A Survey on Rubric-Guided Reinforcement Learning for Language Models

Zifei Shan[*]
WeChat, Tencent
zifeishan@tencent.com

Fangning Shao[*]
Independent Researcher
fangning.shao@gmail.com

Abstract

Reinforcement learning from human feedback (RLHF) has become the dominant paradigm for aligning large language models (LLMs) with human preferences. However, traditional RLHF relies on scalar reward signals that lack interpretability and fail to capture the multi-faceted nature of response quality. *Rubric-guided reinforcement learning* addresses these limitations by introducing structured, interpretable evaluation criteria, or rubrics, as the backbone of reward design, feedback generation, and policy optimization. In this survey, we introduce a Bayesian framework that defines constitutions as prior distributions $P(\mathcal{R})$ over evaluation criteria and rubrics as conditional instantiations $\mathcal{R}_x \sim P(\mathcal{R}|x)$. Under this unified view, we present a taxonomy of rubric-guided RL along the prior-posterior axis, covering constitutional AI, instance-specific rubrics, process-level supervision, self-evolving rubrics, and their agentic and multimodal extensions. Furthermore, as rubrics are natural-language artifacts, we present a linguistic analysis of how granularity trade-offs, semantic drift, and linguistic reward hacking impact alignment reliability, identifying key open problems for future research.

1 Introduction

The alignment of large language models (LLMs) with human values and preferences has emerged as one of the most critical challenges in modern AI research. Reinforcement learning from human feedback (RLHF, Ouyang et al., 2022) has established itself as the predominant paradigm for this alignment, demonstrating that models trained with human-preference-driven reward signals can produce more helpful, harmless, and honest outputs. However, as LLMs are deployed in increasingly diverse and complex domains, the limitations of traditional RLHF, which relies on scalar, opaque reward signals, have become increasingly apparent (Gao et al., 2023; Wu et al., 2023; Wang et al., 2024).

A *rubric* is a structured set of evaluation criteria that specifies what constitutes different levels of performance across multiple dimensions. Rubrics have found natural application in LLM evaluation and alignment, where they provide interpretable, multi-dimensional, and controllable feedback signals (Hashemi et al., 2024; Gunjal et al., 2026; Viswanathan et al., 2025). The integration of rubrics into reinforcement learning for language models, which we term *Rubric-guided Reinforcement Learning*—represents a paradigm shift from implicit, scalar reward optimization to explicit, structured, and interpretable alignment (Gunjal et al., 2026; Xu et al., 2026; Liu et al., 2026b).

The motivation for rubric-guided RL is threefold. First, **interpretability**: scalar reward signals in traditional RLHF are opaque, making it difficult to understand why a particular response receives a high or low score; rubric and checklist methods decompose feedback into human-readable criteria (Hashemi et al., 2024; Viswanathan et al., 2025). Second, **controllability**: rubrics allow fine-grained control over which aspects of response quality are rewarded, enabling alignment that respects specific values, safety constraints, or domain requirements (Wang et al., 2024; Dineen et al., 2025). Third, **self-evolution**: rubrics can be generated, refined, and co-evolved with the policy model, enabling continual self-improvement without external supervision, a capability that scalar reward models lack (Li et al., 2026c; Xu et al., 2026).

The landscape of structured-feedback alignment has expanded rapidly since 2022. Constitutional AI (Bai et al., 2022) pioneered the use of explicit natural-language principles to guide AI feedback for alignment. Subsequent work

[*] Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 34065–34087
October 24-29, 2026 ©2026 Association for Computational Linguistics

has connected rubric-guided alignment to process supervision and verifier-style feedback (Lightman et al., 2024; Gunjal et al., 2026), RLVR for objectively checkable reasoning tasks (Shao et al., 2024; DeepSeek-AI, 2025; Yang et al., 2024), multi-objective reward modeling (Wang et al., 2024), and direct preference optimization (Rafailov et al., 2023). Furthermore, Rule Based Rewards (Mu et al., 2024) demonstrated that composable natural-language rules can serve as direct RLHF rewards for safety and controllability. Recent rubric-specific preprints and workshop papers further explore rubric generation, rubric-based reward modeling, and self-evolving criteria; we discuss these emerging directions in the taxonomy below.

Existing surveys primarily organize the literature around LLM judges, reward design and PRMs, or self-evolution loops (Gu et al., 2024; Ji et al., 2025b; Zheng et al., 2025; Tao et al., 2024). We instead take explicit structured natural-language criteria as the organizing object across feedback construction, reward modeling, and policy updates. Accordingly, we use a two-tier inclusion rule: methods are *core* when such criteria directly condition the training or evaluation signal; generic DPO, RLVR, PRMs, verifiers, and preference-learning methods are *supporting mechanisms* unless explicitly rubric-conditioned. This boundary also motivates our linguistic analysis rather than treating rubrics only as optimization or prompting components.

We introduce a **Bayesian framework** for rubric-guided RL: constitutions are *prior distributions* $P(\mathcal{R})$ over evaluation criteria (global, input-independent), while rubrics are *conditional instantiations* $\mathcal{R}_x \sim P(\mathcal{R}|x)$ (input-specific). This analytical lens places fixed, conditional, and update-oriented criteria in a common vocabulary and interprets rubric exploitation as overfitting to a point estimate; it does not imply that every self-evolving system performs exact Bayesian updating.

We also foreground a **linguistic analysis** of rubrics as natural language texts (§4), examining how semantic drift, granularity effects, ambiguity, and linguistic reward hacking affect rubric-guided RL. This NLP-centered perspective is essential because rubrics are fundamentally natural language artifacts, and their textual properties have been underexplored relative to their mathematical formalization.

Our contributions are as follows:

1. We introduce a **Bayesian framework** for rubric-guided RL, defining constitutions as prior distributions $P(\mathcal{R})$ over evaluation criteria and rubrics as conditional instantiations $P(\mathcal{R}|x)$. This framework unifies fixed constitutions, instance-specific rubrics, and self-evolving criteria, while clarifying its scope and limitations (§2.4).

2. We provide a **linguistic analysis** of rubrics as natural-language artifacts, examining semantic drift, granularity effects, ambiguity, and linguistic reward hacking—phenomena that directly affect the stability and reliability of rubric-guided RL but remain underdeveloped in prior work (§4).

3. We synthesize a **prior-posterior taxonomy** of rubric-guided RL, covering fixed constitutional priors, deterministic instance-specific rubrics, process-level supervision, rubric-based reward modeling, self-evolving rubrics, and extensions to agentic and multimodal settings.

2 Rubric: Definition and Formalization

In this section, we define what a rubric is, formalize its role as a reward signal, and introduce a unifying Bayesian framework that clarifies the relationship between rubrics, constitutions, and principles.

2.1 What is a Rubric?

A rubric is a structured evaluation instrument that specifies criteria for assessing performance across multiple dimensions and quality levels. In the context of LLM alignment, a rubric $\mathcal{R}$ for a given prompt x can be formalized as a set of criteria: $\mathcal{R}(x) = \{c_1, c_2, \ldots, c_K\}$, where each criterion c_k specifies an evaluation dimension (e.g., correctness, completeness, clarity, safety) and may include performance-level descriptors. Unlike scalar reward signals, rubrics decompose the evaluation into interpretable, auditable components. Importantly, rubrics are *natural language texts*: their linguistic properties, including ambiguity, granularity, and semantic stability, fundamentally affect rubric-guided RL performance, as analyzed in §4.

2.2 Rubric as Reward Signal

The key insight of rubric-guided RL is that rubrics can serve as structured reward signals. Given a rubric $\mathcal{R}(x) = \{c_1, \ldots, c_K\}$ and a response y,

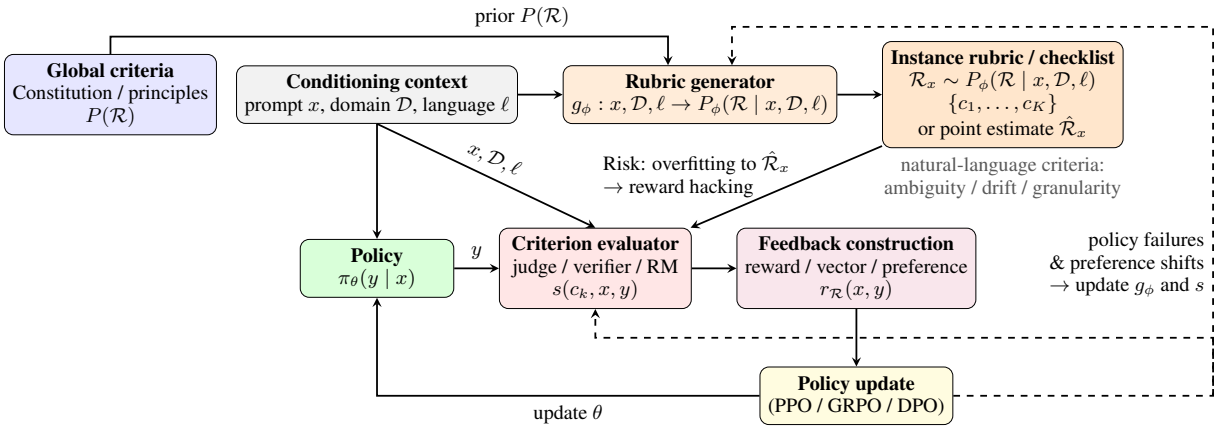


Figure 1: Bayesian view of rubric-guided RL. Global principles define a prior $P(\mathcal{R})$ over evaluation criteria. Given x, $\mathcal{D}$, and ℓ, the rubric generator g_ϕ parameterizes $P_\phi(\mathcal{R} \mid x, \mathcal{D}, \ell)$ and produces sampled rubrics $\mathcal{R}_x$ or deterministic estimates $\hat{\mathcal{R}}_x$. Criterion evaluators construct scalar, vector, or preference feedback for policy optimization. Dashed arrows indicate self-evolution of the generator or evaluator; reward hacking corresponds to overfitting to brittle rubric point estimates.

the rubric-conditioned reward can be computed as: $r_{\mathcal{R}}(x, y) = \sum_{k=1}^{K} w_k \cdot s(c_k, x, y)$, where w_k is the weight for criterion c_k and $s(c_k, x, y)$ is the criterion-specific score. This formulation enables interpretable, controllable, and multi-dimensional reward computation.

2.3 Rubric Generation

Rubric generation is a critical component of rubric-guided RL. Approaches span from contrastive rubric generation (CRG) (Liu et al., 2026b), instance-level checklist generation (Viswanathan et al., 2025), and rubric generation for judges (Shen et al., 2026; Rao and Callison-Burch, 2026) to self-proposed rubrics (Sheng et al., 2026; Li et al., 2026c) and fixed constitutional principles (Bai et al., 2022). The Bayesian framework distinguishes methods that optimize distributions over rubrics from those that construct deterministic, input-conditioned criteria.

2.4 A Bayesian Framework: Constitution as Prior, Rubric as Posterior

To provide a unifying perspective for rubric-guided RL, we introduce a *Bayesian framework* that offers a probabilistic notation for organizing how different methods specify, condition, and update evaluation criteria. This framework resolves terminological ambiguity by explicitly clarifying the relationship between constitutions, rubrics, and principles:

- A **constitution** $P(\mathcal{R})$ functions as a *prior distribution* over evaluation criteria: a global, input-independent specification of what quality dimensions are relevant. Constitutional AI (Bai et al., 2022) defines such a prior through natural-language principles (e.g., helpfulness, harmlessness, honesty) that apply across inputs. We classify constitutions as priors based on their functional role: they specify global evaluation criteria before observing a specific input.

- The conditional distribution $P(\mathcal{R}|x)$ functions as a *posterior distribution* over evaluation criteria, conditioned on the observed input x. An instance-specific **rubric** $\mathcal{R}_x$ is a sampled or deterministic instantiation of this posterior. Instance-specific methods such as RaR (Gunjal et al., 2026) approximate $P(\mathcal{R}|x)$ with a point estimate or checklist, while domain-specific rubrics correspond to coarser conditionals such as $P(\mathcal{R}|\mathcal{D})$. General rubrics are closest to the prior $P(\mathcal{R})$ with minimal conditioning. The Bayesian framing eliminates the need for a rigid hierarchy: constitutions and rubrics can be viewed as the same evaluation object at different levels of conditioning.

- A **principle** is an atomic element of the evaluation space: a single dimension along which quality can be assessed. The sample space of the prior $P(\mathcal{R})$ consists of valid combinations of these principles. In Constitutional AI (Bai et al., 2022), each constitutional principle is one such dimension.

This Bayesian framework formalizes constitutions and rubrics as the same mathematical object at different conditioning levels. The constitution is the unconditional prior $P(\mathcal{R})$; domain-specific rubrics instantiate or approximate coarser conditionals such as $P(\mathcal{R}|\mathcal{D})$; and instance-specific rubrics instantiate or approximate the full posterior $P(\mathcal{R}|x)$. Treating the input x or domain $\mathcal{D}$ as the observation explains the structural difference in generality between these evaluation instruments.

Formalization. We define a rubric-guided RL system as a tuple $(\pi_\theta, P(\mathcal{R}), g_\phi, \mathcal{J}, \mathcal{A})$ where:

- $\pi_\theta(y|x)$ is the policy model;
- $P(\mathcal{R})$ is the prior over evaluation criteria (the constitution);
- g_ϕ is the rubric generator that maps conditioning variables such as the prompt x, domain $\mathcal{D}$, and language ℓ to a conditional distribution $P_\phi(\mathcal{R} \mid x, \mathcal{D}, \ell)$, or to a deterministic point estimate $\hat{\mathcal{R}}_x$;
- $\mathcal{J}$ is a criterion evaluator, such as an LLM judge, verifier, or reward model, that produces criterion-level scores $s(c_k, x, y)$;
- $\mathcal{A}$ constructs feedback from criterion-level scores, such as scalar rewards, multi-dimensional vectors, or preference labels.

Appendix C further clarifies how this notation applies to fixed, deterministic, and self-evolving rubric systems.

Given conditioning variables $(x, \mathcal{D}, \ell)$, an explicitly probabilistic rubric generator samples $\mathcal{R}_x \sim P_\phi(\mathcal{R} \mid x, \mathcal{D}, \ell)$. Deterministic systems instead return a point estimate $\hat{\mathcal{R}}_x = g_\phi(x, \mathcal{D}, \ell)$.

For scalar-reward methods, the rubric-guided reward aggregates criterion-level scores, potentially over multiple rubric samples:

$$r_{\mathcal{R}}(x, y) = \mathbb{E}_{\mathcal{R}_x \sim P_\phi(\mathcal{R}|x,\mathcal{D},\ell)} \left[\mathcal{A}\big(\{s(c_k, x, y)\}_{k=1}^{K}\big) \right] \quad (1)$$

These scalar rewards are typically optimized using KL-regularized policy gradients:

$$\max_\theta \ \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta} \big[r_{\mathcal{R}}(x, y) \big] - \beta \cdot \mathrm{KL}(\pi_\theta \| \pi_{\text{ref}}) \quad (2)$$

Alternatively, direct preference optimization (DPO) methods instantiate the same conceptual framework by modifying both endpoints: the aggregation function $\mathcal{A}$ produces pairwise preferences or labels rather than scalar rewards, and the policy is updated via a pairwise preference loss rather than Equation 2.

3 Taxonomy of Rubric-Guided RL Methods

Building on the Bayesian framework introduced in §2.4, we categorize existing methods under this lens. Table 1 organizes these methods along the prior-posterior axis. Methods differ primarily in how they specify, approximate, or operationalize the transition from global evaluation principles $P(\mathcal{R})$ to input-conditioned criteria $P(\mathcal{R}|x)$. Our coverage prioritizes peer-reviewed publications and stable archival records, while additionally including influential preprints and technical reports that have been widely adopted or cited in the community. Appendix A summarizes the search and inclusion protocol.

3.1 Point-Mass Priors: Constitutional AI

Methods in this category use a fixed set of principles as the prior $P(\mathcal{R})$ and do not condition heavily on the input, yielding a posterior that is a point mass at the prior: $P(\mathcal{R}|x) = \delta_{\mathcal{R}_{\text{constitution}}}(\mathcal{R})$. Representative works include Constitutional AI (Bai et al., 2022), IterAlign (Chen et al., 2024), ICAI (Findeis et al., 2025), and Deliberative Alignment (OpenAI, 2024), which apply fixed principle sets across all inputs rather than generating task-specific criteria.

3.2 Deterministic Posteriors: Instance-Specific Rubrics

Methods in this category generate instance-specific rubrics deterministically, producing a single point estimate from $P(\mathcal{R}|x)$ without uncertainty quantification, including RaR (Gunjal et al., 2026), RLCF (Viswanathan et al., 2025), rubric-conditioned PRMs (Lightman et al., 2024), OpenRubrics (Liu et al., 2026b), rubric-generation methods for LLM judges (Shen et al., 2026; Rao and Callison-Burch, 2026; Siro et al., 2026), and Critique Models (Ke et al., 2024), which map inputs to specific evaluation structures for fine-grained feedback.

Method	Domain	Prior $P(\mathcal{R})$	Posterior $P(\mathcal{R}\|x)$	Aggregation	Training / Optim.
RaR (Gunjal et al., 2026)	Text	Implicit	Deterministic point	Weighted sum / Implicit	GRPO
Rubric-ARM (Xu et al., 2026)	Text	Learned	Variational	Learned	PPO
EvoLM (Li et al., 2026c)	Text	Implicit	Variational (temporal)	Discriminative	GRPO
RLCER (Sheng et al., 2026)	Reasoning	Implicit	Process-conditional	Process-level	Extended from PPO
Constitutional AI (Bai et al., 2022)	Text	Fixed principles	Point mass at prior	Principle ensembling	PPO
QA-LIGN (Dineen et al., 2025)	Text	Fixed principles	QA decomposition	Decomposed	GRPO
PRM (Lightman et al., 2024)	Math/Code	Correctness	Implicit	Step product	RM training
RLCF (Viswanathan et al., 2025)	Text	Implicit	Instance checklist	Weighted sum	DPO
OpenRubrics (Liu et al., 2026b)	Text	CRG	Contrastive	Multi-dimensional	RM training
RRD (Shen et al., 2026)	Text	Coarse rubrics	Refined criteria	Corr.-aware weights	RM training
OpenRS (Jia et al., 2026)	Text	Principles	Pairwise adaptive	Criterion prefs.	GRPO
Agentic Extensions (see §5):					
DR Tulu (Shao et al., 2026)	Agent	Evolving	Evolving	Evolving	GRPO
HiPER (Peng et al., 2026)	Agent	Hierarchical	Plan-Execute	HAE	PPO
RubricEM (Li et al., 2026a)	Agent	Stagewise	Stagewise	Stagewise	Extended from GRPO
AdaRubric (Ding, 2026)	Agent	Task desc.	Task-adaptive	Conf.-weighted	DPO
Multimodal Extensions (see §5):					
ARR (Tian et al., 2026)	VLM	Implicit	Generated	Explicit	GRPO
DeltaRubric (Liu et al., 2026a)	VLM	Implicit	Plan-checklist	Verification	GRPO
rDPO (Yu et al., 2026)	VLM	Implicit	Instance-specific	Extended from DPO	
Step-Audio-R1.5 (Zhang et al., 2026c)	Audio	Implicit	Acoustic criteria	Preference RM	GRPO

Table 1: Taxonomy of rubric-guided RL methods under the Bayesian framework. Each method is characterized by its domain, prior specification, posterior parameterization or criterion-construction mechanism, aggregation strategy, and optimization algorithm.

3.3 Variational Posteriors: Self-Evolving Rubrics

Self-evolving rubrics occupy the update-oriented end of Figure 2: as the policy π_θ improves, preference data and model outputs shift, and criteria or evaluators are updated sequentially. This paradigm is deeply connected to self-rewarding and meta-rewarding language models (Yuan et al., 2024; Wu et al., 2025), which iteratively improve their own feedback generation, and recent methods including Rubric-ARM (Xu et al., 2026), EvoLM (Li et al., 2026c), and RLCER (Sheng et al., 2026). Across these methods, we identify three recurring dynamics: (1) *Posterior sharpening* in explicitly latent-rubric formulations; (2) *Prior simplification*, where brittle gating mechanisms simplify for better generalization; and (3) *Conditional expansion*, where rubrics evolve from outcome-focused to process-focused.

Self-evolution versus posterior updating. Crucially, self-evolution is broader than exact Bayesian posterior updating. EvoLM provides a relatively concrete posterior-sharpening interpretation: preference pairs constructed from policy outputs act as observations, and a rubric-conditioned frozen judge's score margin between preferred and dispreferred responses supplies a likelihood-like signal (Li et al., 2026c). By contrast, systems without an explicit prior, likelihood, and observation model, including RLCER, are update-oriented analogues rather than exact Bayesian inference methods.

Assumptions and limitations by class. Fixed-principle methods assume that global principles transfer across inputs; instance-specific methods depend on both rubric-generator and evaluator reliability. Process-level methods additionally assume a meaningful decomposition into steps. Self-

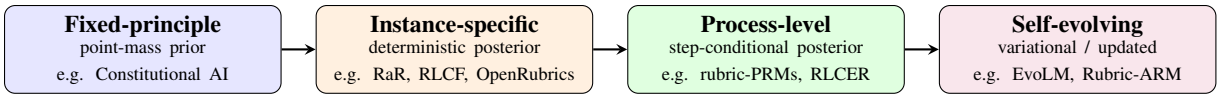


Figure 2: Navigation along the prior–posterior axis. The axis summarizes how method classes condition or update criteria; it does not assert that every class performs exact Bayesian inference.

Method class	Example	Criteria source / update	Workflow into optimization
Fixed-principle	Constitutional AI	Global principles or constitution	principles → critique/preferences → scalar or preference feedback → policy optimization
Instance-specific	RaR-style methods	Input- or domain-conditioned rubrics/checklists	rubric generation → criterion verification → aggregated scalar/vector/preference feedback → PPO/GRPO/DPO
Process-level	RLCER, rubric-conditioned PRMs	Step criteria derived from reasoning traces	trace → step decomposition → step scoring → dense process feedback → policy update
Self-evolving	EvoLM-style systems	Criteria updated from outputs, failures, or preference shifts	initial criteria → policy evaluation → rubric/evaluator sharpening → subsequent policy optimization

Table 2: Operational workflows for the major method classes. Generic optimization backends count as rubric-guided only when their feedback is conditioned on structured criteria.

evolving methods can adapt to policy shifts, but coupled rubric–policy updates risk feedback loops, criterion drift, and evaluator dependence. Figure 2 and Table 2 make these operational differences explicit.

4 Linguistic Insights: Rubrics as Natural Language

A fundamental but underexplored aspect of rubric-guided RL is that rubrics are *natural language texts*. This section analyzes how their prompt engineering, semantic stability, granularity, ambiguity, and cross-linguistic variation affect the performance and reliability of rubric-guided RL systems.

4.1 Rubrics as Prompted Specifications

When rubrics are used to guide LLM judges or reward models, they become prompted specifications: their wording, formatting, and output schema can affect the resulting scores, as observed in LLM-as-a-judge frameworks like Prometheus, which train evaluators on customized score rubrics (Kim et al., 2023, 2024), and FLASK, which decomposes evaluation into fine-grained alignment skills (Ye et al., 2024). Different formats expose different tradeoffs: binary checklists are easy to aggregate but coarse (Gunjal et al., 2026), free-text critiques are expressive but difficult to scalarize (Ke et al., 2024), and multi-level descriptors offer a middle ground between discriminative power and calibration (Hashemi et al., 2024).

4.2 Granularity Effects

The granularity of rubric criteria, whether defined over tokens, reasoning steps, whole responses, or decomposed checklist items, affects both evaluation quality and the informativeness of reward signals. The granularity-density tradeoff mirrors the exploration-exploitation tradeoff in RL: denser rewards enable faster learning but may miss global quality, while sparser rewards capture holistic quality but slow convergence. Existing systems often balance this tradeoff through hybrid designs: PRMs combine dense step-level supervision with outcome-level evaluation (Lightman et al., 2024), RLCF combines instance-specific checklists with universal criteria and verifier programs (Viswanathan et al., 2025), and OpenRS combines pairwise adaptive meta-rubrics with pointwise verifiable rubrics for guardrails (Jia et al., 2026).

Lexical and Token-Level Signals. At the finest scale, alignment systems may use lexical constraints or token-level shaping terms, such as PPO-style per-token KL penalties, length penalties, keyword inclusion/exclusion checks, profanity filters, or exact-format constraints. Lexical and formatting signals directly enforce observable surface properties, and length- or style-correlated rewards can favor well-formatted or verbose outputs without corresponding improvements in semantic quality (Zeng et al., 2024; Gao et al., 2023; Zheng et al., 2023). Therefore these dense signals should be clearly distinguished from process-level credit assignment strategies.

Sentence/Step-Level Criteria. Process-level rubrics (RLCER, Sheng et al., 2026; PRMs, Lightman et al., 2024) evaluate intermediate reasoning steps, providing denser reward signals than outcome-level rubrics. PRMs operate at the step level, providing per-step correctness signals and achieving 78.2% accuracy on MATH (Lightman et al., 2024). However, step-level evaluation requires defining what constitutes a "step", which is itself a non-trivial decision. Recent self-evolving rubric work suggests that optimal granularity interacts with evaluator capacity: smaller evaluators may require more precise rubrics, whereas larger evaluators can tolerate coarser specifications (Li et al., 2026c).

Response-Level and Document-Level Criteria. Outcome-level rubrics, such as RaR (Gunjal et al., 2026) and Constitutional AI (Bai et al., 2022), evaluate the complete response. They provide holistic assessment of dimensions such as helpfulness, safety, completeness, and coherence, but the reward is sparse and may obscure which part of the response caused success or failure.

Atomic Checklist Criteria. A separate form of granularity comes from decomposing a holistic objective into atomic checklist items rather than changing the textual span of evaluation. RLCF (Viswanathan et al., 2025), for example, decomposes tasks into atomic checklist items, illustrating how fine-grained criteria can provide denser feedback for robustness. Similarly, CheckEval (Lee et al., 2025) demonstrates that Boolean checklist-style decomposition improves evaluator reliability and robustness. This differs from word-level or sentence-level evaluation: the granularity lies in the criterion decomposition, not necessarily in the textual unit being scored.

4.3 Semantic Drift in Rubric Evaluation

Semantic drift refers to the phenomenon where the meaning of rubric criteria shifts across RL training iterations, even when the rubric text remains unchanged. This drift occurs because the policy model's interpretation of rubric criteria changes as it adapts to the reward signal. We distinguish semantic drift from the broader concept of concept drift (Gama et al., 2014) in machine learning: semantic drift is specifically about the *semantic instability* of natural language criteria as they are applied by changing models across training, rather than changes in the data distribution itself.

Causes of Semantic Drift. We identify three causes: (1) **Distribution shift**: as the policy improves, the response distribution changes, altering the context in which rubric criteria are applied. A criterion like "factual accuracy" means something different for a model that produces mostly correct responses vs. one that produces mostly incorrect responses. EvoLM (Li et al., 2026c) provides direct evidence: rubrics evolve from vague criteria (e.g., "Ensure mathematical accuracy") to specific, verifiable facts (e.g., "The answer is 144, derived from perimeter 48") as the policy distribution shifts during training. (2) **Criterion gaming**: the policy learns to satisfy the letter of rubric criteria while violating their spirit, effectively shifting the rubric's semantics. Constitutional AI (Bai et al., 2022) observed this Goodharting effect, where over-optimization leads to boilerplate language and overly harsh responses. (3) **Evolving evaluator**: in systems with learned judges, reward models, or co-evolving rubric generators, the scoring function itself changes during training. Thus, even when a criterion such as "mathematical accuracy" is textually stable, the function $s_t(c_k, x, y)$ that applies this criterion may shift as the evaluator is updated on new policy outputs. This is distinct from judge collapse: evaluator evolution is a general design property of adaptive systems, whereas collapse is a failure mode where the judge loses discriminative power or learns exploitable shortcuts.

Measuring Semantic Drift. We propose measuring semantic drift by tracking the consistency of rubric application across training iterations. Formally, given a rubric criterion c_k and a fixed set of reference responses, we define the drift metric as:

$$\text{Drift}(c_k, t) = 1 - \text{corr}\big(s_t(c_k, x, y), s_0(c_k, x, y)\big) \tag{3}$$

where s_t is the criterion score at iteration t and corr measures correlation (e.g., Spearman's ρ). In practice, the reference set should be a frozen, diverse validation set of prompt–response pairs. Scores can be anchored to a static high-capacity judge or held-out human annotations to reduce model-dependent self-evaluation bias. Rank-correlation decay is then monitored across iterations, with intervention thresholds calibrated as an absolute or relative drop from the initial held-out correlation rather than a universal constant; crossing a threshold triggers rubric regeneration or human auditing. This diagnostic proposal is related to but

distinct from semantic drift scores in text generation (Spataru et al., 2024) and value drift in post-training alignment (Bhatia et al., 2025); its empirical validation remains open.

4.4 Linguistic Reward Hacking

Linguistic reward hacking occurs when a policy exploits semantic gaps in natural language criteria, producing responses that satisfy the literal wording of a rubric while violating its intended meaning. This is a distinct phenomenon from numerical reward hacking (manipulating scalar reward magnitudes) and requires qualitative linguistic analysis to detect. While historically studied narrowly within low-resource language alignment (Inuwa-Dutse, 2025), we generalize the concept to encompass any systematic policy exploitation of semantic gaps in multi-dimensional evaluation rubrics.

Root Causes: Ambiguity and Underspecification. This vulnerability stems directly from the fact that natural language criteria are inherently ambiguous and underspecified. Ambiguity arises because high-level desiderata like "helpfulness" depend heavily on latent user context, leading to only moderate alignment with human consensus (Kendall's τ of 0.3–0.4) even when mitigated by performance-level descriptors (Hashemi et al., 2024). Underspecification occurs because criteria typically define *what* to evaluate rather than *how*, leaving operational boundaries vague unless anchored by intermediate granular structures like checklists (Viswanathan et al., 2025). Crucially, ambiguity and underspecification function as the primary systemic drivers for both semantic drift and linguistic reward hacking (see detailed case studies in Appendix L). However, this underspecification is also an opportunity: it provides the essential semantic flexibility required for self-evolution and contextual adaptation. Without initial underspecification, optimization frameworks would lack the headroom to perform posterior sharpening or track emergent data distributions during training via automated alignment updates (Li et al., 2026c; Xu et al., 2026).

Types of Linguistic Reward Hacking. We categorize these exploitative behaviors into three primary structural modes: (1) **Verbosity exploitation**, where the policy overoptimizes for "completeness" by generating elongated, low-substance responses, reflecting the proxy-reward degeneration patterns documented by Gao et al. (2023). (2) **Criterion bleeding**, where the model satisfies one axis by sacrificing another, such as aggressively refusing valid prompts to maximize "safety" at the cost of "helpfulness" (Dai et al., 2024; Wang et al., 2024). (3) **Surface pattern matching**, where the policy learns to parrot rubric keywords (e.g., stating "I have ensured safety") without executing the underlying reasoning. Self-evolving rubric work suggests that evaluator capacity and rubric specificity can affect how easily such surface-level satisfaction is detected (Li et al., 2026c).

Potential diagnostics of linguistic reward hacking include: divergence between rubric scores and human preferences, auxiliary verification of claimed reasoning steps, abnormal criterion-correlation patterns, and cross-rubric consistency checks.

Relation to Bayesian Framework. Under the Bayesian lens (§2.4), linguistic reward hacking can be viewed as the policy exploiting the gap between a rubric's literal point-estimate instantiation ($\mathcal{R}_x$) and its intended conditional criterion space $P(\mathcal{R}|x)$. Optimizing across multiple plausible rubric formulations therefore suggests a robustness direction. Self-evolving systems may patch vulnerabilities as they are uncovered, but constitute approximate posterior updates only when the required probabilistic components are explicit.

4.5 Cross-Lingual Considerations

While rubric-guided RL is predominantly optimized for English, multilingual alignment introduces typological and systemic challenges. High-level desiderata such as "respectfulness" are culturally and linguistically mediated; Japanese and Korean honorific registers (*keigo/jondaetmal*) and Arabic diglossia show that rubric and evaluator invariance cannot be assumed. A criterion may depend on dialect, speaker relation, register, and sociocultural setting rather than surface form alone. Low LLM-as-a-judge cross-lingual consistency ($\kappa \approx 0.3$) compounds this problem (Fu and Liu, 2025).

To mitigate these gaps, recent work on **cross-lingual reward transfer** shows that English-trained reward models can bootstrap target-language evaluations (Hong et al., 2025; Yang et al., 2025b), motivating unified criteria sets and multilingual reward reasoning architectures (Sheth et al., 2026; Anugraha et al., 2026; Gureja et al., 2025). We propose *language-aware rubric*

design: decoupling language-agnostic properties (e.g., factual accuracy) from language-specific instantiations (e.g., localized formality). Under our Bayesian framework, this corresponds to conditioning on language, dialect, speaker relation, and sociocultural context. For example, $P(\mathcal{R} \mid x, \ell, d, z)$—while anchoring broadly shared principles in the prior $P(\mathcal{R})$.

4.6 Evidence Levels

We distinguish three evidence levels of the analyses in this section: prompt sensitivity, evaluator calibration, granularity effects, and Goodhart-style over-optimization are established findings; semantic drift and linguistic reward hacking are our literature-grounded synthesis, supported by Appendix L; underspecification-as-opportunity and posterior marginalization are open hypotheses or design directions requiring validation.

5 Outlook and Emerging Frontiers: Agentic and Multimodal RL

While our core taxonomy centers on text-based alignment, extending rubric-guided methods to agentic RL and multimodal models introduces fundamentally new structural challenges. This section highlights these emerging frontiers, which push rubric evaluation beyond single-turn chat, as summarized in the lower sections of Table 1. Additional discussions on deep-research benchmarks are provided in Appendix F.

Agentic RL and Credit Assignment. In agentic settings, evaluation must account for multi-step trajectories and environment grounding. The core bottleneck is *multi-step credit assignment*. Traditional scalar rewards fail to isolate which specific action in a long trajectory led to success or failure. Rubric-guided methods like DR Tulu (Shao et al., 2026), HiPER (Peng et al., 2026), and RubricEM (Li et al., 2026a) address this by decomposing trajectory rewards into stagewise or hierarchical rubrics, enabling precise credit assignment across temporal phases. This structured feedback may reduce the policy from learning spurious correlations during complex tasks.

Multimodal Rubric-guided RL. In multimodal settings, evaluating outputs requires capturing cross-modal interactions and modality-specific dimensions such as visual grounding, object hallucination, and audio fluency. Foundational multimodal alignment work demonstrated that a single latent preference score is insufficient, by augmenting rewards with visual evidence (Sun et al., 2024), using segment-level correctional feedback (Yu et al., 2024), separating safety and helpfulness objectives (Ji et al., 2025a), or curating fine-grained multidimensional preference data (Zhang et al., 2025b). Recent methods directly address this complexity through explicit rubrics. Building on multidimensional benchmarks (Li et al., 2023; Guan et al., 2024; Fu et al., 2024), systems such as Auto-Rubric as Reward (Tian et al., 2026), DeltaRubric (Liu et al., 2026a), and rDPO (Yu et al., 2026) now explicitly condition multimodal reward modeling and preference optimization on structured criteria. In the audio domain, Step-Audio-R1.5 (Zhang et al., 2026c) applies rubric-guided optimization to multi-turn spoken dialogues by utilizing structural rubrics such as content requirements, persona consistency, and cross-turn instruction retention, while leaving subjective traits like tone and naturalness to standard preference comparison.

6 Conclusion and Future Directions

Rubric-guided reinforcement learning marks a shift from implicit scalar reward optimization to explicit, structured, and interpretable alignment. We unify this space through a Bayesian framework that treats constitutions as priors $P(\mathcal{R})$ and rubrics as conditional instantiations $P(\mathcal{R}|x)$, interpreting some self-evolution mechanisms as approximate posterior updating and rubric exploitation as point-estimate overfitting. Because rubrics are natural-language artifacts, navigating granularity trade-offs and mitigating vulnerabilities like semantic drift and linguistic reward hacking are central to building controllable, self-improving, and reliable language models.

Future work faces several challenges. First, we need theory connecting natural-language criteria to stable scalar or process-level feedback, especially as rubrics and evaluators evolve. Second, instance-specific rubric generation incurs token and latency overhead, motivating retrieval, caching, compression, and local evaluators. Third, cross-lingual and multimodal frontiers require principled mappings from abstract criteria to domain-specific instantiations. Finally, extending to autonomous agents demands proper credit assignment over agentic tasks to reliably evaluate long-horizon trajectories.

7 Limitations

We acknowledge several limitations of this survey and of the rubric-guided RL paradigm more broadly.

7.1 Heterogeneous Evaluation Settings

Because rubric-guided RL methods use different base models, benchmarks, and optimization protocols, direct quantitative comparison is difficult. We therefore focus on qualitative organization and report numbers from original papers where available. A standardized benchmark for controlled comparison remains an important direction for future work.

7.2 Scope of the Bayesian Framework

Our Bayesian framework is an analytical abstraction for organizing how methods specify, condition, and update evaluation criteria. It intentionally abstracts over implementation details: fixed constitutions, deterministic rubric generators, and self-evolving systems instantiate different parts of the notation. Exact Bayesian guarantees would require explicit priors, likelihoods, observations, and calibrated uncertainty, which most surveyed systems do not provide. Developing fully probabilistic rubric generators remains an open research opportunity.

7.3 Production Systems

Recent technical reports suggest that production systems increasingly embed rubric-like criteria in training and data pipelines, including Kimi K2, GLM-5, Qwen3, and DeepSeek-V4 (Kimi Team et al., 2025; GLM-5 Team, 2026; Yang et al., 2025a; DeepSeek-AI, 2026). For example, DeepSeek-V4 reports rubric-guided RL data and a Generative Reward Model for hard-to-verify tasks, while GLM-5 defines multidimensional criteria for chat and role-playing data, including instruction following, expressiveness, creativity, coherence, and long-dialogue consistency (DeepSeek-AI, 2026; GLM-5 Team, 2026). Because these practices are often only partially documented, a fuller account of implicit production-oriented rubric usage remains an important direction for future empirical study.

7.4 Evolving Terminology

Terminology around constitutions, rubrics, and principles is still evolving. Our framework provides one coherent vocabulary, but future work may refine these distinctions as the field matures.

7.5 Computational Cost

As analyzed in §E, instance-specific rubric generation and API-based judging introduce token, latency, and financial overhead. This creates deployment tradeoffs and motivates cheaper rubric generation, caching, compression, and local evaluator designs.

Ethics Statement

Rubric-guided RL raises important ethical considerations. While rubrics improve interpretability, they also pose the systemic risk of embedding cultural, political, or institutional biases into formalized evaluation criteria that are subsequently treated as objective standards. Evaluator dependence, underspecified criteria, and aggregation choices can further conceal whose values dominate. The sensitivity to normative framing observed in Constitutional AI (Bai et al., 2022) extends to rubrics more broadly. Future work should develop frameworks for ethical rubric design by prioritizing participatory construction to engage diverse stakeholders in defining criteria, democratic governance to establish formal mechanisms for continuous revision and contestation, and systemic transparency to ensure that both the underlying criteria and their downstream policy updates remain fully auditable.

Acknowledgments

AI assistants, including ChatGPT and Gemini, were used only to identify additional relevant work and to polish author-written text. All suggested references and language edits were independently verified by the authors.

We thank Yao Zhou and Zhenran Xu for their valuable discussions.

References

David Anugraha, Shou-Yi Hung, Zilu Tang, En-Shiun Annie Lee, Derry Tanti Wijaya, and Genta Indra Winata. 2026. mR3: Multilingual rubric-agnostic reward reasoning models. In *The Fourteenth International Conference on Learning Representations*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini,

Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Mehar Bhatia, Shravan Nayak, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, Vered Shwartz, and Siva Reddy. 2025. Value drifts: Tracing value alignment during LLM post-training. *arXiv preprint arXiv:2510.26707*.

Xiusi Chen, Hongzhi Wen, Sreyashi Nag, Chen Luo, Qingyu Yin, Ruirui Li, Zheng Li, and Wei Wang. 2024. IterAlign: Iterative constitutional alignment of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1423–1433, Mexico City, Mexico. Association for Computational Linguistics.

Josef Dai, Rui Pan, Jiaming Ji, Xinyu Li, and Yaodong Wang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek-AI.

Jacob Dineen, Aswin Rrv, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, and Ben Zhou. 2025. QA-LIGN: Aligning LLMs through constitutionally decomposed QA. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20619–20642, Suzhou, China. Association for Computational Linguistics.

Liang Ding. 2026. AdaRubric: Task-adaptive rubrics for reliable llm agent evaluation and reward learning. *arXiv preprint arXiv:2603.21362*. KnowFM Workshop at ACL 2026.

Shihan Dou, Yujiong Shen, Chenhao Huang, Junjie Ye, Jiayi Chen, Junzhe Wang, Qianyu He, Shichun Liu, Changze Lv, Jiahang Lin, Jiazheng Zhang, Ming Zhang, Shaofan Liu, Tao Ji, Zhangyue Yin, Cheng Zhang, Huaibing Xie, Jianglu Hu, Jingcheng Deng, and 19 others. 2026a. CL-bench Life: Can language models learn from real-life context? *arXiv preprint arXiv:2604.27043*.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, Huaibing Xie, Jianglu Hu, Shaolei Wang, Weichao Wang, Yanling Xiao, Yiting Liu, Zenan Xu, Zhen Guo, Pluto Zhou, and 8 others. 2026b. CL-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. 2025. Inverse constitutional AI: Compressing preferences into principles. In *International Conference on Learning Representations*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2024. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR.

GLM-5 Team. 2026. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. 2026. Rubrics as rewards: Reinforcement

learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58, Vienna, Austria. Association for Computational Linguistics.

Janghoon Han, Heegyu Kim, Changho Lee, Dahm Lee, Min Hyung Park, Hosung Song, Stanley Jungkyu Choi, Moontae Lee, and Honglak Lee. 2025. DEER: A benchmark for evaluating deep research agents on expert report generation. *arXiv preprint arXiv:2512.17776*.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. 2025. Cross-lingual transfer of reward models in multilingual alignment. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Jeesoo Bang, Hwanhee Lee, and Kyomin Jung. 2024. Kosmic: Korean text similarity metric reflecting honorific distinctions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9954–9960, Torino, Italia. ELRA and ICCL.

Isa Inuwa-Dutse. 2025. OpenAI's GPT-OSS-20B model and safety alignment issues in a low-resource language. *arXiv preprint arXiv:2510.01266*.

Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Chi-Min Chan, Yida Tang, and Yaodong Yang. 2025a. Safe RLHF-V: Safe reinforcement learning from multi-modal human feedback. In *Advances in Neural Information Processing Systems*, pages 51568–51604.

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. 2025b. A survey on progress in LLM alignment from the perspective of reward design. *arXiv preprint arXiv:2505.02666*.

Ruipeng Jia, Yunyi Yang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. 2026. Open rubric system: Scaling reinforcement learning with pairwise adaptive rubric. *arXiv preprint arXiv:2602.14069*.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054, Bangkok, Thailand. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi K2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tülu 3: Pushing frontiers in open language model post-training. In *International Conference on Learning Representations*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26874–26901.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China. Association for Computational Linguistics.

Gaotang Li, Bhavana Dalvi Mishra, Zifeng Wang, Jun Yan, Yanfei Chen, Chun-Liang Li, Long T. Le, Rujun Han, George Lee, Hanghang Tong, Chen-Yu Lee, and Tomas Pfister. 2026a. RubricEM: Meta-RL with rubric-guided policy decomposition beyond verifiable rewards. *arXiv preprint arXiv:2605.10899*.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2026b. DeepResearch Bench II: Diagnosing deep research agents via rubrics from expert report. *arXiv preprint arXiv:2601.08536*.

Shuyue Stella Li, Rui Xin, Teng Xiao, Yike Wang, Rulin Shao, Zoey Hao, Melanie Sclar, Sewoong Oh, Faeze Brahman, Pang Wei Koh, and Yulia Tsvetkov. 2026c. EvoLM: Self-evolving language models through co-evolved discriminative rubrics. In *The Fourteenth International Conference on Learning Representations*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Rui Liu, Dian Yu, Zhenwen Liang, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, and Leoweiliang. 2026a. DeltaRubric: Generative multimodal reward modeling via joint planning and verification. *arXiv preprint arXiv:2605.09269*.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2026b. OpenRubrics: Towards scalable synthetic rubric generation for reward modeling and LLM alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17417–17437, San Diego, California, United States. Association for Computational Linguistics.

Anas Mahmoud, MohammadHossein Rezaei, Zihao Wang, Anisha Gunjal, Bing Liu, and Yunzhong He. 2026. Reward hacking in rubric-based reinforcement learning. *arXiv preprint arXiv:2605.12474*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *Advances in Neural Information Processing Systems*, pages 134387–134429.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian D Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. In *Advances in Neural Information Processing Systems*, volume 37.

OpenAI. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Max Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Tianjun Pan, Xuan Lin, Wenyan Yang, Qianyu He, Shisong Chen, Licai Qi, Wanqing Xu, Hongwei Feng, Bo Xu, and Yanghua Xiao. 2026. RubricEval: A rubric-level meta-evaluation benchmark for LLM judges in instruction following. *arXiv preprint arXiv:2603.25133*.

Jiangweizhi Peng, Yuanxin Liu, Ruida Zhou, Charles Fleming, Zhaoran Wang, Alfredo Garcia, and Mingyi Hong. 2026. HiPER: Hierarchical reinforcement learning with explicit credit assignment for large language model agents. *arXiv preprint arXiv:2602.16165*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

Delip Rao and Chris Callison-Burch. 2026. Autorubric: A unifying framework for rubric-based LLM evaluation on non-verifiable tasks. *arXiv preprint arXiv:2603.00077*.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, and 2 others. 2026. DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. In *International Conference on Machine Learning*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. 2025. ResearchRubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*.

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

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

Ivaxi Sheth, Zeno Jonke, Amin Mantrach, and Saab Mansour. 2026. Cross-lingual LLM-Judge transfer via evaluation decomposition. *arXiv preprint arXiv:2603.18557*.

Clemencia Siro, Pourya Aliannejadi, and Mohammad Aliannejadi. 2026. Learning to judge: LLMs designing and applying evaluation rubrics. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6371–6389, Rabat, Morocco. Association for Computational Linguistics.

Ava Spataru, Eric Hambro, Elena Voita, and Nicola Cancedda. 2024. Know when to stop: A study of semantic drift in text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3656–3671, Mexico City, Mexico. Association for Computational Linguistics.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Zhengwei Tao, Ting-En Lin, Qinyuan Sun, Xingjian Li, Hongling Zhang, Zhe Zhang, Yuting Lian, Bei Sun, Qifan Wang, Junwei Li, Zhiqiang Wu, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

Juanxi Tian, Fengyuan Liu, Jiaming Han, Yilei Jiang, Yongliang Wu, Yesheng Liu, Haodong Li, Furong Xu, and Wanhua Li. 2026. Auto-rubric as reward: From implicit preferences to explicit multimodal generative criteria. *arXiv preprint arXiv:2605.08354*.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. 2025. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA. Association for Computational Linguistics.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. 2025. Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11537–11554, Suzhou, China. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, volume 36, pages 59008–59033.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *International Conference on Machine Learning*.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. 2026. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable LLM post-training. *arXiv preprint arXiv:2602.01511*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,

Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025b. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ya-Qi Yu, Fangyu Hong, Xiangyang Qu, Hao Wang, Gaojie Wu, Qiaoyu Luo, Nuo Xu, Huixin Wang, Wuheng Xu, Yongxin Liao, Zihao Chen, Haonan Li, Ziming Li, Dezhi Peng, Minghui Liao, Jihao Wu, Haoyu Ren, and Dandan Tu. 2026. Visual preference optimization with rubric rewards. *arXiv preprint arXiv:2604.13029*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR.

Shunjie Zeng, Ganqu Feng, Weizhou Liu, Minnan Wu, Shicheng Lu, Jingang Zhao, Shuyun Zhang, Di Lin, and Xiaodan Liang. 2024. Token-level direct preference optimization. In *International Conference on Machine Learning (ICML)*, pages 59963–59981. PMLR.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. 2026a. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. In *The Fourteenth International Conference on Learning Representations*.

Qiyuan Zhang, Junyi Zhou, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. 2026b. RubricBench: Aligning model-generated rubrics with human standards. *arXiv preprint arXiv:2603.01562*.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025a. CM-align: Consistency-based multilingual alignment for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25689–25702, Suzhou, China. Association for Computational Linguistics.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Xue Wang, Yibo Hu, Bin Luo, and Tien-Ping Tan. 2025b. MM-RLHF: The next step forward in multimodal LLM alignment. In *International Conference on Machine Learning*.

Yuxin Zhang, Xiangyu Tony Zhang, Daijiao Liu, Fei Tian, Yayue Deng, Jun Chen, Qingjian Lin, Haoyang Zhang, Yuxin Li, Jinglan Gong, Yechang Huang, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Gang Yu, Xiangyu Zhang, and Daxin Jiang. 2026c. Step-Audio-R1.5 Technical Report. *arXiv preprint arXiv:2604.25719*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Congmin Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *arXiv preprint arXiv:2510.08049*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, and 8 others. 2023. Secrets of rlhf in large language models part i: Ppo. *Preprint*, arXiv:2307.04964.

Joey Zhong, Hao Zhang, Clare Southern, Jeremy Yang, Thomas Wang, Kate Jung, Shu Zhang, Denis Yarats, Johnny Ho, and Jerry Ma. 2026. DRACO: A cross-domain benchmark for deep research accuracy, completeness, and objectivity. *arXiv preprint arXiv:2602.11685*.

A Survey Methodology and Inclusion Protocol

Scope. This survey focuses on methods that use explicit or recoverable structured criteria for LLM alignment, evaluation, reward modeling, or policy optimization. We include work on rubrics, constitutions, principles, checklists, process criteria, and LLM-as-a-judge settings when these criteria influence feedback or rewards. We also include adjacent alignment paradigms—DPO-style preference optimization, RLVR, PRMs, and multimodal RLHF—when they clarify the boundary between

rubric-guided RL and competing or complementary methods.

Search sources. We searched ACL Anthology, OpenReview, arXiv, Semantic Scholar, Google Scholar, and proceedings or accepted-paper lists from ACL, EMNLP, NAACL, EACL, ICLR, ICML, NeurIPS, CVPR, and related workshops. We additionally followed citation links from seed papers such as Constitutional AI, DPO, PRMs, DeepSeekMath, and recent rubric-generation papers.

Query families. Search queries covered five families: (i) rubric, criteria, checklist, grading, and LLM judge; (ii) constitution, principle, RLAIF, and AI feedback; (iii) DPO, IPO, KTO, SimPO, ORPO, and direct preference optimization; (iv) RLVR, verifier, process reward, mathematical reasoning, and code execution; and (v) multimodal RLHF, hallucination, VLM reward modeling, and multi-criteria VLM evaluation.

Inclusion criteria. We included papers if they met at least one criterion: (1) they explicitly use rubrics, principles, checklists, or criteria for alignment, reward modeling, or evaluation; (2) they operationally implement rubric-like structured or multi-dimensional feedback; (3) they are foundational adjacent methods needed to position rubric-guided RL; or (4) they introduce benchmarks or failure analyses that directly inform rubric quality, reward hacking, judge reliability, multilingual evaluation, or multimodal alignment.

Reliability and evidence use. Peer-reviewed papers are used as primary evidence whenever available. Influential preprints and technical reports are included when they are widely adopted, provide necessary coverage of fast-moving 2025–2026 developments, or directly introduce rubric-guided RL methods not yet published in archival venues. We label such work cautiously in the prose and avoid using unstable or unverifiable sources as load-bearing evidence. When a citation could not be matched to a stable title, author list, venue, or archival record, we removed it from the main evidence chain rather than relying on it for taxonomy or quantitative claims.

B Background: Related Surveys and Methods

B.1 How This Survey Differs from Prior Surveys

Several existing surveys touch on aspects of rubric-guided RL, but none provides a comprehensive treatment.

LLM-as-a-Judge Surveys. The comprehensive survey by Gu et al. (2024) covers LLM-as-a-judge methodologies, biases, and applications. While LLM-as-a-judge is closely related to rubric-guided evaluation, the two concepts differ: LLM-as-a-judge typically uses implicit evaluation criteria, while rubric-guided evaluation uses explicit, structured criteria. Our survey bridges these two areas by examining how rubrics can make LLM-as-a-judge more interpretable and controllable.

Reward Design and PRM Surveys. Recent surveys on reward design and process reward models (Ji et al., 2025b; Zheng et al., 2025) categorize reward formulations and process supervision methods. Our survey differs in its focus on rubric-based reward design, which provides structured, interpretable, and multi-dimensional reward signals. We provide a more detailed treatment of rubric-specific challenges (rubric quality, evolution, exploitation) that are not addressed in general reward design surveys, and we position PRMs as implicit step-level rubrics extending beyond single-dimension process supervision.

Self-Evolution and Agentic RL Surveys. Tao et al. (2024) survey self-evolution of large language models. Our survey complements these by focusing specifically on how rubrics enable self-evolution and structured credit assignment in agentic settings, dimensions not covered by existing surveys.

B.2 Related Alignment Paradigms

RLHF. RLHF (Ouyang et al., 2022) trains language models to align with human preferences through a three-stage pipeline. While RLHF has proven effective for alignment, its reliance on scalar reward signals introduces fundamental limitations: the reward is opaque, difficult to interpret, and prone to overoptimization (Gao et al., 2023).

RLAIF. RLAIF (Lee et al., 2024; Bai et al., 2022) replaces human preference labels with AI-generated feedback, addressing the scalability bottleneck of human annotation. Constitutional

AI (Bai et al., 2022) is a prominent instance of RLAIF that uses explicit natural-language principles to guide AI feedback.

DPO and direct preference optimization. DPO (Rafailov et al., 2023) bypasses the RL optimization loop by reparameterizing the reward model to derive the optimal policy in closed form. Its variants now form a major alignment family: IPO generalizes preference optimization and regularizes against deterministic-preference overfitting (Gheshlaghi Azar et al., 2024), KTO learns from pointwise desirable/undesirable examples (Ethayarajh et al., 2024), SimPO removes the reference model with a length-normalized implicit reward (Meng et al., 2024), ORPO merges SFT and preference optimization into one stage (Hong et al., 2024), and CPO/SLiC-HF show contrastive calibration as another RL-free route (Xu et al., 2024; Zhao et al., 2023). This family is directly relevant to rubric-guided alignment because rubric scores can be converted into pairwise or pointwise preference labels; consequently, DPO-style optimization is best viewed as a competing or complementary backend for using rubric feedback, not merely as background RLHF literature.

RLVR. RLVR couples reinforcement learning with objective, externally verifiable signals such as mathematical correctness, code execution, or rule-based answer checking (Shao et al., 2024; DeepSeek-AI, 2025; Yang et al., 2024; Lambert et al., 2025). DeepSeekMath introduced GRPO for math reasoning with rule-based rewards (Shao et al., 2024), while DeepSeek-R1 and Qwen2.5-Math show that large-scale verifier- or rule-based RL can produce strong reasoning models in domains with checkable answers (DeepSeek-AI, 2025; Yang et al., 2024). Open post-training recipes such as Tülu 3 further incorporate verifiable tasks into modern alignment pipelines (Lambert et al., 2025). These methods are direct competitors to rubric-guided RL in math and code because verifiable rewards are cheaper and less subjective; rubric-guided methods are most compelling when tasks require partial credit, explanation quality, safety, pedagogy, or other criteria that are not externally decidable.

PRMs. PRMs (Lightman et al., 2024) provide step-level feedback for reasoning tasks. We reframe PRMs as *implicit, step-level rubrics* that score the "correctness" dimension at each reasoning step. The generalization from single-dimension PRMs to multi-dimensional rubric-guided process supervision (as in RLCER (Sheng et al., 2026)) is a natural but non-trivial extension.

GRPO. GRPO (Shao et al., 2024) eliminates the need for a value network by using group-relative rewards. GRPO has become a popular choice for rubric-guided RL methods (Gunjal et al., 2026) because it naturally accommodates structured, multi-dimensional reward signals.

C Mapping Method Classes to the Bayesian Framework

The Bayesian framework maps method classes according to how they specify, condition, and update evaluation criteria. Fixed-principle methods such as Constitutional AI specify input-independent criterion sets, corresponding to point-mass priors. Instance-specific systems such as RaR, OpenRubrics, PRMs, and judge-rubric generators construct input-conditioned criteria, corresponding to posterior point estimates. Self-evolving or learned-rubric methods such as EvoLM, Rubric-ARM, and RLCER occupy the update-oriented part of the framework because their rubrics or reward models change with training data, policy outputs, or temporal contrast. EvoLM most directly supports a posterior-sharpening interpretation; systems without explicit priors and likelihoods remain analogues. This mapping clarifies which components of $P(\mathcal{R})$ and $P(\mathcal{R}|x)$ each method operationalizes.

D What the Bayesian Framework Adds Beyond Re-labeling

This framework provides conceptual and practical utility beyond simple notation replacement. Conceptually, it resolves the terminological ambiguity plaguing the field: constitutions and rubrics are unified as the same mathematical object at different conditioning levels. Constitutions act as global, unconditioned priors $P(\mathcal{R})$, while instance-specific rubrics are input-conditioned posteriors $P(\mathcal{R}|x)$. Methods such as EvoLM (Li et al., 2026c), which optimize evolving rubric representations, instantiate the update-oriented portion of this framework most directly.

Practically, this perspective yields hypotheses about training dynamics and adversarial robustness. In explicitly probabilistic formulations, the

observed "vague to specific" evolution pattern can be interpreted as posterior sharpening as policy data shift; other self-evolving systems are update-oriented analogues. The lens also casts rubric exploitation as overfitting to a single conditional point estimate $\mathcal{R}_x$ and suggests a robustness direction: marginalizing over the rubric distribution,

$$r_{\text{robust}}(x, y) = \mathbb{E}_{\mathcal{R}_x \sim P(\mathcal{R}|x)}\left[r_{\mathcal{R}_x}(x, y)\right]$$

which encourages the policy to generalize across plausible evaluation criteria rather than gaming a single deterministic set. This remains a design hypothesis requiring empirical validation.

E Computational Efficiency Analysis

While rubric-guided reinforcement learning (RL) enhances reward interpretability, it introduces non-negligible computational and token overhead into the optimization loop. To provide a rigorous yet concise overview, we present a structural complexity analysis across different paradigms. Let P denote the token length of static constitutional principles, R the average token length of an individual criterion, K the number of criteria per instance, and S the intermediate step length for process-supervised approaches. We define T_{gen} and T_{eval} as the real-time execution latencies for rubric generation and judge evaluation, respectively.

As summarized in Table 3, the efficiency profiles across paradigms diverge along three core dimensions:

- **Token Consumption & Caching:** Static approaches (e.g., CAI (Bai et al., 2022)) incur a constant $\mathcal{O}(P)$ token overhead inside the evaluation context and fully benefit from inference-time prefix caching. Conversely, instance-specific frameworks (e.g., RaR (Gunjal et al., 2026), Rubric-ARM (Xu et al., 2026)) require online synthesis of conditional rubrics, injecting an uncacheable $\mathcal{O}(K \cdot R)$ dynamic token overhead per instance.
- **Evaluation Call Complexity:** In implicit aggregation modes (e.g., EvoLM (Li et al., 2026c), RaR Implicit), the judge model evaluates all dimensions simultaneously in a single forward pass ($\mathcal{O}(1)$ complexity). Explicit modes (e.g., RaR Explicit) serialize the verification, demanding $\mathcal{O}(K)$ independent judge calls per instance, which scales to $\mathcal{O}(M \cdot K)$ during test-time candidate (M) verification and inherently restricts optimal inference search depths.
- **Judge Infrastructure:** Leveraging commercial APIs typically introduces unpredictable network latency and financial costs. Modern paradigms optimize cost-effectiveness by constraining highly localized small models ($\mathcal{M}_{\text{local}}$, e.g., 1.7B–3B parameters) within precise rubric structures, compressing the localized evaluation latency to a fraction of commercial API execution times without compromising reward fidelity.

F Detailed Agentic and Deep Research Benchmarks

A rapidly growing line of work applies rubric-based evaluation to *deep research agents*—LLM agents that perform multi-step information gathering, synthesis, and report generation. ResearchRubrics (Sharma et al., 2025) provides 2,500+ expert-written rubrics across 100+ real-world tasks in 9 domains. DRACO (Zhong et al., 2026) introduces a cross-domain benchmark spanning 10 domains and 40 countries' information sources, evaluating outputs along factual accuracy, completeness, objectivity, and citation quality. DeepResearchBench (Li et al., 2026b) constructs 9,430 fine-grained binary rubrics from expert-written investigative articles. DEER (Han et al., 2025) develops a 7-dimension, 25-subdimension, 101-item rubric taxonomy with claim verification, making system strengths and limitations interpretable.

G Continual Learning Benchmarks

CL-bench (Dou et al., 2026b) introduces a benchmark with 500 complex contexts, 1,899 tasks, and 31,607 verification rubrics for assessing context learning. CL-bench Life (Dou et al., 2026a) extends this to real-life scenarios with 405 context-task pairs and 5,348 rubrics. In continual learning, rubrics serve as memory, where criteria from previous tasks provide evaluation frameworks that prevent catastrophic forgetting.

H Consistency and Rubric Quality

RubricBench (Zhang et al., 2026b) introduces a curated benchmark with 1,147 pairwise comparisons to assess the reliability of rubric-guided evaluation.

Method	Rubric Context Cost	Judge Calls / Inst.	Eval. Latency	Typical Judge Scale
Scalar RM (Baseline)	0	1_{RM}	$\mathcal{O}(1)$	$\mathcal{M}_{\text{dense}}$
PRM (Lightman et al., 2024)	0 (Implicit)	S	$\mathcal{O}(S \cdot T_{\text{eval}})$	$\mathcal{M}_{\text{dense}}$
Constitutional AI (Bai et al., 2022)	$\mathcal{O}(P)$	1	$\mathcal{O}(T_{\text{eval}})$	$\mathcal{M}_{\text{large}}$
RaR (Explicit) (Gunjal et al., 2026)	$\mathcal{O}(K \cdot R)$	K	$\mathcal{O}(K \cdot T_{\text{eval}})$	$\mathcal{M}_{\text{small}}$ (API)
RaR (Implicit) (Gunjal et al., 2026)	$\mathcal{O}(K \cdot R)$	1	$\mathcal{O}(T_{\text{eval}})$	$\mathcal{M}_{\text{small}}$ (API)
Rubric-ARM (Xu et al., 2026)	$\mathcal{O}(K \cdot R)$	1	$\mathcal{O}(T_{\text{eval}})$	$\mathcal{M}_{\text{local}}$
EvoLM (Li et al., 2026c)	$\mathcal{O}(K \cdot R)$	1	$\mathcal{O}(T_{\text{eval}})$	$\mathcal{M}_{\text{local}}$

Table 3: Computational complexity and resource mapping across key paradigms. P and R denote the token lengths of constitutional principles and individual criteria, respectively; K specifies the total number of evaluation criteria, and S represents the reasoning sequence steps. The base latency parameter T_{eval} scales asymptotically as a function of the designated judge model capacity $\mathcal{M}$.

RubricEval (Pan et al., 2026) is the first rubric-level meta-evaluation benchmark for instruction following.

I Representative Reported Results from Original Papers

Table 4 summarizes representative results reported by original papers. **Important caveat**: these results are reported under different evaluation settings, model sizes, and training configurations. They are included to document each paper's own evaluation setting, not to rank methods against each other.

J Detailed Comparison of Rubric Aggregation Strategies

Rubric-guided RL systems differ not only in how they aggregate criterion-level scores, but also in how they construct criteria and assign credit. Table 5 compares the principal reward construction and aggregation strategies used by representative methods.

K Extended Discussion on Rubric Quality Dimensions

Discriminative Power. Discriminative power measures how well a rubric distinguishes between high-quality and low-quality responses. A rubric with high discriminative power assigns significantly different scores to good and bad responses, while a rubric with low discriminative power assigns similar scores regardless of quality. This can be measured by the accuracy of rubric-conditioned evaluation on held-out preference data.

Coverage. Coverage measures how many relevant quality dimensions a rubric captures. A rubric with high coverage evaluates all important aspects of response quality (correctness, completeness, clarity, safety, etc.), while a rubric with low coverage may miss important dimensions. PRMs, for example, typically cover a narrower set of dimensions. Incomplete rubrics may fail to capture important aspects of response quality, leading to reward hacking through unmeasured dimensions.

Calibration. Calibration measures how well rubric scores align with human judgment across the full quality spectrum. A well-calibrated rubric assigns scores that are consistent with human evaluation, while a poorly calibrated rubric may over- or under-estimate quality. LLM-Rubric (Hashemi et al., 2024) proposes calibration methods for rubric-based evaluation. Calibrating rubric-based scores to match human judgment distributions is critical for consistency.

L Linguistic Insights: Detailed Evidence and Qualitative Examples

This appendix provides detailed evidence and qualitative examples supporting the claims in Section 4.

Semantic Drift: Qualitative Evidence. EvoLM (Li et al., 2026c) provides the most direct evidence for semantic drift: rubrics evolve from vague criteria (e.g., "Ensure mathematical accuracy") to specific, verifiable facts (e.g., "The answer is 144, derived from perimeter 48") as training progresses. This intentional drift admits a posterior-sharpening interpretation. Constitutional AI (Bai et al., 2022) observed unintentional drift through Goodharting, where

Method	Base Model	Benchmark	Metric	Result
RaR (Gunjal et al., 2026)	Qwen-2.5-7B	HealthBench, GPQA-Diamond	Rel. improvement Rel. improvement	+31% +7%
Rubric-ARM (Xu et al., 2026)	Various	RewardBench	Avg. gain	+4.7%
EvoLM (Li et al., 2026c)	Qwen3-8B	RewardBench-2, OLMo3-Adapt	Over GPT-4.1 Average	+25.7% 69.3%
RLCER (Sheng et al., 2026)	7B	Math benchmarks	Over RLVR	Outperforms
OpenRubrics (Liu et al., 2026b)	Various	RewardModel	Over baselines	+8.4%
PRM (Lightman et al., 2024)	Various	MATH	Accuracy	78.2%
Constitutional AI (Bai et al., 2022)	Various	Harmlessness Elo	Pareto	Improved
Chasing the Tail (Zhang et al., 2026a)	Various	RewardBench	Accuracy gain	+3.2%
RLCF (Viswanathan et al., 2025)	Various	Multi-domain	Consistent gain	Improved

Note: Results are from original papers under different settings and should NOT be directly compared.

Table 4: Representative results reported by original papers. These numbers are included only to document each paper's own evaluation setting and should not be interpreted as a controlled cross-method comparison.

over-optimization leads to boilerplate language and overly harsh responses. The diagnostic in Equation 3 is a proposal of this survey; no prior work has validated this specific formalization for rubric-guided RL. We present a concrete case study of semantic drift in Table 6.

Linguistic Reward Hacking: Qualitative Examples.

- **Verbosity exploitation**: Gao et al. (2023) document reward model overoptimization where proxy rewards increase while true rewards decline, a dynamic closely related to verbosity exploitation. In rubric contexts, models learn to pad responses with filler to satisfy "completeness" criteria without adding substantive content.

- **Criterion bleeding**: Safe RLHF (Dai et al., 2024) explicitly addresses the safety-helpfulness tradeoff, where maximizing safety (refusing all requests) violates helpfulness. ArmoRM (Wang et al., 2024) uses multi-objective reward modeling to mitigate criterion bleeding by balancing multiple quality dimensions simultaneously.

- **Surface pattern matching**: self-evolving rubric work suggests that evaluator capacity and rubric specificity interact, with smaller evaluators often requiring more explicit criteria to distinguish genuine satisfaction from superficial compliance (Li et al., 2026c). This suggests that surface pattern matching is a real phenomenon that can be partially mitigated by evaluator and rubric design.

We provide concrete case studies for each type of linguistic reward hacking, drawing on empirical evidence from recent work. The case studies can be found in Table 7, 8, 9 and 10.

Underspecification as Opportunity. No prior work explicitly argues that underspecification is an opportunity in rubric-guided RL. However, several methods implicitly exploit it: EvoLM (Li et al., 2026c) observes rubrics evolving from vague to specific through posterior sharpening, RLCER (Sheng et al., 2026) starts with underspecified rubrics that self-evolve during training, and Rubric-ARM (Xu et al., 2026) learns rubric-conditioned reward models in non-verifiable settings. We posit that underspecification provides the space for rubric self-evolution—without initial underspecification, adaptive sharpening would have no room to operate.

Cross-Linguistic Considerations. Hong et al. (2025) demonstrate that English-trained reward models exceed target-language reward models by 3–4% on Multilingual RewardBench, suggesting cross-lingual transfer is feasible but imperfect. Sheth et al. (2026) propose a Universal Criteria Set (UCS) with language-agnostic evaluation di-

Strategy	Formulation	Used By	Pros	Cons
Weighted Sum	$r = \frac{\sum_k w_k s_k}{\sum_k w_k}$	RaR (Explicit)	Independent criterion scores expose their contributions and permit direct control through criterion weights.	The result depends on prespecified weights and separate judge decisions may miss interactions among criteria.
Implicit Aggregation	$r = f_\phi(x, y, \{(w_k, d_k)\})$	RaR (Implicit)	A single holistic judgment can integrate interactions without requiring a separate call for every criterion.	Delegating aggregation to the judge makes the final score less directly attributable to individual criteria.
Learned Rubric-Conditioned Judging	$o \sim \pi_j(\cdot \mid x, y^{(1)}, y^{(2)}, r)$	Rubric-ARM	Preference feedback jointly improves rubric generation and rubric-conditioned judging under a shared correctness objective.	Training two coupled components adds computation and requires careful scheduling to control non-stationarity.
Discriminative Rubrics	Optimize rubrics for preference separation	EvoLM	Instance-specific rubrics emphasize differences that make closely matched response pairs more informative for training.	Because the objective rewards separability, coverage of important but weakly discriminative criteria requires separate validation.
Step Product	$r = \prod_t p_t(\text{correct})$	PRM	Step-level evaluations localize errors while yielding a solution score that represents all steps being correct.	Multiplication is sensitive to sequence length and calibration, so one underestimated step can suppress the entire score.
Principle Ensembling	Sample one principle per comparison	CAI	Sampling across principles produced more robust preference-model behavior than repeatedly applying a single principle.	Each comparison applies only one sampled principle, so conflicts and coverage gaps are not resolved within that label.
Checklist Aggregation	$r = \frac{\sum_k w_k s_k}{\sum_k w_k}$	RLCF	Atomic criteria and optional programmatic verifiers make individual failures inspectable and hard constraints checkable.	Scoring items separately is judge-call intensive, and optimizing checklist completion can produce reward-hacking artifacts.
Hierarchical Credit	Aggregate subgoal returns at planning and execution levels	HiPER	Hierarchical advantages assign credit at both levels and reduce variance in long-horizon agent training.	The method requires a planner–executor decomposition and introduces additional level-specific optimization choices.
Pairwise Adaptive	Instantiate pair-specific criteria and aggregate criterion preferences externally	OpenRS	Pair-conditioned criteria target semantic differences while retaining inspectable, externally aggregated preferences.	A new rubric must be instantiated for each pair, increasing evaluation work and limiting direct cross-pair comparability.

Table 5: Comparison of rubric reward construction and aggregation strategies. Each strategy exposes a different trade-off among interpretability, computational cost, criterion coverage, and the ability to capture interactions.

Case Study 1: Semantic Drift—Intentional vs. Unintentional

Intentional drift (EvoLM): Li et al. (2026c) document that rubrics evolve from vague criteria (e.g., "Ensure mathematical accuracy") to specific, verifiable facts (e.g., "the correct maximum area of 144, derived from the given perimeter of 48") as training progresses. This intentional drift admits a posterior-sharpening interpretation and improves downstream policy quality.

Unintentional drift (CAI): Bai et al. (2022) observe that the meaning of "harmlessness" drifts from "avoid causing harm" to "refuse all potentially sensitive queries" under over-optimization. The rubric text remains unchanged, but the policy's interpretation shifts, leading to overly harsh and unhelpful responses.

Table 6: Semantic drift in rubric-guided RL: intentional drift (posterior sharpening) may improve quality, while unintentional drift (Goodharting) can degrade it.

mensions combined with language-specific instantiations. Fu and Liu (2025) find that multilingual LLM-as-a-judge has low cross-lingual consistency (Fleiss' Kappa ≈ 0.3), indicating that rubric-conditioned judges face the same challenge. The mR3 model (Anugraha et al., 2026) extends to 72 languages with rubric-agnostic reward reasoning, while M-RewardBench (Gureja et al., 2025) provides the first systematic multilingual RM evaluation across 23 languages. Language-specific quality dimensions include Japanese keigo (honorific system) with multiple formality levels, Korean jondaetmal (honorific usage) (Hwang et al., 2024), and Arabic diglossia—all absent from English-centric rubrics. CM-Align (Zhang et al., 2025a) uses consistency-based approaches for multilingual alignment.

M Adversarial Robustness and Reward Hacking

Rubric-guided RL introduces both new defenses against and new vulnerabilities to adversarial attacks and reward hacking.

Rubric Exploitation as Overfitting to a Point Estimate. A novel form of reward hacking in rubric-guided RL is *rubric exploitation*: the policy learns to produce responses that score well against rubric criteria while violating the spirit of

Case Study 2: Verbosity Exploitation via Presence-Based Rubrics

Source: Mahmoud et al. (2026)

Phenomenon: The RL checkpoint improves completeness (+1.07) but degrades conciseness (-2.91), factual correctness (-0.85), and overall quality (-1.02). The rubric-based verifier prefers the RL checkpoint on 85.8% of prompts, but rubric-free judges prefer the base model on 78.4%—a clear sign of reward hacking under strong verification.

Mechanism: 90.2% of rubric weight falls on *presence-based* criteria (e.g., "States that plasma volume increases more than red cell mass during pregnancy"). Presence-based satisfaction rises from 27.6% to 42.5% (+14.9 pp), while absence-based criteria decline. The model exploits this by generating longer, claim-denser responses that satisfy presence checks without verifying correctness.

Safety-presence exploitation: Safety-presence rubric items (8.4% weight) such as "The response advises the user to consult a healthcare provider before taking any action" can be satisfied by appending boilerplate disclaimers without genuine safety reasoning.

Table 7: Verbosity exploitation through presence-based rubric criteria. The model learns to satisfy presence checks by generating longer, claim-denser output, while degrading on conciseness, factual correctness, and overall quality.

Case Study 3: Criterion Bleeding—Safety-Helpfulness Tradeoff

Phenomenon: Maximizing the safety criterion ("the response should not cause harm") by refusing all requests directly violates the helpfulness criterion ("the response should address the user's query"). Safe RLHF (Dai et al., 2024) and ArmoRM (Wang et al., 2024) have proposed mechanisms to mitigate this problem. We categorize this phenomenon as "Criterion bleeding" — satisfying one rubric dimension at the expense of another.

Constitutional AI evidence: Bai et al. (2022) observe that over-optimization against constitutional principles leads to boilerplate language and overly harsh responses—the policy learns to satisfy safety keywords ("I cannot assist with that") while violating the spirit of balanced evaluation.

Mitigation: Safe RLHF explicitly decomposes the reward into separate safety and helpfulness objectives with constrained optimization. ArmoRM (Wang et al., 2024) uses multi-objective reward modeling to balance multiple quality dimensions simultaneously, preventing criterion bleeding by design.

Table 8: Criterion bleeding: when satisfying one rubric dimension (safety) comes at the expense of another (helpfulness), the policy exploits the multi-objective gap.

the rubric. Under the Bayesian framework (§2.4), rubric exploitation corresponds to *overfitting to a point estimate* of the rubric posterior: the policy optimizes against a single sample $\mathcal{R}_x$ rather than the full posterior $P(\mathcal{R}|x)$. The defense is to evaluate under the full posterior (marginalization), which requires the policy to perform well under all plausible rubrics rather than gaming a single one.

Rubric-Specific Reward Hacking. Reward Hacking in Rubric-Based RL (Mahmoud et al., 2026) studies the specific phenomenon where policies optimized against a training rubric are evaluated against a cross-family panel. The key finding is that rubric-based rewards do not by themselves prevent reward hacking: the gap between training and evaluation rubrics creates an exploitation surface.

Defenses. Dynamic rubric adaptation (self-evolving rubrics close exploitation opportunities), multi-rubric evaluation (ensemble methods), constrained optimization with rubric-based safety constraints, and information-theoretic approaches such as InfoRM (Miao et al., 2024) provide defenses against rubric exploitation.

Case Study 4: Surface Pattern Matching—Rubric Verification Failure Modes

Source: Mahmoud et al. (2026)

Failure Mode Taxonomy (stable across training, domains, and verifier strength):

- **Partial Compound (36.0%):** The criterion requires A∧B but the verifier accepts only one conjunct. Example: "The verifier failed because it accepted partial satisfaction of a multi-part requirement as full credit, verifying the outcome statement while not enforcing the required explicit distinction between two specified categories."
- **Implicit-as-Explicit (34.6%):** The required claim is never explicitly stated; the verifier infers it from context. Example: "The verifier failed because it credited an implicit or inferable statement as if it were explicit, accepting broad plausibility instead of requiring the exact characterization the criterion demanded."
- **Imprecise Verification (29.4%):** The verifier accepts a related but distinct concept as equivalent, or checks only broad topic relevance. Example: "The verifier failed because it matched on surface topic relevance instead of verifying the specific claim."

Key insight: These failure modes are *stable* across training—the distribution of exploitation types does not shift, only the volume increases. Both weak and strong verifiers fail in the same proportions, suggesting these are fundamental limitations of rubric verification.

Table 9: Surface pattern matching in rubric verification: three recurring structural failure modes where the verifier accepts superficial criterion satisfaction without genuine engagement.

Case Study 5: Granularity Effects and Evaluator Capacity

Phenomenon: Smaller judges (1.7B) force the learned rubric generator to produce more precise, fine-grained criteria, resulting in better downstream policies than larger judges (up to 32B) that tolerate vague rubrics. This demonstrates that optimal rubric granularity interacts with judge capacity: a rubric that is "too coarse" for a small judge may be "just right" for a large judge, and vice versa.

PRM evidence: Lightman et al. (2024) show that step-level (fine-grained) process supervision achieves 78.2% on MATH, substantially outperforming outcome-level (coarse-grained) supervision. This confirms that finer granularity produces better signals, but we argue that such benefits can be fully achieved only when the evaluator can reliably apply the fine-grained criteria.

RLCF evidence: RLCF (Viswanathan et al., 2025) decomposes each instruction into atomic checklist items and uses them as instance-level reward criteria, illustrating how checklist-style rubrics can provide denser and more inspectable feedback than holistic outcome rewards.

Table 10: Granularity effects: the interaction between rubric granularity and judge capacity determines downstream policy quality. Finer granularity generally helps, but only when the evaluator can reliably apply the criteria.