# Novelty Assessment Report

**Paper**: From Verifiable Dot to Reward Chain: Harnessing Verifiable Reference-based Rewards for Reinforcement Learning of Open-ended Generation

**PDF URL**: https://openreview.net/pdf?id=ZumVIktGbt

**Venue**: ICLR 2026 Conference Submission

**Year**: 2026

**Report Generated**: 2026-01-01

## Abstract

Reinforcement learning with verifiable rewards (RLVR) succeeds in reasoning tasks (e.g., math and code) by checking the final verifiable answer (i.e., a verifiable dot signal). However, extending this paradigm to open-ended generation is challenging because there is no unambiguous ground truth. Relying on single-dot supervision often leads to inefficiency and reward hacking. To address these issues, we propose reinforcement learning with verifiable reference-based rewards (RLVRR). Instead of checking the final answer, RLVRR extracts an ordered linguistic signal from high-quality references (i.e, reward chain). Specifically, RLVRR decomposes rewards into two dimensions: content, which preserves deterministic core concepts (e.g., keywords), and style, which evaluates adherence to stylistic properties through LLM-based verification. In this way, RLVRR combines the exploratory strength of RL with the efficiency and reliability of supervised fine-tuning (SFT). Extensive experiments on more than 10 benchmarks with Qwen and Llama models confirm the advantages of our approach. RLVRR (1) substantially outperforms SFT trained with ten times more data and advanced reward models, (2) unifies the training of structured reasoning and open-ended generation, and (3) generalizes more effectively while preserving output diversity. These results establish RLVRR as a principled and efficient path toward verifiable reinforcement learning for general-purpose LLM alignment.

**Disclaimer**

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

## Core Task Landscape

This paper addresses: **Reinforcement Learning for Open-Ended Text Generation with Verifiable Rewards**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Reward Signal Design and Verification**
- **Training Frameworks and Optimization Methods**
- **Domain-Specific Applications**
- **Quality Assurance and Evaluation**
- **Theoretical Foundations and Surveys**
- **Open-Ended Learning and Autonomy**
- **Specialized Techniques and Auxiliary Methods**

### Complete Taxonomy Tree

- Reinforcement Learning for Open-Ended Text Generation with Verifiable Rewards Survey Taxonomy
- Reward Signal Design and Verification
  - Verifiable Outcome-Based Rewards (6 papers)
  - [2] Lessons from Training Grounded LLMs with Verifiable Rewards (Pala Tej Deep, 2025) View paper
  - [6] Reasoning-SQL: Reinforcement Learning with SQL Tailored Partial Rewards for Reasoning-Enhanced Text-to-SQL (Pourreza, 2025) View paper
  - [20] NOVER: Incentive Training for Language Models via Verifier-Free Reinforcement Learning (Wei Liu, 2025) View paper
  - [29] Factually Consistent Summarization via Reinforcement Learning with Textual Entailment Feedback (Aharoni, 2023) View paper
  - [31] Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models (Song, 2025) View paper
  - [34] VerIF: Verification Engineering for Reinforcement Learning in Instruction Following (Hao Peng, 2025) View paper
  - Structured Evaluation-Based Rewards (4 papers)
  - [1] Reinforcement learning with rubric anchors (Huang Zenan, 2025) View paper
  - [9] Self-Rewarding Rubric-Based Reinforcement Learning for Open-Ended Reasoning (Ye, 2025) View paper
  - [19] Semantically-Aware Rewards for Open-Ended R1 Training in Free-Form Generation (Li, 2025) View paper
  - [39] Think-RM: Enabling Long-Horizon Reasoning in Generative Reward Models (Hong, 2025) View paper
  - Reference-Based and Decomposed Rewards ★ (3 papers)
  - [0] From Verifiable Dot to Reward Chain: Harnessing Verifiable Reference-based Rewards for Reinforcement Learning of Open-ended Generation (Anon et al., 2026) View paper
  - [23] Beyond sparse rewards: Enhancing reinforcement learning with language model critique in text generation (Cao Meng, 2024) View paper
  - [37] Optimizing Long-Form Clinical Text Generation with Claim-Based Rewards (Samyak Jhaveri, 2025) View paper
  - Multi-Dimensional and Adaptive Rewards (4 papers)
  - [5] Rlmr: Reinforcement learning with mixed rewards for creative writing (Liao Jian-xing, 2025) View paper
  - [36] Effective and Transparent RAG: Adaptive-Reward Reinforcement Learning for Decision Traceability (Ren Jing-yi, 2025) View paper

  - [42] Dynamic Reward Adjustment in Multi-Reward Reinforcement Learning for Counselor Reflection Generation (Min, 2024) View paper
  - [46] Multi-Dimensional Optimization for Text Summarization via Reinforcement Learning (Sangwon Ryu, 2024) View paper
- Training Frameworks and Optimization Methods
  - Policy Optimization Algorithms (3 papers)
  - [4] Ar-grpo: Training autoregressive image generation models via reinforcement learning (YUAN Shihao, 2025) View paper
  - [12] RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment (Dong, 2023) View paper
  - [15] Score as Action: Fine-Tuning Diffusion Generative Models by Continuous-time Reinforcement Learning (Zhao, 2025) View paper
  - Unified Training Paradigms (3 papers)
  - [24] URPO: A Unified Reward & Policy Optimization Framework for Large Language Models (Wang Hua, 2025) View paper
  - [30] Med-U1: Incentivizing Unified Medical Reasoning in LLMs via Large-scale Reinforcement Learning (Zhang Xiao-tian, 2025) View paper
  - [35] Omni-Thinker: Scaling Multi-Task RL in LLMs with Hybrid Reward and Task Scheduling (Zhou Jia-ming, 2025) View paper
  - Exploration and Efficiency Enhancement (1 papers)
  - [50] From General to Targeted Rewards: Surpassing GPT-4 in Open-Ended Long-Context Generation (Zhihan Guo, 2025) View paper
  - Verbal and Non-Scalar Feedback Learning (1 papers)
  - [33] Language Models Can Learn from Verbal Feedback Without Scalar Rewards (Luo, 2025) View paper
- Domain-Specific Applications
  - Structured Reasoning Tasks (1 papers)
  - [32] Learning to Reason for Long-Form Story Generation (Gurung, 2025) View paper
  - Creative and Open-Ended Generation (6 papers)
  - [10] Writing-Zero: Bridge the Gap Between Non-verifiable Problems and Verifiable Rewards (Lu, 2025) View paper
  - [22] Language Models that Think, Chat Better (Bhaskar, 2025) View paper
  - [26] Reverse-engineered reasoning for open-ended generation (Wang, 2025) View paper
  - [38] Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks (Xu Yifei, 2025) View paper
  - [49] Jointly reinforcing diversity and quality in language model generations (Li Tianjian, 2025) View paper
  - Domain-Specific Knowledge Tasks (3 papers)
  - [13] Crossing the Reward Bridge: Expanding RL with Verifiable Rewards Across Diverse Domains (Su Yi, 2025) View paper
  - [16] Reason Like a Radiologist: Chain-of-Thought and Reinforcement Learning for Verifiable Report Generation (Peiyuan Jing, 2025) View paper
  - [27] End-to-End Optimization for Multimodal Retrieval-Augmented Generation via Reward Backpropagation (Zhiyuan Fan, 2025) View paper
  - Multimodal and Cross-Domain Generation (2 papers)
  - [28] Generalizable Geometric Image Caption Synthesis (Xin Yue, 2025) View paper
  - [41] Tower+: Bridging Generality and Translation Specialization in Multilingual LLMs (Rei, 2025) View paper
  - Dialogue and Conversational Systems (1 papers)
  - [14] Dynamic planning in open-ended dialogue using reinforcement learning (Cohen, 2022) View paper
- Quality Assurance and Evaluation
  - Confidence Estimation and Calibration (1 papers)
  - [40] Reinforcement Learning for Better Verbalized Confidence in Long-Form Generation (Zhang, 2025) View paper
- Theoretical Foundations and Surveys (4 papers)
  - [11] Reinforcement learning in the era of large language models: Challenges and opportunities (Qianyue Hao, 2024) View paper
  - [17] Deep generative models for offline policy learning: Tutorial, survey, and perspectives on future directions (Chen Jiayu, 2024) View paper
  - [25] A technical survey of reinforcement learning techniques for large language models (Aggarwal, 2025) View paper
  - [43] From self-learning to self-evolving architectures in large language models: A short survey (Ranjan Sapkota, 2025) View paper
- Open-Ended Learning and Autonomy
  - Autotelic and Intrinsically Motivated Learning (2 papers)
  - [7] Endless minds most beautiful: building open-ended linguistic autotelic agents with deep reinforcement learning and language models (Teodorescu, 2023) View paper
  - [48] Autotelic Reinforcement Learning: Exploring Intrinsic Motivations for Skill Acquisition in Open-Ended Environments (Prakhar Srivastava, 2025) View paper
  - Open-Ended Environment Design (2 papers)
  - [18] Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions (Rui Wang, 2020) View paper
  - [47] Open: An open-ended physics environment for learning without a task (Gan, 2021) View paper
  - Creativity and Generative AI (1 papers)
  - [3] Deep reinforcement learning and creativity (Franceschelli, 2025) View paper
- Specialized Techniques and Auxiliary Methods (3 papers)
  - [8] ACTG-ARL: Differentially Private Conditional Text Generation with RL-Boosted Control (Hu, 2025) View paper
  - [44] Goal-Directed Search Outperforms Goal-Agnostic Memory Compression in Long-Context Memory Tasks (Yicong Zheng, 2025) View paper
  - [45] O-Searcher: A Searching-based Agent Model for Open-Domain Open-Ended Question Answering (J Mei, 2025) View paper

## Narrative

Core task: Reinforcement learning for open-ended text generation with verifiable rewards. This field addresses the challenge of training language models to produce creative, diverse, or task-specific outputs while ensuring that quality can be objectively measured. The taxonomy organizes research into several main branches: Reward Signal Design and Verification explores how to construct and validate reward functions, including reference-based metrics and decomposed signals that break complex objectives into verifiable components; Training Frameworks and Optimization Methods covers algorithmic innovations such as policy gradient techniques and online learning schemes; Domain-Specific Applications targets areas like creative writing, dialogue, and specialized domains (e.g., medical or code generation); Quality Assurance and Evaluation develops benchmarks and robustness checks; Theoretical Foundations and Surveys provide conceptual grounding; Open-Ended Learning and Autonomy investigates agents that set their own goals or explore without fixed endpoints; and Specialized Techniques and Auxiliary Methods encompass supporting tools like data augmentation or auxiliary losses. Representative works such as Grounded LLMs Verifiable Rewards[2] and Rubric Anchors[1] illustrate efforts to ground reward signals in interpretable criteria, while Deep RL Creativity[3] and Mixed Rewards Creative Writing[5] highlight domain-specific challenges in balancing novelty with coherence.

A particularly active line of work focuses on decomposing holistic quality judgments into verifiable sub-rewards, enabling more transparent and stable training. For instance, some studies use rubric-based or claim-level decompositions (Rubric Anchors[1], Claim-Based Clinical Rewards[37]) to provide fine-grained feedback, while others explore hybrid signals that combine rule-based checks with learned evaluators. The original paper, Verifiable Dot Reward Chain[0], sits within the Reference-Based and Decomposed Rewards cluster, emphasizing structured reward decomposition to improve verifiability and interpretability. Compared to neighbors like Beyond Sparse Rewards[23], which addresses the broader challenge of reward sparsity across tasks, and Claim-Based Clinical Rewards[37], which targets domain-specific clinical text, Verifiable Dot Reward Chain[0] appears to focus on chaining intermediate verification steps to ensure that each component of the generation process receives clear, actionable feedback. This approach contrasts with end-to-end learned reward models and reflects ongoing debates about the trade-offs between automation, interpretability, and generalization in open-ended generation settings.

## Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. Beyond sparse rewards: Enhancing reinforcement learning with language model critique in text generation

**Authors**: Cao Meng, Shu, Lei, Yu Lei, Zhu Yun, et al. (9 authors total) | **Year/Venue**: 2024 | **URL**: View paper

### Abstract

Reinforcement learning (RL) can align language models with non-differentiable reward signals, such as human preferences. However, a major challenge arises from the sparsity of these reward signals - typically, there is only a single reward for an entire output. This sparsity of rewards can lead to inefficient and unstable learning. To address this challenge, our paper introduces an novel framework that utilizes the critique capability of Large Language Models (LLMs) to produce intermediate-step ...

### Relationship Analysis

Both papers belong to the Reference-Based and Decomposed Rewards category, extracting reward signals from reference data through decomposition strategies. They overlap in addressing sparse reward problems in open-ended text generation by creating intermediate-step rewards: the original paper decomposes rewards into content (deterministic keywords) and style dimensions from high-quality references, while the candidate paper uses LLM critique to generate token/span-level intrinsic rewards from holistic environment feedback. The key difference is that the original paper extracts verifiable linguistic signals directly from reference data in a structured reward chain, whereas the candidate paper relies on a separate critic LLM to retrospectively evaluate and decompose the policy model's outputs into dense rewards.

### 2. Optimizing Long-Form Clinical Text Generation with Claim-Based Rewards

**Authors**: Samyak Jhaveri, Kimï¼□ Jang-Won, Praphul Singh, Taghavi, Tara, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

### Abstract

Automating clinical documentation with large language models requires precise alignment with priorities such as completeness and factual grounding. We present an evaluation-integrated reinforcement learning framework for long-form clinical text generation that couples Group Relative Policy Optimization (GRPO) with DocLens, a claim-level evaluator that provides deterministic, dialogue-grounded rewards. Our method directly optimizes factual grounding and completeness without training a separate re...

### Relationship Analysis

Both papers belong to the Reference-Based and Decomposed Rewards category, extracting verifiable reward signals from reference data through decomposition. They overlap in using reference-based decomposition for RL optimization: the original paper decomposes rewards into content (deterministic keywords) and style dimensions from high-quality references, while the candidate paper decomposes clinical note quality into claim-level completeness (recall) and factual grounding (precision) from dialogue references. The key difference is that the original paper targets general open-ended generation with a reward chain approach across diverse tasks, whereas the candidate paper focuses specifically on long-form clinical text generation using claim-based rewards derived from doctor-patient dialogues.

## Contributions Analysis

**Overall novelty summary.** The paper proposes RLVRR, a framework that extends reinforcement learning with verifiable rewards from reasoning tasks to open-ended generation by decomposing rewards into content and style dimensions extracted from reference data. It resides in the Reference-Based and Decomposed Rewards leaf, which contains only three papers total, indicating a relatively sparse research direction within the broader taxonomy of fifty papers. This leaf focuses specifically on extracting reward signals through decomposition rather than single-dot verification or LLM-judged rubrics, positioning the work at the intersection of verifiable supervision and open-ended generation challenges.

The taxonomy reveals that neighboring leaves address related but distinct approaches: Verifiable Outcome-Based Rewards (six papers) handles deterministic final-answer checking for reasoning tasks, while Structured Evaluation-Based Rewards (four papers) employs LLM-as-judge methods for open-ended evaluation. The paper's approach bridges these directions by maintaining verifiability through reference-based decomposition while targeting open-ended tasks. Nearby branches like Multi-Dimensional and Adaptive Rewards (four papers) explore dynamic reward balancing, and Creative and Open-Ended Generation (six papers) addresses domain-specific challenges, suggesting the work connects reward design innovations with application-driven concerns in creative text generation.

Among twenty-nine candidates examined, the contribution-level analysis shows mixed novelty signals. The RLVRR framework and reward chain decomposition each examined ten candidates with one refutable match, suggesting some prior work exists in reference-based reward extraction or decomposition strategies within the limited search scope. The unified training approach examined nine candidates with no clear refutations, indicating potentially stronger novelty for this aspect. The statistics reflect a focused semantic search rather than exhaustive coverage, meaning the presence of one refutable candidate per contribution does not definitively establish extensive prior overlap but signals areas requiring careful positioning against existing decomposition methods.

Based on the limited search scope of twenty-nine candidates, the work appears to occupy a moderately explored niche within reward design for open-ended generation. The sparse Reference-Based and Decomposed Rewards leaf and the presence of some overlapping prior work suggest the contribution lies in synthesizing and extending existing ideas—combining reference-based signals with content-style decomposition—rather than introducing entirely unprecedented concepts. The analysis captures top semantic matches and immediate neighbors but does not exhaustively survey all possible related work in reward shaping or open-ended RL.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

## Contribution 1: RLVRR framework for open-ended generation

**Description**: The authors propose a new reinforcement learning framework called RLVRR that extends verifiable reward-based RL from reasoning tasks to open-ended generation by using verifiable reference-based rewards instead of single-dot supervision.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Reinforcement learning with rubric anchors

**URL**: View paper

**Prior Art Analysis**

Rubric Anchors[1] demonstrates that extending verifiable reward-based RL to open-ended generation was already explored prior to the original paper's submission. Both papers address the same core challenge: moving beyond single-dot verification in reasoning tasks to handle open-ended generation where no unambiguous ground truth exists. Rubric Anchors[1] explicitly states it extends RLVR 'beyond strictly verifiable domains by integrating open-ended tasks into the framework through rubric-based reward,' which directly parallels the original paper's claim of extending 'verifiable reward-based RL from reasoning tasks to open-ended generation.' The candidate paper uses structured rubrics as verifiable signals, while the original uses reference-based rewards, but both fundamentally solve the identical problem of creating verifiable rewards for open-ended tasks where traditional dot-supervision fails.

**Evidence**

Evidence 1 - **Rationale**: Both papers acknowledge RLVR as an established paradigm for reasoning tasks with verifiable dot signals, establishing the common starting point. - **Original**: reinforcement learning with verifiable rewards (rlvr) (shao et al., 2024; yu et al., 2025a; team, 2025) has emerged as a promising paradigm for enhancing large language models (llms) in reasoning tasks such as mathematics and code generation. at its core, rlvr sidesteps the complicated chain-of-thou... - **Candidate**: reinforcement learning from verifiable rewards (rlvr) has emerged as a powerful paradigm for enhancing large language models (llms), exemplified by the success of openai's o-series. in rlvr, rewards are derived from deterministic, programmatically verifiable signals-such as passing unit tests in cod...

Evidence 2 - **Rationale**: Both papers identify the identical limitation of RLVR (failure on open-ended tasks without ground truth) and propose extending it to open-ended generation, demonstrating prior work on this exact contribution. - **Original**: while rlvr is simple yet effective for reasoning tasks (e.g., math and code generation), it fails in open-ended generation tasks, where no unambiguous ground truth exists and reliable verification cannot be reduced to a single dot. - **Candidate**: while effective, this requirement for unambiguous correctness largely confines rlvr to domains with clear, automatically checkable outcomes. to overcome this limitation, we extend the rlvr paradigm beyond strictly verifiable domains by integrating open-ended tasks into the framework through rubric-b...

Evidence 3 - **Rationale**: Both papers frame the core research question identically: extending RL to open-ended generation beyond single-dot supervision, with Rubric Anchors[1] providing a solution through rubric-based rewards. - **Original**: this motivates a critical research question:how can we extend rl optimization to open-ended generation by moving beyond single-dot supervision? - **Candidate**: this shift, however, introduces a fundamental challenge: how to construct reward signals that are both reliable and scalable in the absence of explicit ground truth. rubric-based reward offers a promising path forward: by defining structured, interpretable criteria for assessment, it can capture mul...

Evidence 4 - **Rationale**: Both papers explicitly claim to extend RLVR to open-ended generation, with Rubric Anchors[1] demonstrating this extension was already accomplished prior to the original paper's work. - **Original**: to this end, we introducerlvrr(reinforcement learning with verifiable reference-based rewards), a framework that extends rlvr to open-ended generation. instead of relying on a single verifiabledot, rlvrr extracts an ordered sequence of verifiable linguistic signals from high-quality references - **Candidate**: we address this limitation by extending rlvr to incorporate open-ended tasks and other forms of non-verifiable data, thereby broadening its applicability to a much wider range of real-world scenarios.

### 2. Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks

**URL**: View paper

**Brief Assessment**

Direct Reasoning Optimization[38] focuses on long-form reasoning tasks using reasoning reflection rewards (R3), while RLVRR addresses open-ended generation through reference-based content and style rewards extracted from high-quality references.

### 3. Reinforcement learning with token-level feedback for controllable text generation

**URL**: View paper

**Brief Assessment**

Token-Level Feedback[69] focuses on controllable text generation with token-level rewards for attribute control (sentiment, toxicity), not on extending verifiable reward-based RL from reasoning tasks to open-ended generation using reference-based rewards.

### 4. Text2reward: Reward shaping with language models for reinforcement learning

**URL**: View paper

**Brief Assessment**

Text2Reward[70] focuses on reward shaping for RL using LLMs to generate executable reward code for robotic manipulation and locomotion tasks, not on extending verifiable reward-based RL from reasoning to open-ended generation with reference-based rewards.

### 5. NOVER: Incentive Training for Language Models via Verifier-Free Reinforcement Learning

**URL**: View paper

**Brief Assessment**

NOVER[20] focuses on eliminating the need for external verifiers by using perplexity-based rewards, while the original paper extends verifiable reward-based RL to open-ended generation using reference-based rewards with content and style dimensions. These are distinct technical approaches to different aspects of the problem.

### 6. Semantically-Aware Rewards for Open-Ended R1 Training in Free-Form Generation

**URL**: View paper

**Brief Assessment**

Semantically-Aware Rewards[19] focuses on designing reward models (PrefBERT) for evaluating open-ended generation quality in GRPO, not on extending verifiable reward-based RL frameworks from reasoning to open-ended tasks using reference-based rewards as RLVRR does.

### 7. Learning to Reason for Long-Form Story Generation

**URL**: View paper

**Brief Assessment**

Long-Form Story Reasoning[32] focuses on long-form story generation with chapter prediction tasks, not general open-ended generation. The candidate uses completion likelihood improvement as rewards for story-specific reasoning, while the original proposes reference-based content and style rewards for diverse open-ended tasks.

### 8. Dynamic Multi-Reward Weighting for Multi-Style Controllable Generation

**URL**: View paper

**Brief Assessment**

Dynamic Multi-Reward Weighting[72] focuses on multi-style controllable text generation using multi-objective RL with discriminator-based rewards, not on extending verifiable reward-based RL from reasoning tasks to open-ended generation using reference-based rewards.

### 9. RLAC: Reinforcement Learning with Adversarial Critic for Free-Form Generation Tasks

**URL**: View paper

**Brief Assessment**

RLAC[73] focuses on dynamic rubric verification using an adversarial critic-generator framework, while the original paper proposes reference-based verifiable rewards with content and style decomposition. These represent distinct technical approaches to open-ended generation.

### 10. Teacher Forcing Recovers Reward Functions for Text Generation

**URL**: View paper

**Brief Assessment**

Teacher Forcing Rewards[71] focuses on deriving reward functions from teacher-forced models for semi-supervised learning, not on extending verifiable reward-based RL from reasoning tasks to open-ended generation using reference-based rewards as proposed in RLVRR.

## Contribution 2: Reward chain decomposition into content and style

**Description**: The method decomposes rewards into content dimension that captures deterministic core concepts like keywords, and style dimension that evaluates stylistic properties using LLM-based verification, creating an ordered sequence of verifiable linguistic signals.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Reinforcement Learning-Guided Large Language Model Fine-Tuning for Privacy-Preserving Text Rewriting

**URL**: View paper

**Brief Assessment**

Privacy-Preserving Text Rewriting[67] focuses on privacy-preserving text rewriting with disentangled semantic and stylistic representations for anonymization, not on open-ended generation tasks with verifiable reference-based rewards.

### 2. Bone Soups: A Seek-and-Soup Model Merging Approach for Controllable Multi-Objective Generation

**URL**: View paper

**Brief Assessment**

Bone Soups[61] addresses controllable multi-objective generation through model merging with combined reward functions, not reward decomposition into content/style dimensions for language model training.

### 3. Aligning Dialogue Agents with Global Feedback via Large Language Model Multimodal Reward Decomposition

**URL**: View paper

**Brief Assessment**

Multimodal Reward Decomposition[68] appears to focus on dialogue agents with multimodal feedback, while the original paper addresses open-ended text generation with reference-based rewards. The candidate's full text context is empty, preventing detailed comparison.

### 4. Efficient controlled language generation with low-rank autoregressive reward models

**URL**: View paper

**Brief Assessment**

Low-Rank Autoregressive Rewards[63] decomposes reward models into baseline and marginal components for efficient decoding, not into content and style dimensions for capturing linguistic signals from references.

### 5. Reinforced rewards framework for text style transfer

**URL**: View paper

**Brief Assessment**

Reinforced Rewards Framework[65] focuses on text style transfer tasks (formal-to-informal, excitement levels, Shakespearean English), not general open-ended generation with verifiable reference-based rewards. The candidate does not demonstrate prior work on reward chain decomposition for open-ended LLM alignment.

### 6. On learning text style transfer with direct rewards

**URL**: View paper

**Prior Art Analysis**

Direct Rewards Style Transfer[64] demonstrates prior work that decomposes rewards into content and style dimensions for text style transfer. The candidate paper explicitly describes decomposing rewards into content preservation (measuring semantic similarity between source and output) and style dimensions (evaluating stylistic properties). This directly challenges the novelty claim of the original paper's reward decomposition approach, as the candidate paper published this method in 2021, before the original submission.

**Evidence**

Evidence 1 - **Rationale**: Both papers decompose rewards into content and style dimensions. The candidate explicitly describes using separate reward functions for content preservation and style transfer accuracy, establishing this decomposition approach prior to the original paper. - **Original**: rlvrr decomposes rewards into two dimensions:content, which preserves deterministic core concepts (e.g., keywords), andstyle, which evaluates adherence to stylistic properties through llm-based verification. - **Candidate**: we use four reward functions to control the quality of the system outputs. the quality of the outputs is assessed in three ways: style transfer accuracy, content preservation, and fluency. we attend to each of these factors with their respective rewards.

Evidence 2 - **Rationale**: Both papers use content rewards to verify preservation of core concepts. The candidate uses semantic similarity metrics to assess content preservation, while the original uses keywords - both are reference-based approaches to measuring content fidelity. - **Original**: thecontentreward uses reference-derived key points (e.g., key entities or keywords) to score a rollout by whether those deterministic core concepts are present - **Candidate**: to ensure that the system outputs still preserve the basic semantics of the source sentences, we use the pretrained sim model introduced in wieting et al. (2019b,a) to measure the semantic similarity between the source sentences and system outputs.

Evidence 3 - **Rationale**: Both papers evaluate stylistic properties through verification mechanisms. The candidate uses a style classifier to assess style transfer accuracy, while the original uses LLM-generated verifiable checks - both approaches verify adherence to stylistic properties. - **Original**: the stylereward runs a small set of llm-generated, verifiable python checks on the rollout to confirm adherence to reference-specific stylistic properties (e.g., length, format). - **Candidate**: we use a style classifier to provide the supervision signal to the generator with respect to the style transfer accuracy. the min-max game between the generator gand the classifier fcls

### 7. Alarm: Align language models via hierarchical rewards modeling

**URL**: View paper

**Brief Assessment**

ALARM[62] decomposes rewards into holistic and aspect-specific dimensions (e.g., factuality, grammar) for different tasks, not specifically into content and style dimensions as defined in the original paper. The original paper's content reward uses reference-derived keywords and style reward uses verifiable Python checks, which differs from ALARM's task-specific aspect rewards.

### 8. Learning goal-conditioned representations for language reward models

**URL**: View paper

**Brief Assessment**

Goal-Conditioned Representations[59] focuses on learning goal-conditioned representations via contrastive learning for reward models in RL settings, not on decomposing rewards into content and style dimensions for language generation tasks.

### 9. Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier

**URL**: View paper

**Brief Assessment**

Unified Preference Optimization[66] focuses on decomposing preference and auxiliary objectives for LLM alignment, not on decomposing rewards into content and style dimensions with verifiable linguistic signals.

### 10. USO: Unified Style and Subject-Driven Generation via Disentangled and Reward Learning

**URL**: View paper

**Brief Assessment**

USO[60] focuses on image generation with style-content disentanglement for visual customization tasks, not language model training with verifiable linguistic rewards as in the original paper.

## Contribution 3: Unified training approach for reasoning and generation

**Description**: The framework provides a unified approach that can handle both structured reasoning tasks and open-ended generation tasks within a single training paradigm, combining the exploratory strength of RL with the efficiency of supervised fine-tuning.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. A review on synergizing knowledge graphs and large language models: Z. Yang et al.

**URL**: View paper

**Brief Assessment**

Knowledge Graphs LLMs Review[58] is a survey paper on synergizing knowledge graphs and LLMs. The provided context fragments mention text generation, logical reasoning, and joint training of entity/word representations, but do not describe a unified RL-based training framework for structured reasoning and open-ended generation tasks as proposed in the original paper.

### 2. Complex Reasoning over Logical Queries on Commonsense Knowledge Graphs

**URL**: View paper

**Brief Assessment**

Logical Queries Commonsense[56] focuses on complex commonsense reasoning over knowledge graphs using multi-hop logical queries, not on unifying structured reasoning and open-ended generation tasks within a single RL training paradigm.

### 3. Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks

**URL**: View paper

**Brief Assessment**

Low-Resource Transfer Methods[57] focuses on transfer learning for low-resource text generation tasks using instruction tuning and parameter-efficient fine-tuning (LoRA, Adapter). It does not address unified training for both structured reasoning tasks (e.g., math, code) and open-ended generation within a single RL-based framework, which is the core novelty of the original paper's RLVRR approach.

### 4. Enhancing multimodal analogical reasoning with Logic Augmented Generation

**URL**: View paper

**Brief Assessment**

Logic Augmented Generation[51] focuses on enhancing multimodal analogical reasoning for metaphor detection and understanding through semantic knowledge graphs and conceptual blending theory. It does not address unified training paradigms combining RL with supervised fine-tuning for both structured reasoning and open-ended generation tasks.

### 5. O-Searcher: A Searching-based Agent Model for Open-Domain Open-Ended Question Answering

**URL**: View paper

**Brief Assessment**

O-Searcher[45] focuses on unified training for open-ended and closed-ended question answering tasks using search-based retrieval, not on combining structured reasoning with open-ended generation within a single RL paradigm as described in the original paper.

### 6. Structured path guidance for logical coherence in large language model generation

**URL**: View paper

**Brief Assessment**

Structured Path Guidance[52] focuses on controlling generation paths through structural encoding and dynamic state mechanisms for chain-of-thought reasoning, not on unifying RL-based training for both reasoning and open-ended generation tasks as proposed in the original paper.

### 7. MURMUR: Modular multi-step reasoning for semi-structured data-to-text generation

**URL**: View paper

**Brief Assessment**

MURMUR[53] focuses on modular multi-step reasoning for data-to-text generation from semi-structured data (graphs/tables), not on unified training paradigms combining RL and supervised fine-tuning for general-purpose LLMs.

### 8. SORTIE: Dependency-Aware Symbolic Reasoning for Logical Data-to-text Generation

**URL**: View paper

**Brief Assessment**

SORTIE[54] focuses on logical data-to-text generation using symbolic reasoning with a table-compatible programming language. It does not address unified training for both structured reasoning and open-ended generation tasks within a single RL paradigm as described in the original paper.

### 9. Omni-Thinker: Scaling Multi-Task RL in LLMs with Hybrid Reward and Task Scheduling

**URL**: View paper

**Brief Assessment**

Omni-Thinker[35] focuses on multi-task RL with task scheduling and hybrid rewards across domains, while the original paper specifically addresses extending verifiable rewards from reasoning to open-ended generation through reference-based reward chains. The technical approaches differ fundamentally.

## Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

### 1. Reinforcement learning with rubric anchors

**Detected in**: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

## References

- [0] From Verifiable Dot to Reward Chain: Harnessing Verifiable Reference-based Rewards for Reinforcement Learning of Open-ended Generation View paper
- [1] Reinforcement learning with rubric anchors View paper
- [2] Lessons from Training Grounded LLMs with Verifiable Rewards View paper
- [3] Deep reinforcement learning and creativity View paper
- [4] Ar-grpo: Training autoregressive image generation models via reinforcement learning View paper
- [5] Rlmr: Reinforcement learning with mixed rewards for creative writing View paper
- [6] Reasoning-SQL: Reinforcement Learning with SQL Tailored Partial Rewards for Reasoning-Enhanced Text-to-SQL View paper
- [7] Endless minds most beautiful: building open-ended linguistic autotelic agents with deep reinforcement learning and language models View paper
- [8] ACTG-ARL: Differentially Private Conditional Text Generation with RL-Boosted Control View paper
- [9] Self-Rewarding Rubric-Based Reinforcement Learning for Open-Ended Reasoning View paper
- [10] Writing-Zero: Bridge the Gap Between Non-verifiable Problems and Verifiable Rewards View paper
- [11] Reinforcement learning in the era of large language models: Challenges and opportunities View paper
- [12] RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment View paper
- [13] Crossing the Reward Bridge: Expanding RL with Verifiable Rewards Across Diverse Domains View paper
- [14] Dynamic planning in open-ended dialogue using reinforcement learning View paper
- [15] Score as Action: Fine-Tuning Diffusion Generative Models by Continuous-time Reinforcement Learning View paper
- [16] Reason Like a Radiologist: Chain-of-Thought and Reinforcement Learning for Verifiable Report Generation View paper
- [17] Deep generative models for offline policy learning: Tutorial, survey, and perspectives on future directions View paper
- [18] Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions View paper
- [19] Semantically-Aware Rewards for Open-Ended R1 Training in Free-Form Generation View paper
- [20] NOVER: Incentive Training for Language Models via Verifier-Free Reinforcement Learning View paper
- [21] Writing-Zero: Bridge the Gap Between Non-verifiable Tasks and Verifiable Rewards View paper
- [22] Language Models that Think, Chat Better View paper
- [23] Beyond sparse rewards: Enhancing reinforcement learning with language model critique in text generation View paper
- [24] URPO: A Unified Reward & Policy Optimization Framework for Large Language Models View paper
- [25] A technical survey of reinforcement learning techniques for large language models View paper
- [26] Reverse-engineered reasoning for open-ended generation View paper
- [27] End-to-End Optimization for Multimodal Retrieval-Augmented Generation via Reward Backpropagation View paper
- [28] Generalizable Geometric Image Caption Synthesis View paper
- [29] Factually Consistent Summarization via Reinforcement Learning with Textual Entailment Feedback View paper
- [30] Med-U1: Incentivizing Unified Medical Reasoning in LLMs via Large-scale Reinforcement Learning View paper
- [31] Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models View paper

- [32] Learning to Reason for Long-Form Story Generation View paper
- [33] Language Models Can Learn from Verbal Feedback Without Scalar Rewards View paper
- [34] VerIF: Verification Engineering for Reinforcement Learning in Instruction Following View paper
- [35] Omni-Thinker: Scaling Multi-Task RL in LLMs with Hybrid Reward and Task Scheduling View paper
- [36] Effective and Transparent RAG: Adaptive-Reward Reinforcement Learning for Decision Traceability View paper
- [37] Optimizing Long-Form Clinical Text Generation with Claim-Based Rewards View paper
- [38] Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks View paper
- [39] Think-RM: Enabling Long-Horizon Reasoning in Generative Reward Models View paper
- [40] Reinforcement Learning for Better Verbalized Confidence in Long-Form Generation View paper
- [41] Tower+: Bridging Generality and Translation Specialization in Multilingual LLMs View paper
- [42] Dynamic Reward Adjustment in Multi-Reward Reinforcement Learning for Counselor Reflection Generation View paper
- [43] From self-learning to self-evolving architectures in large language models: A short survey View paper
- [44] Goal-Directed Search Outperforms Goal-Agnostic Memory Compression in Long-Context Memory Tasks View paper
- [45] O-Searcher: A Searching-based Agent Model for Open-Domain Open-Ended Question Answering View paper
- [46] Multi-Dimensional Optimization for Text Summarization via Reinforcement Learning View paper
- [47] Open: An open-ended physics environment for learning without a task View paper
- [48] Autotelic Reinforcement Learning: Exploring Intrinsic Motivations for Skill Acquisition in Open-Ended Environments View paper
- [49] Jointly reinforcing diversity and quality in language model generations View paper
- [50] From General to Targeted Rewards: Surpassing GPT-4 in Open-Ended Long-Context Generation View paper
- [51] Enhancing multimodal analogical reasoning with Logic Augmented Generation View paper
- [52] Structured path guidance for logical coherence in large language model generation View paper
- [53] MURMUR: Modular multi-step reasoning for semi-structured data-to-text generation View paper
- [54] SORTIE: Dependency-Aware Symbolic Reasoning for Logical Data-to-text Generation View paper
- [55] Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective View paper
- [56] Complex Reasoning over Logical Queries on Commonsense Knowledge Graphs View paper
- [57] Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks View paper
- [58] A review on synergizing knowledge graphs and large language models: Z. Yang et al. View paper
- [59] Learning goal-conditioned representations for language reward models View paper
- [60] USO: Unified Style and Subject-Driven Generation via Disentangled and Reward Learning View paper
- [61] Bone Soups: A Seek-and-Soup Model Merging Approach for Controllable Multi-Objective Generation View paper
- [62] Alarm: Align language models via hierarchical rewards modeling View paper
- [63] Efficient controlled language generation with low-rank autoregressive reward models View paper
- [64] On learning text style transfer with direct rewards View paper
- [65] Reinforced rewards framework for text style transfer View paper
- [66] Unified Preference Optimization: Language Model Alignment Beyond the Preference Frontier View paper
- [67] Reinforcement Learning-Guided Large Language Model Fine-Tuning for Privacy-Preserving Text Rewriting View paper
- [68] Aligning Dialogue Agents with Global Feedback via Large Language Model Multimodal Reward Decomposition View paper
- [69] Reinforcement learning with token-level feedback for controllable text generation View paper
- [70] Text2reward: Reward shaping with language models for reinforcement learning View paper
- [71] Teacher Forcing Recovers Reward Functions for Text Generation View paper
- [72] Dynamic Multi-Reward Weighting for Multi-Style Controllable Generation View paper
- [73] RLAC: Reinforcement Learning with Adversarial Critic for Free-Form Generation Tasks View paper