HARPO: Hallucination-Aware Reinforcement Learning for Faithful and Creative Language Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are prone to generating hallucinated content, which compromises their reliability in knowledge-intensive tasks. To address this challenge without sacrificing creativity, we propose **HARPO**, a reinforcement learning framework designed to jointly optimize faithfulness and creativity. HARPO incorporates a Generative Reward Model (GRM), trained via verifiable feedback, to simultaneously assess faithful adherence and writing quality. Crucially, we employ a Selective Activation Mechanism (SAM) that acts as a conditional gate, incentivizing creativity only when outputs are hallucination-free. To further stabilize training, we implement a curriculum learning scheme that progressively shifts from creative writing data to hallucination-centric samples. Extensive experiments demonstrate that HARPO significantly improves faithfulness while preserving expressiveness, outperforming strong baselines.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in generating fluent, human-like text. However, their reliability remains a critical concern due to hallucinated outputs that are unfaithful with respect to a given source (Ji et al., 2022). Such inaccuracies undermine the trustworthiness and utility of LLMs, particularly in high-stakes applications.

While reinforcement learning (RL) techniques have been widely adopted to boost the performance of large language models (LLMs) (Wei et al., 2025), optimizing for a single capability often comes at the cost of degraded performance in other dimensions. For instance, the RLMR framework (Liao et al., 2025) underscores the inherent challenge of preserving subjective textual quality while adhering to rigid objective constraints. Similarly, RLCR (Damani et al., 2025) demonstrates that RL training tailored to specific question-answering (QA) tasks can drastically impair a model's calibration on out-of-domain benchmarks.

In this paper, we address the conflict between **hallucination mitigation** and **creative generation**. Mitigating hallucinations demands precision and caution, whereas creative writing requires richness and novelty. Existing methods struggle to balance these competing objectives: single-reward strategies fail to improve both simultaneously, while naive data mixing often leads to interference between faithful adherence and expressive quality.

To address this problem, we introduce **HARPO** (Hallucination-Aware Reinforcement for Policy Optimization), a framework designed to jointly optimize faithfulness and creativity. Central to our approach is the Hallucination-Aware Generative Reward Model (HA-GRM), trained using Reinforcement Learning with Verifiable Rewards (RLVR) (Shao et al., 2024). Beyond standard binary classification, we introduce an auxiliary reward based on hallucinated-span prediction. This provides the model with dense, fine-grained supervision, enabling it to pinpoint unfaithful segments. Furthermore, by incorporating pairwise preference data for creative writing, the HA-GRM learns to evaluate stylistic quality alongside faithfulness.

To effectively utilize these signals during policy optimization, we propose a Selective Activation Mechanism (SAM). Unlike traditional fixed-weight scalarization, prone to reward conflicts, SAM acts as a conditional gate: the creative writing reward is activated only if the response is judged to be hallucination-free. Finally, to stabilize the training process, we design a data curriculum strategy that gradually shifts the data distribution from general writing instructions to hallucination-focused scenarios. This approach mitigates the catastrophic forgetting often observed in single-objective alignment, preserving broad linguistic proficiency while sharpening evidential rigor.

Our results demonstrate that HA-GRM achieves superior generalization on unseen hallucination tasks, thanks to reinforcement learning with fine-grained rewards. Furthermore, we validate the effectiveness of HARPO through extensive experiments on the Qwen2.5 and Qwen3 model families (parameter scales from 1.7B to 8B). When applied to policy optimization, HARPO reduces the hallucination rate by over 85% on Qwen3-4B while simultaneously boosting creative writing scores. Comparative analyses confirm that SAM significantly outperforms linear reward mixing, whereas our curriculum learning strategy effectively balances the dual optimization of faithfulness and creativity.

The contributions are as follows:

- We identify the optimization conflict between hallucination mitigation and creative generation, and introduce HARPO, a framework that jointly optimizes both capabilities.
- We develop the HA-GRM and SAM to hierarchically resolve reward conflicts, supported by a curriculum strategy that prevents catastrophic forgetting.
- Extensive experiments demonstrate that HARPO consistently enhances performance across various model families and scales, confirming its effectiveness.

2 Methodology

We introduce HARPO, a framework designed to mitigate hallucinations in large language models (LLMs) while preserving creative generation capabilities. Our approach integrates a Hallucination-Aware Generative Reward Model (HA-GRM) with a curriculum-driven reinforcement learning strategy. The optimization process features two core mechanisms: a Selective Activation Mechanism (SAM) that gates creativity rewards based on faithful accuracy, and a dynamic curriculum that progressively shifts the training distribution from creative to hallucination-sensitive tasks.

2.1 HA-GRM: Reward Model Training

Task Formulation. We train the HA-GRM to detect hallucinations at both the response and span levels, following the paradigm of generative reward models (Wu et al., 2023). Given the input context c and the generated response $y = (y_1, y_2 ... y_T)$ consisting of T characters, the model need to first identify whether the response y is hallucinated and then locate all hallucinated spans S, defined as portions of text in y that lack support from c. Additionally, to maintain expressive quality, we incorporate pairwise preference data, encouraging the model to evaluate responses based on helpfulness, relevance, conciseness, creativity, and completeness, similar to Arena-Hard (Li et al., 2024a).

Rule-Based Reinforcement Learning. We fine-tune the HA-GRM using rule-based online RL, adopting the GRPO setting (Shao et al., 2024). During rollout, the model predicts unfaithful spans and pointwise response-level labels for hallucination detection, while generating pairwise comparisons for creative writing. The reward signal combines format validity and prediction accuracy.

For hallucination detection, we incorporate the span-F1 metric alongside response-level accuracy. Let $\hat{S}$ be the predicted hallucination spans and S be the ground-truth spans. The hallucination reward r_h is

$$r_h = \begin{cases} 1 + \textit{span-F1}(S, \hat{S}), & \text{correct format \& acc.} \\ 1, & \text{correct format only} \\ 0, & \text{incorrect format} \end{cases} \tag{1}$$

As for the creative writing, r_w is

$$r_w = \begin{cases} 2, & \text{correct format \& acc.} \\ 1, & \text{correct format only} \\ 0, & \text{incorrect format} \end{cases} \tag{2}$$

These reward schemes are applied separately to the pointwise hallucination data and pairwise creative writing data.

2.2 HARPO: Hallucination-Aware Reinforcement for Policy Optimization

We now present the details of HARPO, a framework designed to balance the competing objectives of hallucination reduction and creative generation. We employ reinforcement learning rather than direct Supervised Fine-Tuning. This approach enables us to optimize the model using unlabelled prompts without requiring ground-truth reference texts. To guide this process, we introduce two novel components: first, the Selective Activation Mechanism, which functions as the core scoring metric during optimization; and second, a data curriculum strategy that schedules the training distribution to

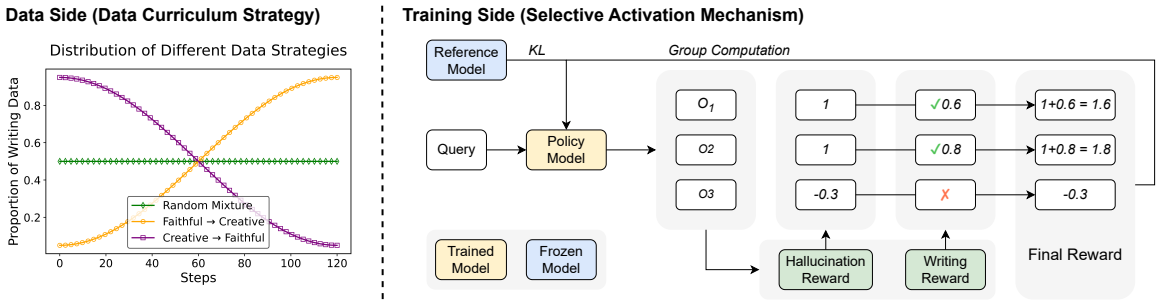

Figure 1: **The HARPO Framework.** Left: The **Data Curriculum Strategy** smoothly shifts the training distribution from creative to faithful tasks via cosine annealing. Right: The **Selective Activation Mechanism** gates optimization, ensuring creativity rewards are only assigned to hallucination-free responses.

smooth the transition between hallucination-centric and creativity-centric tasks.

Reinforcement Learning with GRPO. As our reinforcement learning framework, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a variant of REINFORCE (Williams, 1992) that utilizes group-based advantage normalization. To optimize the LLM policy π_θ, given a prompt x, we sample a group of K rollouts $\{\tau_k\}_{k=1}^{K}$ from the current policy π_θ. Each rollout τ_k is assigned a reward score $R_k = R(\tau_k)$ calculated by our reward system. The policy is updated via the clipped policy gradient objective:

$$\mathcal{L}_{\text{GRPO}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\tau_k|}\sum_{t=1}^{|\tau_k|}\min\Big[w_{k,t}(\theta)\cdot\hat{A}_{k,t},\ \text{clip}(w_{k,t}(\theta),1-\epsilon_{\text{low}},1+\epsilon_{\text{high}})\cdot\hat{A}_{k,t}\Big] - \beta\cdot D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}). \quad (3)$$

where the token-level importance weight is defined as the probability ratio:

$$w_{k,t}(\theta) = \frac{\pi_\theta(\tau_{k,t}|x,\tau_{k,<t})}{\pi_{\theta_{\text{old}}}(\tau_{k,t}|x,\tau_{k,<t})}. \quad (4)$$

with $\tau_{k,t}$ denoting the token generated at step t in rollout k, and $\tau_{k,<t}$ representing the preceding context. The group-normalized advantage $\hat{A}_{k,t}$ is computed using the rewards of the K rollouts sampled for the same input:

$$\hat{A}_{k,t} = \frac{R(\tau_k) - \text{mean}(\{R_i\}_{i=1}^{K})}{\text{std}(\{R_i\}_{i=1}^{K}) + \delta}. \quad (5)$$

where δ is a small constant for numerical stability. Following recent advances in RL for LLM reasoning such as DAPO (Yu et al., 2025), we adopt asymmetric clipping (clip-higher) with $\epsilon_{\text{low}} < \epsilon_{\text{high}}$. This strategy encourages the exploration of novel token sequences while maintaining training stability. The coefficient β scales the KL divergence term $D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$, ensuring the policy remains close to the reference model π_{ref} to prevent mode collapse.

Selective Activation Mechanism (SAM). A core challenge in aligning LLMs for both reliability and quality is the potential conflict between objectives. Standard approaches typically employ a linear scalarization of rewards, $R_{\text{total}} = w_1 R_{\text{hallu}} + w_2 R_{\text{creat}}$. However, this fixed-weight strategy creates an optimization landscape where the model may learn to exploit the reward function. To prevent this, we propose the Selective Activation Mechanism (SAM), a hierarchical reward aggregation strategy that treats factual correctness as a strict prerequisite for creativity.

Let x denote the input prompt and y denote the generated response corresponding to a rollout τ_k. Our HA-GRM, described in Section 2.1, provides two distinct scalar signals:

1. **Hallucination Reward (R_{hallu}):** A score indicating whether the response y is grounded in the context. Let $\hat{S}$ denote the set of hallucinated spans identified by the model, and let $|\cdot|$ denote the text length. The hallucination reward is calculated as:

$$R_{\text{hallu}}(x,y) = \begin{cases} 1, & \text{if } \hat{S} = \emptyset \\ -\frac{|\hat{S}|}{|y|}, & \text{if } \hat{S} \neq \emptyset \end{cases} \quad (6)$$

This formulation imposes a penalty proportional to the density of hallucinated content when errors occur.

2. **Creativity Reward (R_{creat}):** A score evaluating the helpfulness, relevance, conciseness, and novelty of the response. This score is derived from a pairwise evaluation comparing the model's output y against a reference answer r.

Under the SAM framework, the creativity reward is strictly gated by the factuality assessment. The final reward R_{SAM} assigned to a rollout is defined as:

$$R_{\text{SAM}}(x, y, r) = R_{\text{hallu}}(x, y) + \mathbb{I}[y \text{ is faithful}] \cdot R_{\text{creat}}(x, y, r). \quad (7)$$

where $\mathbb{I}[\cdot]$ is the indicator function. In our implementation, the condition "y is faithful" is satisfied only if no hallucinations are detected (i.e., $\mathbb{I} = 1$ if $\hat{S} = \emptyset$, and 0 otherwise). By dynamically decoupling the two objectives, SAM ensures that the pursuit of high-quality writing never comes at the cost of trustworthiness.

Data Curriculum Strategy. The distinct nature of our two types of training data, creative writing and hallucination mitigation, presents a significant optimization challenge. Creative writing tasks demand open-ended generation with diverse vocabulary, whereas hallucination mitigation requires strict adherence to evidence and factual constraints. Directly mixing these datasets in a static ratio often leads to optimization interference.

To address this, we design a dynamic data curriculum that modulates the training distribution over time. We adopt a *Creative* $\rightarrow$ *Faithful* progression strategy for two reasons: (1) **Model Affinity:** We initialize training with a higher proportion of creative, which is more general. This leverages the pre-trained model's inherent linguistic capabilities, stabilizing the initial policy optimization before introducing strict factual constraints. (2) **Prevention of Catastrophic Forgetting:** Rather than abruptly switching tasks, we employ a smooth decay schedule. This ensures that even as the focus shifts toward hallucination mitigation, a minimum threshold of general writing data is maintained to preserve language proficiency.

Formally, let T denote the total number of training steps. We define $r(t)$ as the sampling ratio of creative writing data at step t. We utilize a cosine annealing schedule to smoothly reduce this ratio from an initial high to a fixed minimum:

$$r(t) = r_0 \cdot \cos\left(\frac{\pi}{2T} \cdot t\right) + r_{\min}. \quad (8)$$

where $r_{\min} = 0.05$ represents the minimum proportion of creative data maintained at the end of training to prevent forgetting, and $r_0 = 0.9$ represents the magnitude of the decay. Under this schedule, the training begins with a creative data ratio of $r(0) = r_0 + r_{\min}$ and gradually converges to $r(T) = r_{\min}$ as the model becomes more robust to hallucination-sensitive tasks.

3 Experiments on Hallucination-Aware Generative Reward Model

In this section, we mainly answer this research question: (**RQ1**) What are the advantages of using reinforcement learning with fine-grained rewards for training generative reward models?

3.1 Experimental Setup

Training Details. We initialize HA-GRM from Qwen3-4B and train it on two public datasets. Specifically, for hallucination detection, we use the summarization and Question Answering subset of RAGTRUTH (Wu et al., 2023) training set. For creative writing, we utilize ARENA-HARD-V2.0 (Li et al., 2024b). We employ verl (Sheng et al., 2025) as our RL training framework, using a global batch size of 128 and a learning rate of 1e-6. Additionally, we apply a clipping strategy with the clip ratio high set to 0.28. All experiments are run once.

Evaluation Datasets and Metrics. We evaluate the Reward Models on the RagTruth test set for hallucination detection and Auto-J Eval (Li et al., 2023a) for general preference evaluation. In addition to response-level metrics such as precision, recall, and F1 scores, we conduct span-level detection by calculating the overlap between detected and human-labeled spans, reporting the character-level F1 score. For the pairwise evaluation on Auto-J, we utilize Accuracy. All datasets used in this evaluation are human-annotated.

3.2 Results and Analysis

Main Results. We present the response-level performance of various models in Table 1. Our proposed HA-GRM (initialized from Qwen3-4B) achieves an average F1 score of 78.08%, significantly outperforming the base model (74.58%) and the supervised fine-tuning baseline (66.37%). Notably, while standard SFT surpass ours in some

Model Name	RAGTruth(Summ)			RAGTruth(QA)			RAGTruth(D2T)			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Open-Source Models (8B - 671B)												
Qwen3-8B	50.14	87.25	63.69	36.81	88.12	51.93	79.76	91.19	85.09	60.50	89.82	72.30
Qwen3-32B	42.76	91.18	58.22	33.64	91.88	49.25	80.12	94.28	**86.62**	56.54	93.20	70.39
Qwen3-235B-A22B-Instruct	40.25	96.55	56.81	28.65	95.62	44.09	66.55	99.48	79.75	48.99	98.19	65.37
Qwen3-235B-A22B-Thinking	37.94	97.01	54.55	28.95	95.00	44.38	71.84	98.44	83.07	50.03	97.55	66.14
DeepSeek-V3.2	32.05	96.97	48.18	26.83	96.86	42.02	65.75	99.83	79.28	44.97	98.71	61.79
Proprietary Models												
GPT-4o	50.29	86.43	63.59	32.48	86.39	47.21	69.60	96.98	81.04	55.67	92.95	69.64
doubao-seed-1-6-thinking	41.81	95.10	58.08	37.29	96.25	53.75	74.47	98.26	84.73	55.83	97.23	70.94
claude-4-sonnet	46.32	86.27	60.27	36.65	87.50	51.66	72.95	94.91	82.49	54.83	91.21	68.49
Open-Source Models (Qwen3-4B based)												
Qwen3-4B	60.92	71.43	65.76	43.66	73.12	54.67	84.75	86.09	85.41	60.92	80.67	74.58
HA-GRM (*vanilla SFT*)	78.91	56.86	66.10	71.61	69.38	**70.48**	81.61	52.33	65.09	81.04	56.20	66.37
HA-GRM (*Reeponse Only*)	77.44	62.25	69.02	63.79	69.38	66.47	91.22	77.34	83.71	82.73	72.72	77.40
HA-GRM (*Full*)	81.05	60.78	**69.47**	70.75	65.00	67.75	91.98	77.20	83.94	85.88	71.58	**78.08**

Table 1: Performance (Precision, Recall and F1) of hallucination evaluation on the test split of RAGTruth. We separate Qwen3-4B based models, larger open-source models, and proprietary models into different blocks.

areas, it suffers from catastrophic forgetting, particularly on the Data-to-Text (D2T) task, a domain not included in our training set, where the score drops precipitously from 85.41% to 65.09%. In contrast, our RL-based approach maintains robust generalization on the unseen D2T dataset (83.94%), demonstrating that the HA-GRM learns generalized hallucination patterns rather than merely overfitting to the training distribution. Furthermore, despite its smaller size (4B), HA-GRM yields competitive performance against significantly larger open-source models (e.g., Qwen3-32B) and proprietary models (e.g., GPT-4o) across the RAGTruth benchmarks, validating the effectiveness of our specialized reinforcement learning framework.

Model Name	Summ	QA	D2T	Overall
Qwen3-4B	41.43	34.00	38.08	37.33
HA-GRM (*vanilla SFT*)	45.94	61.26	30.13	44.58
HA-GRM (*Reeponse Only*)	48.53	48.74	39.31	44.43
HA-GRM (*Full*)	**52.10**	**53.63**	**47.26**	**50.38**

Table 2: Span-level hallucination detection performance (Span-F1) on the RAGTruth test set across different training strategies.

Can span level hallucination prediction help? To investigate the impact of fine-grained supervision, we compare the HA-GRM *(Full)* against an ablation variant, HA-GRM *(Response Only)*, which is trained solely with response-level binary rewards. As shown in Table 2, incorporating span-level rewards yields substantial improvements. The full model achieves a significantly higher Span-F1 score across all test subsets (Average: 50.38% vs. 44.43%). Crucially, this granular feedback also enhances response-level classification, as seen in Table 1, where the Average F1 score rises from 77.40% to 78.08%. This suggests that training the model to explicitly localize hallucinated spans acts as a dense reward signal, reducing the likelihood of it learning spurious correlations (such as response length or style) and forcing it to ground its judgments in hallucination detection.

Model Name	RAGTruth	Auto-J	Average
Qwen3-4B	74.58	57.22	65.90
w/ Hallucination	**78.41**	57.31	67.86
w/ Hallucination and General	78.08	**59.91**	**69.00**

Table 3: Impact of multi-objective training on Hallucination Detection (RAGTruth) and General Preference Evaluation (Auto-J) testsets.

How writing evaluation affect hallucination detection? A core premise of the HA-GRM is that hallucination detection should not come at the expense of general preference evaluation. In Table 3, we analyze the trade-off between these objectives. Adding the creative writing preference data (*w/ Hallucination and General*) boosts the Auto-J score from 57.31% to 59.91%, indicating a marked improvement in the model's ability to judge helpfulness and creativity. Importantly, this gain incurs a negligible penalty on hallucination detection performance, with the average Hallucination F1 dropping only slightly from 78.41% to 78.08%, confirming that HA-GRM successfully unifies faithful con-

Model Name	Hallucination Control ↓		Creative Writing ↑	General ↑
	HHEM	MULTIHOPRAG	ARENA-HARD (Creative)	ARENA-HARD (Hard Prompt)
Qwen2.5 series				
Qwen2.5-3B	11.07	4.69	2.33	2.07
w/ HARPO(Ours)	2.41	2.86	2.66	2.30
Qwen2.5-7B	5.05	1.88	4.68	4.63
w/ HARPO(Ours)	1.56	0.86	6.78	4.93
Qwen3 series				
Qwen3-1.7B	11.07	5.38	7.04	3.87
w/ HARPO(Ours)	3.97	2.03	8.53	4.05
Qwen3-4B	7.34	3.29	16.95	11.16
w/ Hallucination-Only	0.72	0.98	9.74	5.29
w/ Creativity-Only	24.31	2.66	43.35	11.45
w/ Linear Mixture	3.73	0.90	28.83	5.63
w/ HARPO (Ours)	1.08	1.02	27.54	12.35
Qwen3-8B	4.93	2.90	33.76	16.08
w/ HARPO(Ours)	1.56	1.17	41.43	16.52

Table 4: Comparison of HARPO against base models and baselines. (↓) indicates lower is better, (↑) indicates higher is better. HARPO achieves low hallucination rates while simultaneously improving creative writing scores.

straint checking with creative quality assessment without significant compromise.

Takeaway(RQ1) Reinforcement learning with fine-grained rewards prevents the catastrophic forgetting observed in supervised fine-tuning, enabling a compact model to learn generalized hallucination patterns and outperform significantly larger models across diverse domains.

4 Experiments on Hallucination-Aware Reinforcement for Policy Optimization

We aim to answer the research question: (RQ2) Can we leverage the feedback from HA-GRM to effectively mitigate the LLM's hallucinations while incentivizing its creative writing capabilities?

4.1 Experimental Setup

Dataset and Evaluation Metric. Our training corpus consists of three distinct datasets selected to balance faithful grounding with creative expression. For hallucination mitigation, we utilize the summarization subset of RAGTRUTH (Wu et al., 2023) and the question answering subset of HALUEVAL (Li et al., 2023b). To support creative generation, we curate a subset of the ARENA-HUMAN-PREFERENCE-140K (Chiang et al., 2024) dataset, filtering specifically for prompts related to creative writing. All experiments are run once.

For evaluation, we employ HHEM-2.1 (Bao et al., 2024) and MULTIHOPRAG to assess hallucination rates in summarization and QA tasks, respectively. General and creative writing capabilities are evaluated using ARENA-HARD-V2.0 (Li et al., 2024b). We report the hallucination rate using our best-performing HA-GRM as an automated judge, and we utilize the official ARENA-HARD-V2.0 pipeline for writing quality scores. Table 5 summarizes the statistics for all datasets.

Dataset	DOMAIN	# OF SAMPLES
Training data		
RagTruth	Hallucination-Summ	2,217
HaluEval	Hallucination-QA	1,452
arena-human-preference	Creative Writing	3,000
Testing data		
HHEM	Hallucination-Summ	831
MultiHopRAG	Hallucination-QA	2,556
Arena-Hard (Hard Prompt)	General	500
Arena-Hard (Creative)	Creative Writing	250

Table 5: Statistics of the training and testing datasets utilized in the HARPO experiments.

Baselines and Training Details. To evaluate HARPO, we compare it against three baselines: (1) ***Hallucination-Only***, trained exclusively on RAGTRUTH and HALUEVAL using only R_{hallu}; (2) ***Creativity-Only***, trained solely on ARENA-HUMAN-PREFERENCE-140K using only R_{creat}; and (3) ***Linear Mixture***, a combined-objective baseline. Following Peng et al. (2025), the Lin-

ear Mixture approaches multi-objective alignment via fixed scalarization. We normalize the creative reward to $[0, 1]$ and compute the arithmetic mean with the hallucination reward: $(R^{\text{norm}}_{\text{creat}} + R_{\text{hallu}})/2$. This serves as the primary state-of-the-art comparison for mixed-objective training. We use the same training data as HAPRO but different reward strategy for this baseline. As for the hyperparameters of reinforcement learning, we adopt the same configuration as HA-GRM for training.

4.2 Results and Analysis

Main Results. Table 4 presents the performance of HARPO across different model families (Qwen2.5, Qwen3) and parameter scales (1.7B to 8B). Our analysis yields two key observations:

1. Pitfalls of Single-Objective Training: The ablation on Qwen3-4B highlights the dangers of training on isolated task and objectives. The *Hallucination-Only* baseline achieves the lowest hallucination rates (0.72% on HHEM) but suffers from catastrophic forgetting in generation tasks, dropping the General score from 11.16% to 5.29%. Conversely, the *Creativity-Only* baseline improves writing style significantly but induces severe reward hacking, causing the hallucination rate to spike to 24.31%. This confirms that faithful grounding and creative expression are competing objectives that require careful balancing.

2. Universal Improvement across Scales: HARPO consistently achieves improvement over base models across all tested sizes. For instance, on Qwen2.5-7B, HARPO reduces the hallucination rate on HHEM from 5.05% to 1.56% while simultaneously boosting the Creative Writing score from 4.68% to 6.78%. This demonstrates that our framework effectively mitigates hallucinations without imposing the "alignment tax" typically associated with safety fine-tuning.

Effect of Selective Activation Mechanism (SAM). Comparing HARPO with the *Linear Mixture* baseline reveals the critical role of the Selective Activation Mechanism (SAM). While the Linear Mixture approach (standard scalarization) reduces hallucinations, it fails to preserve general capabilities, resulting in a General score of only 5.63% compared to HARPO's 12.35%. By linearly combining rewards, the baseline creates a landscape where the model struggles to resolve gradient conflicts between creativity and faithfulness. In contrast, SAM's hierarchical gating ensures that creativity is optimized only within the subspace of faithfully correct responses, allowing HARPO to maintain the highest general language proficiency among all fine-tuned models.

Curriculum Schedule	Hallucination Control ↓		Creative Writing ↑
	HHEM	MULTIHOPRAG	ARENA-HARD (Creative)
Qwen3-4B	7.34	3.29	16.95
Random Mixture (Static)	4.09	1.17	23.65
Faithful → Creative (Reverse)	9.64	1.72	**34.30**
Creative → Faithful (**Ours**)	**1.08**	**1.02**	27.54

Table 6: Ablation on Data Curriculum Strategies.

Effect of Data Curriculum Strategy. To validate our data curriculum strategy, we compare it against two alternative schedules as shown in the left of Figure 1: (1) *Random Mixture*, which samples tasks with a fixed probability throughout training; and (2) *Faithful* → *Creative*, a reverse schedule that begins with hallucination mitigation and transitions to creative writing.

Table 6 demonstrates the necessity of our proposed ordering. The *Random Mixture* approach achieves simultaneous hallucination reduction and creative writing enhancement, but its performance gains are smaller than those of our curriculum learning scheme. Besides, The *Faithful* → *Creative* strategy performs poorly on HHEM with even a higher hallucination rate than original Qwen3-4B.

In contrast, our *Creative* → *Faithful* strategy achieves the best balance. By initializing training with a high ratio of creative tasks, we leverage the model's pre-trained linguistic priors to stabilize the policy. As the curriculum progresses, the gradual introduction of faithful constraints acts as a refinement stage, pruning hallucinations without degrading the stylistic quality established in the early phases.

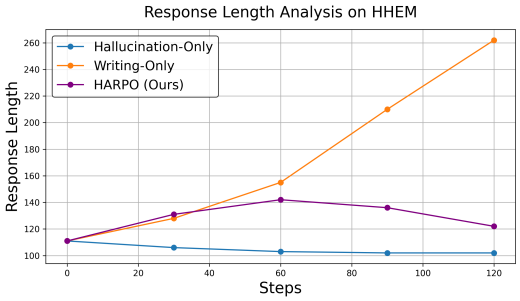

Figure 2: Response length analysis on HHEM.

Analysis of Response Length and Verbosity To investigate whether our framework mitigates "re-

ward hacking", specifically the tendency of RL-tuned models to inflate scores through verbosity, we track the evolution of response length throughout the training process (Figure 2). The *Writing-Only* baseline exhibits a linear, unbounded increase in token count, confirming that optimizing solely for creativity encourages excessive length. Conversely, the *Hallucination-Only* baseline rapidly converges to shorter, risk-averse responses, potentially sacrificing necessary detail to minimize hallucination rate. HARPO, however, demonstrates a distinct trajectory that balances these extremes. Initially, the response length increases, aligning with the high ratio of creative tasks in the early curriculum. However, as the training progresses beyond step 60, the length peaks and gradually stabilizes. This shift corresponds directly with our curriculum schedule and the imposition of the Selective Activation Mechanism; as the distribution shifts toward faithful tasks and the SAM strictness takes effect, the model is regularized against gratuitous verbosity while retaining sufficient length to remain expressive.

Takeaway(RQ2) HARPO effectively resolves the grounding-creativity conflict by gating rewards based on faithfulness. Combined with a dynamic curriculum, this approach minimizes hallucinations while boosting writing quality.

5 Related Work

Close-Domain Hallucination Detection Close-domain hallucination refers to generating content that is inconsistent with or unsupported by the context or knowledge provided in the input prompt (Jaech et al., 2024). Traditional detection methods frame this as a Natural Language Inference (NLI) task (Lattimer et al., 2023; Bao et al., 2024), where the source document serves as the premise and the generated text as the hypothesis. With the advancement of generative model performance, hallucination detection based on question answering arises (Honovich et al., 2021; Cattan et al., 2024). More recent methods explore intrinsic LLM metrics, analyzing features like token probabilities (Ridder and Schilling, 2025), hidden states (Zhang et al., 2025), and embedding distances (Ricco et al., 2025). The use of LLM-as-a-Judge has also become a prominent approach, employing both off-the-shelf and fine-tuned models (Bang et al., 2025; Wu et al., 2023). To our knowledge, we are the first to address this task using Reinforcement Learning with Verifiable Rewards.

Reinforcement Learning with Verifiable Rewards Reinforcement Learning with Verifiable Rewards (RLVR) is a prominent method for enhancing LLM reasoning (Guo et al., 2025), initially successful in structured domains like mathematics and programming (Wen et al., 2025) using algorithms such as PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024). While recent work has extended RLVR to complex fields like medicine and economics, it has not yet been applied to closed-domain hallucination detection, which we address here.

Reducing Hallucination for LLMs Research into mitigating LLM hallucinations has progressed across the model lifecycle. Initial pretraining strategies focus on curating high-quality corpora to reduce exposure to false or outdated knowledge (Penedo et al., 2023; Zhou et al., 2023). During fine-tuning, honesty-oriented SFT uses refusal and uncertainty samples (Sun et al., 2023; Wan et al., 2024) to teach models to acknowledge their limits. In the alignment stage, RLHF (Ouyang et al., 2022; Lightman et al., 2023) rewards factuality but can sometimes lead to over-conservative models that avoid answering due to fear of error (Wei et al., 2025). Finally, at inference time, model-agnostic techniques—such as factual-aware decoding (Lee et al., 2022), contrastive decoding (Chuang et al., 2023; Shi et al., 2024), and Chain-of-Verification (Dhuliawala et al., 2023)—provide dynamic ways to reduce hallucinations without retraining.

6 Conclusion

We presented HARPO, a reinforcement learning framework designed to reconcile the conflict between hallucination mitigation and creative generation. By integrating a fine-grained Hallucination-Aware Generative Reward Model (HA-GRM) with a Selective Activation Mechanism (SAM), our approach establishes faithfulness as a prerequisite for creative expression. Furthermore, our data curriculum strategy effectively prevents the catastrophic forgetting often observed in single-objective alignment. Extensive empirical results demonstrate that HARPO significantly reduces hallucination rates while simultaneously enhancing writing quality.

Limitations

Despite the promising results of HARPO, several limitations remain. First, training the Hallucination-Aware Generative Reward Model (HA-GRM) relies on high-quality, span-level hallucination annotations, which are significantly more labor-intensive and costly to acquire than standard binary labels. Second, the inference overhead introduced by the generative reward process increases the computational cost of training compared to traditional discriminator-based reward models. Finally, while we demonstrate effectiveness on models up to 8B parameters within general instruction and RAG domains, we have not yet verified the scalability of our framework on significantly larger models (e.g., 70B+) or its transferability to specialized reasoning tasks such as mathematics and code generation, which we leave for future investigation.

Ethics Statement

No datasets or scientific artifacts requiring dedicated ethical review, data privacy safeguards, or licensing arrangements were utilized or developed in this study, and we affirm that our work complies with the conference's ethical standards and carries no direct adverse social implications.

References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan. 2024. Localizing factual inconsistencies in attributable text generation. *arXiv preprint arXiv:2410.07473*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

S Dhuliawala, M Komeili, J Xu, R Raileanu, X Li, A Celikyilmaz, and J Weston. 2023. Chain-of-verification reduces hallucination in large language models, arxiv, 2023. *arXiv preprint arXiv:2309.11495*, 10.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Barrett Lattimer, Patrick H Chen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1703.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, and 1 others. 2023a. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024a. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Jianxing Liao, Tian Zhang, Xiao Feng, Yusong Zhang, Rui Yang, Haorui Wang, Bosi Wen, Ziying Wang, and Runzhi Shi. 2025. Rlmr: Reinforcement learning with mixed rewards for creative writing. *arXiv preprint arXiv:2508.18642*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. 2025. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. *arXiv preprint arXiv:2502.19328*.

Emanuele Ricco, Lorenzo Cima, and Roberto Di Pietro. 2025. Hallucination detection: A probabilistic framework using embeddings distance analysis. *arXiv preprint arXiv:2502.08663*.

Fabian Ridder and Malte Schilling. 2025. The hallurag dataset: Detecting closed-domain hallucinations in rag applications using an llm's internal states. In *AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, and 1 others. 2023. Moss: Training conversational language models from synthetic data. *arXiv preprint arXiv:2307.15020*, 7(3):3.

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge verification to nip hallucination in the bud. *arXiv preprint arXiv:2401.10768*.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, and 1 others. 2025. Truthrl: Incentivizing truthful llms via reinforcement learning. *arXiv preprint arXiv:2509.25760*.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, and 1 others. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Annual Meeting of the Association for Computational Linguistics*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *Annual Meeting of the Association for Computational Linguistics*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.