, Y.; Yao, Z.; Min, R.; Cao, Y.; Hou, L.; Li, J. RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
78. Tan, S.; Zhuang, S.; Montgomery, K.; Tang, W.Y.; Cuadron, A.; Wang, C.; Popa, R.; Stoica, I. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
79. Gureja, S.; Miranda, L.J.V.; Islam, S.B.; Maheshwary, R.; Sharma, D.; Winata, G.T.; Lambert, N.; Ruder, S.; Hooker, S.; Fadaee, M. M-RewardBench: Evaluating Reward Models in Multilingual Settings. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume

 doi:10.20944/preprints202605.1624.v1

1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 43–58. https://doi.org/10.18653/v1/2025.acl-long.3.

80. Jin, Z.; Yuan, H.; Men, T.; Cao, P.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; Zhao, J. RAG-RewardBench: Benchmarking Reward Models in Retrieval Augmented Generation for Preference Alignment. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 17061–17090. https://doi.org/10.18653/v1/2025.findings-acl.877.
81. Kim, S.; Kang, D.; Kwon, T.; Chae, H.; Won, J.; Lee, D.; Yeo, J. Evaluating robustness of reward models for mathematical reasoning. arXiv preprint arXiv:2410.01729 **2024**.
82. Malik, S.; Pyatkin, V.; Land, S.; Morrison, J.; Smith, N.A.; Hajishirzi, H.; Lambert, N. Rewardbench 2: Advancing reward model evaluation. arXiv preprint arXiv:2506.01937 **2025**.
83. Zhou, E.; Zheng, G.; Wang, B.; Xi, Z.; Dou, S.; Bao, R.; Shen, W.; Xiong, L.; Fan, J.; Mou, Y.; et al. Rmb: Comprehensively benchmarking reward models in llm alignment. arXiv preprint arXiv:2410.09893 **2024**.
84. Frick, E.; Li, T.; Chen, C.; Chiang, W.L.; Angelopoulos, A.N.; Jiao, J.; Zhu, B.; Gonzalez, J.E.; Stoica, I. How to Evaluate Reward Models for RLHF. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
85. Li, L.; Wei, Y.; Xie, Z.; Yang, X.; Song, Y.; Wang, P.; An, C.; Liu, T.; Li, S.; Lin, B.Y.; et al. VL-RewardBench: a challenging benchmark for vision-language generative reward models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 24657–24668.
86. Yasunaga, M.; Zettlemoyer, L.; Ghazvininejad, M. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. arXiv preprint arXiv:2502.14191 **2025**.
87. Chen, Z.; Wen, Z.; Du, Y.; Zhou, Y.; Cui, C.; Han, S.; Weng, Z.; Wang, C.; Tong, Z.; HUANG, L.; et al. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.
88. Wang, B.; Liu, Y.; Liu, Y.; Tang, T.; Wang, S.; Gao, C.; Zheng, C.; Zhang, Y.; Yu, L.; Liu, S.; et al. Outcome Accuracy is Not Enough: Aligning the Reasoning Process of Reward Models. arXiv preprint arXiv:2602.04649 **2026**.
89. Mahmoud, A.; Rezaei, M.; Wang, Z.; Gunjal, A.; Liu, B.; He, Y. Reward Hacking in Rubric-Based Reinforcement Learning. arXiv preprint arXiv:2605.12474 **2026**.
90. Vygotsky, L.S. Mind in society: The development of higher psychological processes; Vol. 86, Harvard university press, 1978.
91. Wood, D.; Bruner, J.S.; Ross, G. The role of tutoring in problem solving. Journal of child psychology and psychiatry **1976**, 17, 89–100.
92. Andrade, H.G. Teaching with rubrics: The good, the bad, and the ugly. College teaching **2005**, 53, 27–31.
93. Zhang, W.; Zhang, K.; Qi, J.; Lai, B.; Huang, J. Experience is the Best Teacher: Motivating Effective Exploration in Reinforcement Learning for LLMs. arXiv preprint arXiv:2603.20046 **2026**.
94. Yu, J.; Xu, Z.; Wang, J.; Yang, Y. Think-with-Rubrics: From External Evaluator to Internal Reasoning Guidance. arXiv preprint arXiv:2603.07461 **2026**.
95. Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; Deng, Y. WildChat: 1M ChatGPT Interaction Logs in the Wild. In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
96. Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; et al. ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
97. Liu, C.Y.; Zeng, L.; Liu, J.; Yan, R.; He, J.; Wang, C.; Yan, S.; Liu, Y.; Zhou, Y. Skywork-reward: Bag of tricks for reward modeling in llms. arXiv preprint arXiv:2410.18451 **2024**.
98. Wang, Z.; Zeng, J.; Delalleau, O.; Shin, H.C.; Soares, F.; Bukharin, A.; Evans, E.; Dong, Y.; Kuchaiev, O. HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.
99. Xu, Z.; Jiang, F.; Niu, L.; Deng, Y.; Poovendran, R.; Choi, Y.; Lin, B.Y. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
100. Fan, R.Z.; Wang, Z.; Liu, P. MegaScience: Pushing the Frontiers of Post-Training Datasets for Science Reasoning. arXiv preprint arXiv:2507.16812 **2025**.

 doi:10.20944/preprints202605.1624.v1

101. Bai, Y.; Zhang, J.; Lv, X.; Zheng, L.; Zhu, S.; Hou, L.; Dong, Y.; Tang, J.; Li, J. LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
102. Wu, Y.; Bai, Y.; Hu, Z.; Lee, R.K.W.; Li, J. Longwriter-zero: Mastering ultra-long text generation via reinforcement learning. arXiv preprint arXiv:2506.18841 **2025**.
103. Bai, Y.; Lv, X.; Zhang, J.; He, Y.; Qi, J.; Hou, L.; Tang, J.; Dong, Y.; Li, J. LongAlign: A Recipe for Long Context Alignment of Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 1376–1395. https://doi.org/10.18653/v1/2024.findings-emnlp.74.
104. Zheng, L.; Chiang, W.L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E.; et al. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
105. Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; Wang, B. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs, 2024, [arXiv:cs.CL/2412.18925].
106. Yu, T.; Zhang, H.; Li, Q.; Xu, Q.; Yao, Y.; Chen, D.; Lu, X.; Cui, G.; Dang, Y.; He, T.; et al. RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 19985–19995. https://doi.org/10.1109/CVPR52734.2025.01861.
107. Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.; Wang, Y.X.; Yang, Y.; et al. Aligning Large Multimodal Models with Factually Augmented RLHF. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 13088–13110. https://doi.org/10.18653/v1/2024.findings-acl.775.
108. Xiong, T.; Wang, X.; Guo, D.; Ye, Q.; Fan, H.; Gu, Q.; Huang, H.; Li, C. LLLaVA-Critic: Learning to Evaluate Multimodal Models. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 13618–13628. https://doi.org/10.1109/CVPR52734.2025.01271.
109. Zhang, Y.; Yu, T.; Tian, H.; Fu, C.; Li, P.; Zeng, J.; Xie, W.; Shi, Y.; Zhang, H.; Wu, J.; et al. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
110. Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J.S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 91–104. https://doi.org/10.1109/CVPR52734.2025.00018.
111. Wang, H.; Qu, C.; Huang, Z.; Chu, W.; Lin, F.; Chen, W. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.