From Verification to Internalization: A Cognitive Science Perspective on Verifier-Free Reinforcement Learning

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has advanced Large Language Models (LLMs) reasoning by providing objective feedback, yet it remains fundamentally dependent on external verifiers, which limits self-regulated reasoning and generalization. We propose a shift toward *internalization*, relocating verification from external infrastructure into model-internal signals. We formalize this paradigm through a four-dimensional taxonomy: Probabilistic, Uncertainty, Process, and Interaction Internalization. This taxonomy captures how verifier-free reinforcement learning (VFRL) derives evaluative signals from likelihood, uncertainty, intermediate reasoning steps, and candidate interactions. This perspective enables dense, scalable, and model-driven supervision while highlighting characteristic failure modes such as proxy misalignment[1], miscalibration[2], local-process errors[3] and preference drift[4]. Our analysis systematizes recent VFRL methods, delineates their strengths and limitations, and outlines research directions for building reliable, auditable, and self-supervised reasoning agents.

1 Introduction

Recent advances in Large Language Models (LLMs) have increasingly motivated research on Reinforcement Learning with Verifiable Rewards (RLVR) to enhance reasoning capabilities(Wen et al., 2026b). Compared with Reinforcement Learning from Human Feedback (RLHF), which relies heavily on subjective human preferences(Kaufmann et al., 2023), RLVR introduces more objective feedback sources, such as code execution and formal proof verification(Hu et al., 2025; Fan et al.). These verifiable signals have improved the reliability and scalability of reinforcement learning for reasoning-oriented tasks. Nevertheless, RLVR remains fundamentally dependent on external verifiers: the model is optimized to satisfy externally supplied reward signals rather than to construct or regulate its own criteria of correctness.

This dependence creates several limitations. First, external verifiers are often domain-specific, computationally expensive, and difficult to generalize across open-ended tasks(Lightman et al., 2024). Second, optimizing against fixed external rewards can encourage shallow trial-and-error behavior, fragmented reasoning chains, or reward hacking, especially when the reward signal is only weakly aligned with the intended reasoning process(Lu et al., 2026; Ackermann et al., 2026). Third, because the verifier remains outside the model, the resulting system has limited capacity for autonomous self-evaluation and self-correction(Huang et al., 2024). In this sense, RLVR improves the reliability of external supervision but does not fully address the deeper question of how verification can become part of the model's own reasoning mechanism.

To address this issue, we draw on the cognitive science notion of internalization(Vygotsky and Cole, 1978). In this paper, internalization refers to the transformation of external verification, feedback, and error-correction procedures into intrinsic signals that can guide generation, reasoning, and self-regulation. This perspective aligns with the emerging field of Verifier-Free Reinforcement Learning (VFRL), where learning signals are derived not from task-specific external verifiers, but from model-internal quantities such as likelihood, uncertainty, reasoning trajectories, or interactions among candidate outputs. Rather than treating verifier-free learning as the simple removal of veri-

[1]Proxy misalignment occurs when VFRL optimizes an internal reward that only weakly tracks the true task objective.

[2]Miscalibration occurs when internal confidence, entropy, or consistency only weakly predicts empirical correctness.

[3]Local process error occurs when locally rewarded reasoning steps only weakly support a globally correct solution.

[4]Preference drift occurs when relative judgments among candidates only weakly remain grounded in true task utility.

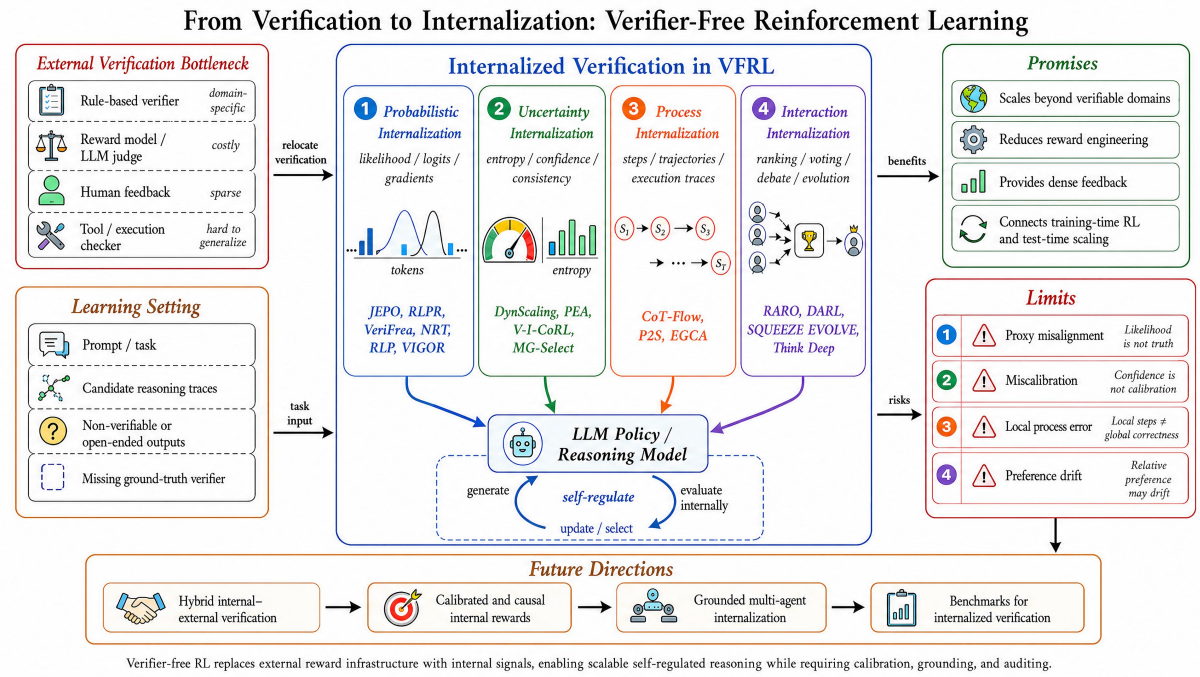

Figure 1: Framework of Verifier-Free Reinforcement Learning as Internalized Verification.

fication, we interpret it as a relocation of verification from external reward infrastructure to internal model-derived signals.

To organize this emerging research direction, we propose a Four-Dimensional Internalization Taxonomy. The taxonomy characterizes VFRL according to the type of internal signal used to replace or approximate external verification:

- **Probabilistic Internalization**: replacing binary external correctness with likelihood, log-probability, or gradient-based signals derived from the model's generative distribution.
- **Uncertainty Internalization**: using confidence, entropy, calibration, or consistency as internal signals for sample selection, curriculum design, or compute allocation.
- **Process Internalization**: transforming outcome-level verification into step-level or trajectory-level signals that evaluate intermediate reasoning processes.
- **Interaction Internalization**: deriving evaluative signals from ranking, voting, debate, adversarial comparison, or multi-agent interaction among candidate solutions.

As illustrated in Figure 1, the paper is organized around a progressive shift from external verification to internalized evaluation in VFRL.

2 Background

2.1 The Bottlenecks of External Verification

RLVR can be understood as a **generator–verifier** framework. The language model serves as the generator, producing candidate behaviors, while external verifiers, reward models, or rule-based checkers assign rewards (Wen et al., 2026b). This separation improves reliability compared with RLHF (Kaufmann et al., 2023), which relies on subjective human preferences. Nevertheless, it introduces structural dependencies: the optimization trajectory of the model is constrained by the availability, cost, coverage, and bias of the verifier.

2.2 Verifier-Free Reinforcement Learning

As illustrated in Figure 2, VFRL addresses the limitations of external verification[5] by **deriving evaluative signals directly from the model's internal generative and reasoning processes**, rather than relying entirely on external judges or reward infrastructures. Such signals may include likelihood, log-probability, entropy, confidence, reasoning-path statistics, and interactions among multiple candidate solutions. In this paradigm, verification is

[5]In this paper, External verification refers to independent correctness or reward oracles (e.g., symbolic verifiers, execution-based checkers, or reward models). References, retrieved evidence, demonstrations, or tool outputs are not treated as external verification when they only provide contextual grounding for internally derived evaluation signals.

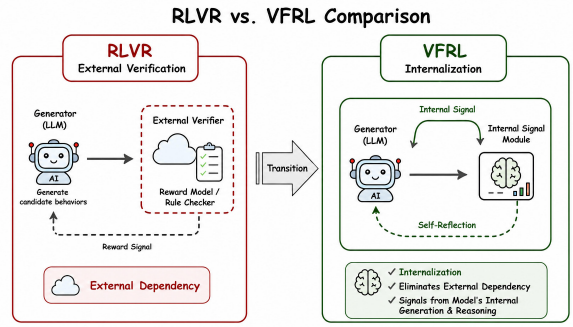

Figure 2: Comparison between RLVR and VFRL verification mechanisms.

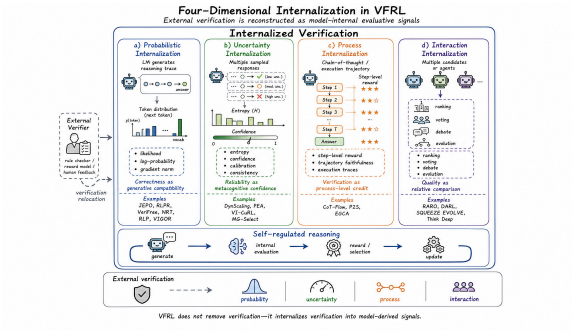

Figure 3: **Four-dimensional internalization of verification in VFRL.** For readability, we show only a subset of representative examples in the figure; detailed paper-level descriptions and coding rationales are provided in Appendix A.1.

Table 1: **Correspondence between cognitive internalization and VFRL mechanisms.**

Internalization	VFRL Counterpart
External guidance	*External rewards or verifier*
Internalization	*Internalized training signals*
Mediational signals	*CoT, likelihood, confidence, etc.*
Self-regulation	*Self-correction and adaptive control*

progressively shifted **from external reward infrastructures toward model-internal or model-derived signals**, enabling a more self-regulated and scalable reinforcement learning framework.

2.3 Internalization in Cognitive Science

This internalization process aligns with cognitive science principles (Vygotsky and Cole, 1978; Miller, 2000), where **externally guided actions or social feedback are transformed into intrinsic regulatory mechanisms**. Analogously, in VFRL, external verification and error correction are progressively reconstructed as internal signals guiding generation, selection, and self-correction. In Table 1, we provide a mapping between VFRL and internalization theory.

3 The Four-Dimensional Internalization

As shown in Figure 3, VFRL shifts verification from external correctness signals to internally derived training signals. Instead of relying on task-specific verifiers, reward models, or human feedback, it leverages signals already present in the model's likelihood, uncertainty, reasoning trajectory, or candidate interactions. Cognitively, this resembles internalization: external evaluative functions are progressively reconstructed as internal control mechanisms. Recent work can therefore be organized into four complementary dimensions, namely probabilistic, uncertainty, process, and interaction internalization, each capturing a distinct mode of internal verification. Table 2 further summarizes the signals and mechanisms across these dimensions.

3.1 Probabilistic Internalization

Core Concept. Probabilistic internalization replaces external verification with signals derived from the model's own likelihood landscape. Instead of optimizing against an external reward, these methods construct an **intrinsic reward** from the model-induced conditional distribution:

$$r_{\text{int}}(x, z, a^\star) = f(\pi_\theta(a^\star \mid x, z)) \,. \qquad (1)$$

where x is the prompt, z a reasoning trace, $a^\star$ a target answer, π_θ the policy distribution parameterized by θ, and f a *probability-to-reward transformation*[6].

Cognitively, this corresponds to an **internal compatibility judgment** (Topolinski and Strack, 2009; Koriat and Sorka, 2015): the model approximates correctness by assessing whether a reasoning path makes the target continuation more predictable.

Representative Mechanisms. Reference-based methods reward reasoning traces that increase the likelihood of a target answer, including JEPO (Tang et al., 2026), Native Reasoning Training (Wang et al., 2026), VeriFree (Zhou et al., 2025), RLPR (Yu et al., 2025), and Likelihood-Based Reward Designs (Kwiatkowski et al., 2026). A common instantiation is the reference-answer

[6]The transformation f may instantiate log-probability, normalized likelihood, distribution sharpening, predictive gain, or gradient-derived compatibility.

Table 2: **Summary of the four-dimensional internalization paradigm.** Internalized signals and corresponding mechanisms for each dimension.

Feature	*Internalized Mechanism*	*Internalized Signal*
Probabilistic	Estimates correctness via probability	Likelihood, logits, log-probability, gradient norm
Uncertainty	Guides reasoning via uncertainty	Entropy, confidence, calibration, consistency
Process	Rewards intermediate reasoning steps	Step-level reward, trajectory faithfulness, execution traces
Interaction	Evaluates quality via comparison	Ranking, voting, debate, adversarial comparison, evolution

log-probability reward:

$$r_{\log p}(x, z, a^\star) = \sum_{t=1}^{|a^\star|} \log \pi_\theta \left(a_t^\star \mid x, z, a_{<t}^\star\right). \quad (2)$$

where $a_t^\star$ denotes the t-th token of the reference answer, $a_{<t}^\star$ denotes its prefix before position t, and $|a^\star|$ is the answer length.

This replaces hard verification with a likelihood-based compatibility test: a reasoning trace is rewarded when it makes the target answer more probable under the current policy.

Probabilistic internalization also extends beyond reference-answer scoring. RLP (Hatamizadeh et al., 2025) treats chain-of-thought as an exploratory action before next-token prediction and rewards the predictive gain induced by the reasoning trace:

$$\begin{aligned} r_{\text{gain}}(x, z, y_t) = &\log \pi_\theta \left(y_t \mid x, z, y_{<t}\right) \\ &- \log \pi_\theta \left(y_t \mid x, y_{<t}\right). \end{aligned} \quad (3)$$

where r_{gain} measures the predictive utility of the generated thought. y_t is the t-th future token, $y_{<t}$ is its preceding context, and the two terms compare next-token log-likelihood with and without the reasoning trace z.

This measures the predictive utility of z by comparing next-token likelihood with and without the reasoning trace. Other methods instantiate the same principle through proxy likelihoods, distribution sharpening, near-policy targets, or gradient-norm smoothness (Liu et al., 2025; Ji et al., 2026; Khan et al., 2026; Wen et al., 2026a).

Discussion. Probabilistic internalization is general because likelihoods, logits, gradients, and distributional transformations are readily available in autoregressive models. Its main risk is **proxy misalignment**: probability may reflect imitation, majority patterns, spurious correlations, or model bias rather than semantic correctness. It is therefore most reliable when likelihood is used comparatively, anchored by trustworthy references, or complemented by uncertainty, process, or interaction-based signals.

3.2 Uncertainty Internalization

Core Concept. Uncertainty internalization uses the model's own confidence, entropy, consistency, or distributional dispersion to guide training or inference. Instead of asking whether an answer is externally correct, these methods ask whether the model is **internally decisive, stable, or uncertain**. Formally, an uncertainty-based internal signal can be written as

$$u_\theta(x, z) = \mathcal{U}(\pi_\theta(\cdot \mid x, z)). \quad (4)$$

where z is a reasoning trace, action sequence, or partial output, and $\mathcal{U}$ denotes an *uncertainty functional*[7].

Cognitively, this resembles **metacognitive monitoring** (Flavell, 1979; Livingston, 2003): the model regulates sampling, learning, and compute allocation using its own uncertainty state.

Representative Mechanisms. At inference time, DynScaling (Wang et al., 2025a) allocates additional sampling budget to queries whose previous responses indicate higher uncertainty. MG-Select (Jang et al., 2025) applies a related idea to Vision-Language-Action models by comparing a conditional action distribution with a high-uncertainty reference distribution obtained through condition masking:

$$C(z) = \sum_{t=1}^{T} D_{\text{KL}} \left(p_t^{\text{cond}} \parallel p_t^{\text{mask}}\right). \quad (5)$$

where $z = (z_1, \ldots, z_T)$, and p_t^{cond} and p_t^{mask} denote the conditional and masked token distributions at step t. Thus, uncertainty becomes an operational signal for Best-of-N selection and adaptive budget allocation.

[7]The uncertainty functional may be instantiated as entropy, confidence, response dispersion, agreement rate, or divergence from an uncertainty-aware reference distribution.

At training time, VI-CuRL (Cai and Sugiyama, 2026) uses intrinsic confidence, defined as the complement of normalized generation entropy, to prioritize high-confidence samples in early RL phases. Let $s_t = (x, z_{<t})$. A simplified confidence score is

$$c(x) = 1 - \mathbb{E}_{z \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{T} \sum_{t=1}^{T} \frac{H(\pi_{\theta_{\text{old}}}(\cdot \mid s_t))}{\log |V|} \right]. \tag{6}$$

where $H(\cdot)/\log|V|$ is normalized token entropy, so higher $c(x)$ indicates that the old policy is more decisive along the sampled trajectory.

Prototype Entropy Alignment (Pan et al., 2026) further aligns model uncertainty with expert-derived entropy signatures rather than minimizing entropy uniformly.

Discussion. Uncertainty internalization improves efficiency by allocating computation or learning to regions where confidence or uncertainty is informative. Its main risk is **miscalibration**: models can be confidently wrong, and entropy may reflect distributional artifacts rather than epistemic competence. It is therefore most reliable when uncertainty signals are calibrated, used comparatively, or combined with external checks in regimes where verifier-free scaling may be suboptimal (Setlur et al., 2025).

3.3 Process Internalization

Core Concept. Process internalization shifts the evaluative signal from the final answer to **the intermediate reasoning trajectory**. Rather than assigning a single sparse reward to an entire response, these methods decompose reasoning into steps, spans, or state transitions and estimate which parts of the process contribute to progress or failure. Let $z = (z_1, \ldots, z_T)$ denote a reasoning trajectory and $I_t = (x, z_{\leq t})$ the reasoning state after step t. A generic process-level internal reward can be written as

$$\begin{aligned} R_{\text{proc}}(x, z, a^\star) &= \sum_{t=1}^{T} \alpha_t r_t, \\ r_t &= \Phi(I_t, a^\star) - \Phi(I_{t-1}, a^\star). \end{aligned} \tag{7}$$

where α_t denotes the weighting coefficient for the reward at reasoning step t, r_t denotes the local process reward at reasoning step t, Φ denotes a *process-progress functional*[8].

[8]The process-progress functional may measure answer reachability, suffix likelihood, execution consistency, semantic adequacy, or other step-level progress signals.

Cognitively, this resembles **process monitoring** (Botvinick et al., 2001; Evans and Stanovich, 2013): the model evaluates intermediate operations rather than only final outcomes.

Representative Mechanisms. CoT-Flow (Liu et al., 2026) models reasoning as a probabilistic flow and measures the information gain contributed by each step toward the target answer:

$$r_t^{\text{flow}} = \log p_\theta(a^\star \mid I_t) - \log p_\theta(a^\star \mid I_{t-1}). \tag{8}$$

where $a^\star$ is the target answer, I_t is the reasoning state after step t, and r_t^{flow} denotes the stepwise log-likelihood gain toward $a^\star$[9].

A positive value indicates that the current step makes the target answer more reachable, while a near-zero or negative value suggests redundancy or deviation. P2S (Zhong et al., 2026) similarly scores each prefix by the conditional probability of completing the remaining gold-CoT suffix, making probability a local process signal rather than only a final-answer reward.

EGCA (Kumar et al., 2026) provides a complementary execution-grounded variant. It executes both a candidate program and a reference implementation, identifies the earliest semantic divergence, and assigns advantage only to the responsible token span. This converts coarse outcome feedback into localized process credit without requiring a dense external verifier. Survey work on verification design for test-time scaling (Venktesh et al., 2025) further clarifies this distinction between outcome verification and process verification.

Discussion. Process internalization reduces reward sparsity and improves credit assignment for long reasoning trajectories. Its main risk is **local reliability**: locally useful steps may not support global correctness, and process rewards can inherit biases from synthesized traces, execution instruments, or suffix-likelihood proxies. It is therefore most reliable when local credit is grounded in downstream reachability, execution behavior, or other evidence that links intermediate steps to final task success.

3.4 Interaction Internalization

Core Concept. Interaction internalization relaxes absolute verification into relational evaluation over multiple outputs, agents, demonstrations, or

[9]Positive values indicate progress; near-zero or negative values indicate redundancy or deviation.

candidate populations. Instead of evaluating an isolated response against a fixed verifier, these methods derive signals from **comparison, disagreement, aggregation, competition, or iterative refinement**. Formally, an interaction-based internal signal can be written as

$$r_{\mathrm{rel}}(x, y_i) = \Psi(y_i; \mathcal{Y}_{-i}, \mathcal{R}, x)\,, \tag{9}$$

where y_i is a candidate response, $\mathcal{Y}_{-i}$ denotes alternatives, $\mathcal{R}$ denotes references, demonstrations, or population memory, and Ψ is a *relational scoring function*[10].

Cognitively, this resembles **social or dialogic reasoning** (Kuhn, 1993; Mercier and Sperber, 2011): evaluation emerges through comparison, disagreement, aggregation, and revision.

Representative Mechanisms. RARO (Cai and Provilkov, 2025) instantiates this principle through adversarial imitation, where a relativistic critic compares policy outputs with expert demonstrations rather than assigning absolute correctness. Reference-guided alignment similarly replaces hard verification with reference-grounded preference comparison in non-verifiable domains: LLM judges compare candidate outputs with the aid of reference responses, construct preference data, and update the policy through DPO-style optimization (Shi et al., 2026).

Voting-based inference instead aggregates multiple sampled answers. To avoid reducing voting to exact string matching, let $\mathcal{C}$ denote answer clusters obtained through normalization or semantic grouping. Majority voting selects

$$c^\star = \arg\max_{c \in \mathcal{C}} \sum_{i=1}^{N} \mathbf{1}[y_i \in c], \qquad y^\star \in c^\star. \tag{10}$$

where $y^\star$ is selected from the winning cluster $c^\star$.

This relational view also covers majority-vote test-time scaling, evolutionary recombination, multi-model orchestration, and reference-constrained diversity optimization (Wang et al., 2025b; Maheswaran et al., 2026; Huang et al., 2026).

Discussion. Interaction internalization supports open-ended tasks by exploiting comparison, disagreement, diversity, and aggregation when no single verifier fully captures quality. Its main risk is **preference drift**: pairwise judges, majority votes, debates, and evolutionary loops may reinforce their own internal biases rather than true task utility. It is therefore most reliable when relational signals are anchored by high-quality references, diversity-preserving sampling, independent candidate generation, or sparse external validation.

Appendix A.2 presents a method-level functional mapping, and Appendix A.3 summarizes the coding protocol and decision rules.

[10]The relational scoring function may instantiate pairwise preference, adversarial discrimination, consensus aggregation, evolutionary selection, or controlled deviation from references.

4 Promises and Limits of Internalization

The four-dimensional taxonomy clarifies how VFRL replaces external verification with model-internal signals. This shift is more than a workaround for missing verifiers; it relocates supervision and redefines evaluative feedback. Rather than receiving correctness signals from rule-based checkers, reward models, human annotators, or LLM judges, VFRL derives optimization signals from the model's likelihood landscape, uncertainty structure, reasoning process, or interactions among candidate solutions. While this internalization provides significant benefits, it introduces a new class of potential failure modes. The central trade-off is therefore not "verification versus no verification," but external verification versus internal proxy-based verification. Table 3 summarizes this promise–limit coupling across the four internalization dimensions.

4.1 Strengths of Verifier-Free Internalization

Verifier-free internalization offers four major advantages, corresponding to the four internalization dimensions introduced earlier. Probabilistic internalization supports scalability by reusing likelihood, logits, or gradient-derived signals that are already available from the policy model. This allows efficient training across numerous tasks without requiring task-specific verifiers.

Uncertainty internalization enhances efficiency by leveraging confidence, entropy, or calibration signals to guide sample selection and allocate computational resources. Models can prioritize informative reasoning paths and reduce unnecessary computation during both training and inference.

Process internalization provides denser feedback and improved credit assignment by decomposing reasoning trajectories into intermediate steps, token spans, or execution traces. This ensures that intermediate contributions receive targeted learning

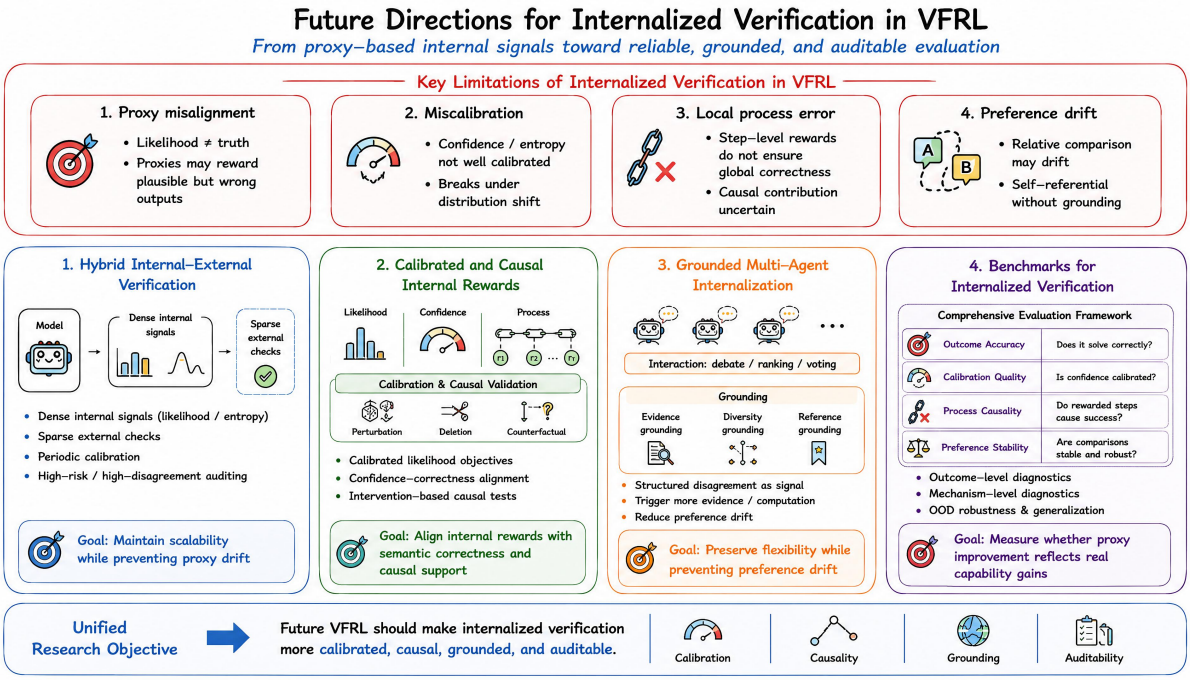

Figure 4: **Future directions for internalized verification in VFRL.**

Table 3: Strengths and limitations of internalization dimensions.

Dimension	*Strength*	*Limitation*
Probabilistic	Scalable	Likelihood is not truth
Uncertainty	Efficient	Confidence is not calibration
Process	Denser feedback	Local steps do not ensure global correctness
Interaction	Supports open-ended tasks	Relative preference may drift

signals, improving convergence and generalization.

Interaction internalization enables evaluation through candidate comparisons, ranking, debate, or multi-agent interaction. By assessing relative quality rather than absolute correctness, models can learn flexible strategies for open-ended tasks while maintaining robust internal evaluation.

4.2 Limits of Verifier-Free Internalization

Despite its advantages, internalization introduces four distinct limitations. First, probabilistic signals are vulnerable to **proxy misalignment**: high likelihood can correspond to plausible but incorrect outputs rather than truth. Second, uncertainty signals can suffer from **miscalibration**, especially under distribution shift, even when they are useful for resource allocation.

Third, process-level signals may induce **local-process errors**: dense step-level feedback can reward internally coherent steps that do not lead to correct overall outcomes. Fourth, interaction-based internalization may exhibit **preference drift**, where relative preferences among candidates gradually lose grounding without references, diversity constraints, or external anchoring.

Collectively, these limitations suggest that reliable VFRL should not depend on a single internal signal. Instead, robust internalized verification requires combining complementary signals and, when necessary, sparse external checks. This motivates future work on calibration, causal attribution, grounding, and benchmarks for evaluating whether internal signals remain aligned with true task utility.

5 Future Directions

Verifier-free internalization in VFRL offers scalability, dense feedback, and open-ended applicability by replacing external verification with internal proxies. However, this shift introduces characteristic risks: likelihood may diverge from truth, confidence may be miscalibrated, local process rewards

may not guarantee global correctness, and relative preferences may drift. These limitations do not undermine VFRL, but indicate that the next stage should focus on reliability, calibration, causal reasoning, and auditability. As shown in Figure 4, we outline four research directions that respond to these limitations at different levels. A more explicit mapping between the four future directions and the corresponding limitations is provided in Appendix A.4.

5.1 Hybrid Internal–External Verification

Internal proxies can drift when unanchored. Probabilistic signals may reinforce plausible but false answers, confidence may be overconfident under distribution shift, process rewards may optimize locally convincing steps, and interaction-based comparison may become self-referential. Hybrid verification addresses this by combining dense internal signals with sparse external checks. Likelihood- or entropy-based rewards guide most training, while external validation periodically calibrates internal rewards. Similarly, interaction-based inference can use voting or debate internally, with external verification for high-risk or high-disagreement cases. This maintains VFRL's scalability while preventing proxy drift.

5.2 Calibrated and Causal Internal Rewards

Internal rewards must align with the target property. For probabilistic internalization, calibrated likelihood objectives distinguish plausible outputs from semantically correct ones, including under out-of-distribution prompts. For uncertainty internalization, confidence and entropy should be empirically calibrated to predict correctness, not merely serve as unexamined proxies. For process internalization, rewards should reflect causal contributions, evaluated through interventions such as perturbations, deletions, or counterfactual reasoning steps. These measures ensure intermediate rewards meaningfully support final outcomes.

5.3 Grounded Multi-Agent Internalization

Interaction-based VFRL risks preference drift. Grounded multi-agent internalization preserves flexibility while anchoring comparisons. Approaches include evidence grounding (requiring retrieved evidence or tools), diversity grounding (maintaining multiple reasoning styles or solution paths), and reference grounding (anchoring judgments to high-quality demonstrations or sparse external feedback). Disagreement among agents can serve as a diagnostic signal, prompting additional evidence, computation, or auditing. This transforms interaction into a structured process for uncertainty resolution.

5.4 Benchmarks for Internalized Verification

Current benchmarks evaluating only final outcomes are insufficient. Effective evaluation should distinguish improvements in internal proxies from true capability gains. For probabilistic methods, assess whether likelihood-based rewards improve reasoning accuracy and out-of-distribution robustness. For uncertainty-based methods, measure calibration and confidence-accuracy correlation. For process-based methods, evaluate causal contribution of rewarded steps. For interaction-based methods, examine preference stability, consensus bias, and robustness to stylistic artifacts. A comprehensive framework should combine outcome-level and mechanism-level diagnostics to ensure internalized signals remain trustworthy approximations of reasoning quality.

6 Conclusion

This paper has argued that verifier-free reinforcement learning is best understood not as the removal of verification, but as its internalization into model-derived signals. Drawing on the cognitive-science notion of internalization, we introduced a four-dimensional taxonomy, namely probabilistic, uncertainty, process, and interaction internalization, to organize recent VFRL methods according to how they replace or approximate external evaluative functions.

This perspective highlights both the promise and the limits of VFRL. Internalized verification can expand RL beyond strictly verifiable domains, reduce dependence on external reward infrastructure, provide denser feedback, and connect training-time optimization with test-time scaling. However, these gains also introduce risks such as proxy misalignment, miscalibration, local-process errors, and preference drift. The central challenge for future VFRL research is therefore to make internal signals not only useful for optimization, but also calibrated, causal, grounded, and auditable as approximations of reasoning quality.

Limitations

This paper provides a conceptual and taxonomic analysis of VFRL, rather than proposing a new algorithm or empirical benchmark. Its primary contribution is therefore organizational: we reinterpret recent VFRL methods through the lens of cognitive internalization and classify them into probabilistic, uncertainty, process, and interaction internalization. While this framework helps clarify the structure of the emerging literature, it should not be viewed as exhaustive or definitive. As VFRL develops, future methods may combine multiple internal signals or introduce mechanisms that do not fit cleanly into the four dimensions proposed here.

A second limitation concerns the granularity of classification. Many methods contain hybrid components, yet we assign each paper to its dominant internalization mechanism. For instance, a process-level method may still rely on likelihood-based scoring, and an interaction-based method may also use uncertainty or reference signals. We therefore use the primary optimization, selection, or credit-assignment signal as the coding criterion. The resulting categories should be understood as analytical distinctions rather than strict ontological boundaries.

Finally, the cognitive-science analogy of internalization is necessarily partial. Human internalization involves developmental, social, embodied, and normative dimensions that current language-model training pipelines do not fully capture. In this paper, we use internalization as a functional analogy for the relocation of evaluative control from external feedback to internal regulatory signals. This analogy helps organize VFRL mechanisms, but it should not be taken to imply that LLMs possess human-like metacognition, self-understanding, or autonomous normative judgment.

References

Johannes Ackermann, Michael Noukhovitch, Takashi Ishida, and Masashi Sugiyama. 2026. Gradient regularization prevents reward hacking in reinforcement learning from human feedback and verifiable rewards. *arXiv preprint arXiv:2602.18037*.

Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. 2001. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624.

Locke Cai and Ivan Provilkov. 2025. Escaping the verifier: Learning to reason via demonstrations. *arXiv preprint arXiv:2511.21667*.

Xin-Qiang Cai and Masashi Sugiyama. 2026. Vicurl: Stabilizing verifier-independent rl reasoning via confidence-guided variance reduction. *arXiv preprint arXiv:2602.12579*.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Zhiyuan Fan, Dadi Guo, Zhitao He, Elias Stengel-Eskin, Mohit Bansal, and Yi R Fung. Proof-verifier: Enabling reinforcement learning from verifiable rewards for mathematical theorem proving.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Ali Hatamizadeh, Syeda Nahida Akter, Shrimai Prabhumoye, Jan Kautz, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. 2025. Rlp: Reinforcement as a pretraining objective. *arXiv preprint arXiv:2510.01265*.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. 2025. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *Preprint*, arXiv:2405.11143.

Chongxuan Huang, Lei Lin, Xiaodong Shi, Wenping Hu, and Ruiming Tang. 2026. Darl: Encouraging diverse answers for general reasoning without verifiers. *arXiv preprint arXiv:2601.14700*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *Preprint*, arXiv:2310.01798.

Suhyeok Jang, Dongyoung Kim, Changyeon Kim, Youngsuk Kim, and Jinwoo Shin. 2025. Verifier-free test-time sampling for vision language action models. *arXiv preprint arXiv:2510.05681*.

Xiaotong Ji, Rasul Tutunov, Matthieu Zimmer, and Haitham Bou Ammar. 2026. Scalable power sampling: Unlocking efficient, training-free reasoning for llms via distribution sharpening. *arXiv preprint arXiv:2601.21590*.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Rana Muhammad Shahroz Khan, Zijie Liu, Zhen Tan, Charles Fleming, and Tianlong Chen. 2026. Tms: Trajectory-mixed supervision for reward-free, on-policy sft. *arXiv preprint arXiv:2602.03073*.

Asher Koriat and Hila Sorka. 2015. The construction of categorization judgments: Using subjective confidence and response latency to test a distributed model. *Cognition*, 134:21–38.

Deanna Kuhn. 1993. Science as argument: Implications for teaching and learning scientific thinking. *Science education*, 77(3):319–337.

Abhijit Kumar, Natalya Kumar, and Shikhar Gupta. 2026. Execution-grounded credit assignment for grpo in code generation. *arXiv preprint arXiv:2603.16158*.

Ariel Kwiatkowski, Natasha Butt, Ismail Labiad, Julia Kempe, and Yann Ollivier. 2026. Likelihood-based reward designs for general llm reasoning. *arXiv preprint arXiv:2602.03979*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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*, volume 2024, pages 39578–39601.

Wei Liu, Siya Qi, Xinyu Wang, Chen Qian, Yali Du, and Yulan He. 2025. Nover: Incentive training for language models via verifier-free reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7450–7469.

Yan Liu, Feng Zhang, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, Han Liu, and Yangdong Deng. 2026. Efficient paths and dense rewards: Probabilistic flow reasoning for large language models. *arXiv preprint arXiv:2601.09260*.

Jennifer A Livingston. 2003. Metacognition: An overview.

Jinda Lu, Junkang Wu, Jinghan Li, Kexin Huang, Shuo Yang, Guoyin Wang, Jiancan Wu, Xiang Wang, and Xiangnan He. 2026. Bridging perception and reasoning: Token reweighting for rlvr in multimodal llms. *arXiv preprint arXiv:2603.25077*.

Monishwaran Maheswaran, Leon Lakhani, Zhongzhu Zhou, Shijia Yang, Junxiong Wang, Coleman Hooper, Yuezhou Hu, Rishabh Tiwari, Jue Wang, Harman Singh, Qingyang Wu, Yuqing Jian, Ce Zhang, Kurt Keutzer, Tri Dao, Xiaoxia Wu, Ben Athiwaratkun, James Zou, and Chenfeng Xu. 2026. Squeeze evolve: Unified multi-model orchestration for verifier-free evolution.

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

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

Earl K Miller. 2000. The prefontral cortex and cognitive control. *Nature reviews neuroscience*, 1(1):59–65.

Zhengyuan Pan, Yanhao Chen, Zhongquan Jian, Wanru Zhao, Haonan Ma, Meihong Wang, and Qingqiang Wu. 2026. Prototype entropy alignment: Reinforcing structured uncertainty in llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 24709–24717.

Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. 2025. Scaling test-time compute without verification or rl is suboptimal. *arXiv preprint arXiv:2502.12118*.

Kejian Shi, Yixin Liu, Peifeng Wang, Alexander R Fabbri, Shafiq Joty, and Arman Cohan. 2026. References improve llm alignment in non-verifiable domains. *arXiv preprint arXiv:2602.16802*.

Yunhao Tang, Sid Wang, Lovish Madaan, and Rémi Munos. 2026. Beyond verifiable rewards: Scaling reinforcement learning in language models to unverifiable data. *Advances in Neural Information Processing Systems*, 38:74421–74448.

Sascha Topolinski and Fritz Strack. 2009. The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence and judgments of grammaticality in artificial grammar learning. *Journal of Experimental Psychology: General*, 138(1):39.

V Venktesh, Mandeep Rathee, and Avishek Anand. 2025. Trust but verify! a survey on verification design for test-time scaling. *arXiv preprint arXiv:2508.16665*.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2025a. Dynscaling: Efficient verifier-free inference scaling via dynamic and integrated sampling. *arXiv preprint arXiv:2506.16043*.

Junlin Wang, Shang Zhu, Jon Saad-Falcon, Ben Athiwaratkun, Qingyang Wu, Jue Wang, Shuaiwen Leon Song, Ce Zhang, Bhuwan Dhingra, and James Zou. 2025b. Think deep, think fast: Investigating efficiency of verifier-free inference-time-scaling methods. *arXiv preprint arXiv:2504.14047*.

Yuanfu Wang, Zhixuan Liu, Xiangtian Li, Chaochao Lu, and Chao Yang. 2026. Native reasoning models: Training language models to reason on unverifiable data. *arXiv preprint arXiv:2602.11549*.

Xuexiang Wen, Hang Yu, Linchao Zhu, and Gaoang Wang. 2026a. Verifier-free rl for llms via intrinsic gradient-norm reward. *arXiv preprint arXiv:2605.09920*.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2026b. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *The Fourteenth International Conference on Learning Representations*.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 2025. Rlpr: Extrapolating rlvr to general domains without verifiers. *arXiv preprint arXiv:2506.18254v1*.

Chenyang Yuan and Xiaoyuan Cheng. 2026. Planning with language and generative models: Toward general reward-guided wireless network design. *arXiv preprint arXiv:2602.00357*.

Wenlin Zhong, Chengyuan Liu, Yiquan Wu, Bovin Tan, Changlong Sun, Yi Wang, Xiaozhong Liu, and Kun Kuang. 2026. P2s: Probabilistic process supervision for general-domain reasoning question answering. *arXiv preprint arXiv:2601.20649*.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*.

A Appendix

A.1 Detailed Literature Mapping

Table 4 provides the paper-level coding of representative VFRL and internalized-verification methods. For each paper, we report its dominant internal signal and the rationale used to assign it to one of the four internalization categories.

A.2 Method-Level Functional Mapping

Table 5 maps representative methods to their main functional roles. This matrix complements the taxonomy by showing that VFRL spans training-time RL, test-time scaling, reward construction, sample selection, credit assignment, and open-ended alignment.

A.3 Coding Protocol and Decision Rules

To make the taxonomy reproducible, we classify each paper by its *dominant internalized evaluative signal*. The coding unit is the primary mechanism used to construct rewards, select samples, assign credit, or compare candidate outputs. For hybrid methods, we assign the paper to the category corresponding to the signal that most directly drives optimization, selection, or credit assignment. Table 6 summarizes the resulting decision rules.

A.4 Future Directions and Targeted Limitations

Table 7 clarifies how the proposed future directions respond to the limitations of internalized verification. The mapping is not a strict one-to-one correspondence: each direction addresses a distinct reliability problem, while benchmark design provides a comprehensive framework for evaluating all four risks.

Table 4: **Paper-level coding of verifier-free reinforcement learning and related internalized-verification methods.** Papers are grouped by their dominant internalization mechanism. For each paper, we identify the primary internal signal and the coding rationale used for classification.

	Dominant internal signal	**Coding rationale**
Probabilistic Internalization		
Beyond Verifiable Rewards (Tang et al., 2026)	Jensen-style ELBO; latent CoT likelihood	Latent CoT likelihood drives optimization.
Likelihood-Based Reward Designs (Kwiatkowski et al., 2026)	Reference-answer probability and log-probability	Reference log-probability is the reward.
RLPR (Yu et al., 2025)	Policy probability of reference answer	Reward comes from intrinsic answer probability.
VeriFree (Zhou et al., 2025)	Probability of generating reference answer	Correctness is replaced by reference likelihood.
NOVER (Liu et al., 2025)	Proxy-model or policy-derived reward	Verifier-free incentive comes from model scoring.
Native Reasoning Models (Wang et al., 2026)	Answer likelihood under self-generated reasoning	Reasoning traces are rewarded by answer likelihood.
RLP (Hatamizadeh et al., 2025)	Next-token log-likelihood gain from thought	Thought value is measured by predictive gain.
Scalable Power Sampling (Ji et al., 2026)	Power distribution; distribution sharpening	Quality is induced by sharpening probability mass.
Planning with Language and Generative Models (Yuan and Cheng, 2026)	Diffusion-based generative sampling	Generative inference replaces verifier refinement.
Reward Hacking in Rubric-Based RL (Mahmoud et al., 2026)	Policy log-probability self-internalization gap	Diagnostic signal is policy log-probability.
TMS (Khan et al., 2026)	Near-policy targets from historical checkpoints	Historical policy distribution supplies supervision.
VIGOR (Wen et al., 2026a)	Teacher-forced NLL gradient norm	Small gradient norm acts as intrinsic preference.
Uncertainty Internalization		
DynScaling (Wang et al., 2025a)	Uncertainty from sampled responses	Uncertainty controls inference budget allocation.
Prototype Entropy Alignment (Pan et al., 2026)	Entropy signatures and prototypes	Reward aligns reasoning entropy patterns.
Scaling Test-Time Compute Without Verification (Setlur et al., 2025)	Trajectory heterogeneity; anti-concentration	Core variable is trajectory distribution uncertainty.
MG-Select (Jang et al., 2025)	KL from masked high-uncertainty distribution	KL measures confidence against uncertainty baseline.
VI-CuRL (Cai and Sugiyama, 2026)	Length-normalized negative entropy	Confidence determines curriculum inclusion.
Process Internalization		
CoT-Flow (Liu et al., 2026)	Stepwise probabilistic flow gain	Each step receives information-gain reward.
P2S (Zhong et al., 2026)	Path Faithfulness Reward	Step reward measures suffix reachability.
EGCA (Kumar et al., 2026)	Execution trace and semantic divergence	Credit is localized to divergent token spans.
Trust but Verify! (Venktesh et al., 2025)	Process-verifier design taxonomy	It systematizes process-level verification.
Interaction Internalization		
RARO (Cai and Provilkov, 2025)	Relativistic critic over expert-policy pairs	Reward is produced by relative discrimination.
References Improve LLM Alignment (Shi et al., 2026)	Reference-guided preference comparison	Learning signal is candidate-level preference.
DARL (Huang et al., 2026)	Controlled reference deviation and diversity	Optimization balances diversity and reference alignment.
SQUEEZE EVOLVE (Maheswaran et al., 2026)	Selection, mutation, recombination	Candidate quality emerges through evolution.
Think Deep, Think Fast (Wang et al., 2025b)	Majority voting and response features	Candidate agreement replaces verifier scoring.

	Training-time RL	Test-time scaling	Reward construction	Sample selection	Credit assignment	Open-ended alignment
Probabilistic Internalization						
Beyond Verifiable Rewards (Tang et al., 2026)	✓	–	✓	–	–	✓
Likelihood-Based Reward Designs (Kwiatkowski et al., 2026)	✓	–	✓	–	–	✓
RLPR (Yu et al., 2025)	✓	–	✓	–	–	✓
VeriFree (Zhou et al., 2025)	✓	–	✓	–	–	✓
NOVER (Liu et al., 2025)	✓	–	✓	–	–	✓
Native Reasoning Models (Wang et al., 2026)	✓	–	✓	–	–	✓
RLP (Hatamizadeh et al., 2025)	✓	–	✓	–	✓	–
Scalable Power Sampling (Ji et al., 2026)	–	✓	–	✓	–	–
Planning with Language and Generative Models (Yuan and Cheng, 2026)	–	✓	–	✓	–	✓
Reward Hacking in Rubric-Based RL (Mahmoud et al., 2026)	✓	–	✓	–	–	✓
TMS (Khan et al., 2026)	✓	–	–	✓	–	✓
VIGOR (Wen et al., 2026a)	✓	–	✓	–	–	–
Uncertainty Internalization						
DynScaling (Wang et al., 2025a)	–	✓	–	✓	–	–
Prototype Entropy Alignment (Pan et al., 2026)	✓	–	✓	–	✓	–
Scaling Test-Time Compute Without Verification (Setlur et al., 2025)	–	✓	–	–	–	–
MG-Select (Jang et al., 2025)	–	✓	–	✓	–	–
VI-CuRL (Cai and Sugiyama, 2026)	✓	–	–	✓	–	–
Process Internalization						
CoT-Flow (Liu et al., 2026)	✓	–	✓	–	✓	–
P2S (Zhong et al., 2026)	✓	–	✓	–	✓	–
EGCA (Kumar et al., 2026)	✓	–	–	–	✓	–
Trust but Verify! (Venktesh et al., 2025)	–	✓	–	–	✓	–
Interaction Internalization						
RARO (Cai and Provilkov, 2025)	✓	–	✓	–	–	✓
References Improve LLM Alignment (Shi et al., 2026)	✓	–	✓	✓	–	✓
DARL (Huang et al., 2026)	✓	–	✓	–	–	✓
SQUEEZE EVOLVE (Maheswaran et al., 2026)	–	✓	–	✓	–	✓
Think Deep, Think Fast (Wang et al., 2025b)	–	✓	–	✓	–	–

Table 5: **Method-level functional mapping of VFRL and internalized-verification methods.** A checkmark indicates that the paper substantially contributes to the corresponding functional role. The mapping shows that verifier-free reinforcement learning spans training-time optimization, test-time scaling, reward construction, sample selection, credit assignment, and open-ended alignment.

Category	Primary coding criterion	Typical internal signal	Decision rule for hybrid cases
Probabilistic Internalization	The method primarily replaces external verification with likelihood-, probability-, distribution-, or gradient-derived scoring.	Likelihood, logits, reference-answer log-probability, distribution sharpening, NLL gradient norm.	If likelihood is used as the main scalar reward or selection objective, the paper is coded as probabilistic, even if the method also uses references or generated reasoning traces.
Uncertainty Internalization	The method primarily uses the model's uncertainty structure to guide training, sampling, curriculum construction, or compute allocation.	Entropy, confidence, calibration, consistency, trajectory dispersion, uncertainty-aware KL divergence.	If uncertainty determines which samples, trajectories, or queries receive additional training or inference budget, the paper is coded as uncertainty-based, even when no explicit reward model is trained.
Process Internalization	The method primarily decomposes evaluation into intermediate reasoning steps, token spans, trajectories, or execution states.	Step-level reward, path faithfulness, process reward, execution trace, localized semantic divergence.	If likelihood or execution is used to assign local credit to reasoning steps rather than to score only the final answer, the paper is coded as process-based.
Interaction Internalization	The method primarily infers quality through relative comparison among candidates, agents, demonstrations, or evolving solution populations.	Ranking, voting, debate, adversarial comparison, pairwise preference, evolutionary selection.	If the learning signal is fundamentally relational, the paper is coded as interaction-based, even when references, uncertainty, or likelihood appear as auxiliary signals.

Table 6: **Coding protocol and decision rules for the four-dimensional taxonomy.** Each paper is assigned to the category corresponding to its dominant internalized evaluative signal. Hybrid methods are coded according to the signal that most directly drives reward construction, sample selection, credit assignment, or candidate comparison.

Future direction	Primary limitation addressed
Hybrid internal–external verification	Proxy misalignment
Calibrated and causal internal rewards	Miscalibration; local process error
Grounded multi-agent internalization	Preference drift
Benchmarks for internalized verification	Comprehensive evaluation of all four risks

Table 7: **Mapping future directions to limitations.** The directions are not one-to-one replacements for the four risks; rather, they target different levels of reliability, from anchoring internal proxies to evaluating whether internalized verification improves true task utility.