KnowRL: Exploring Knowledgeable Reinforcement Learning for Factuality

Baochang Ren[♠♡], **Shuofei Qiao**[♠♡], **Ningyu Zhang**[♠♡], **Da Zheng**[♢♡], **Huajun Chen**[♠♡*]
[♠] Zhejiang University [♢] Ant Group
[♡] Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph
{baochang.ren, shuofei, zhangningyu}@zju.edu.cn

Abstract

Slow-thinking Large Language Models (LLMs) have demonstrated strong reasoning capabilities but often suffer from severe hallucinations due to an inability to recognize their knowledge boundaries. Existing Reinforcement Learning (RL) approaches typically rely on outcome-oriented rewards, which can inadvertently reinforce fabricated reasoning paths when the final answer is correct. To address this, we propose **Know**ledge-enhanced **RL**, **KnowRL**, a framework that integrates factual supervision directly into the reasoning process. By decomposing the chain of thought into atomic facts and verifying them against the corresponding ground-truth knowledge, KnowRL performs fine-grained checks to encourage models to reason faithfully. Crucially, this process-oriented supervision teaches the model to identify its knowledge boundaries, learning to say "I don't know" instead of fabricating answers when information is missing. Experimental results demonstrate that KnowRL effectively mitigates hallucinations—reducing the Incorrect Rate on SimpleQA by 20.3% for distillation-based slow-thinking models while maintaining strong performance on complex reasoning benchmarks like GPQA and AIME 2025. Furthermore, our method shows robust transferability to out-of-distribution tasks, indicating that the model learns a generalizable verification behavior[1].

1 Introduction

Recent advancements, represented by models like DeepSeek-R1 (Guo et al., 2025), mark a paradigm shift towards "slow thinking". By leveraging Reinforcement Learning (RL) to encourage extended Chains of Thought (CoT), these models have achieved remarkable breakthroughs in complex reasoning tasks. However, a critical paradox

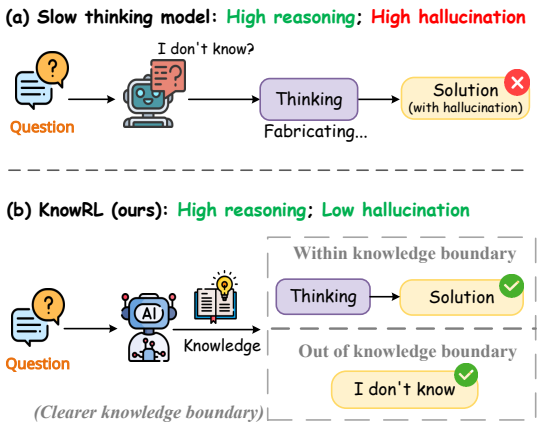


Figure 1: KnowRL reduces hallucinations in slow-thinking models.

has emerged: while scaling reasoning compute significantly boosts problem-solving abilities, it does not naturally align with factual reliability. In fact, slow thinking models often exhibit severe hallucinations (Heyman and Zylberberg, 2025; Patel et al., 2024; Arcuschin et al., 2025). As shown in Figure 2, larger reasoning models achieve higher GPQA (Rein et al., 2024) scores but fail to improve or even regress on hallucination benchmarks like SimpleQA (Wei et al., 2024). For instance, the DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025) achieves an accuracy of only 6.64% on the SimpleQA dataset. This suggests that without proper guidance, the long reasoning process can turn into a "snowball" of errors, where one small mistake leads to a completely fabricated conclusion. This raises a critical question: ***Why do models with such strong reasoning abilities still fail so badly at factual reliability?***

The root cause lies in the way we currently train these models. Standard RL heavily relies on outcome-oriented rewards, which optimize for the final answer while treating the reasoning process as a black box. This approach has two main flaws. *First*, it ignores the Knowledge Boundary—the model's ability to distinguish between what it

*Corresponding author.
[1] https://github.com/zjunlp/KnowRL.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39640–39658
July 2-7, 2026

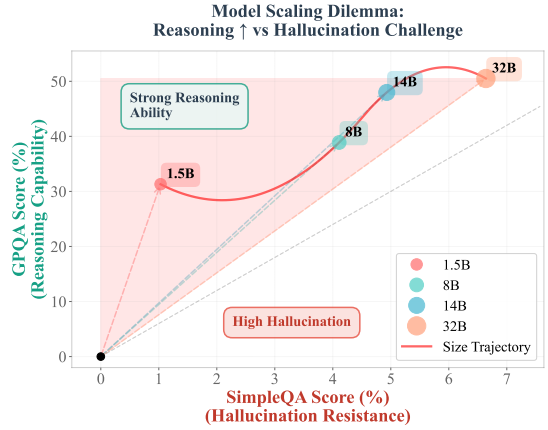


Figure 2: Scaling improves reasoning ability (GPQA) but does not reduce hallucinations (SimpleQA). Results are shown for DeepSeek-R1-Distill models of varying sizes (Qwen-1.5B, Llama-8B, Qwen-14B, Qwen-32B).

knows and what it does not know. Without this boundary, models try to guess the answer to get the reward. *Second*, it creates a supervision gap: model might generate a correct answer using wrong or made up reasoning steps. Since the reward signal depends solely on the final outcome, the model mistakenly learns that fabricating reasoning paths is a valid strategy; therefore, the RL algorithm unwittingly reinforces this hallucinated logic. Consequently, the model becomes "smart" at reasoning but "dishonest" about facts.

Addressing this challenge with existing methods proves difficult. Retrieval Augmented Generation (RAG; Lewis et al., 2020) faces significant retrieval efficiency bottlenecks when integrated into the extensive reasoning steps of slow thinking models. Supervised Fine Tuning (SFT; Ouyang et al., 2022) primarily relies on static knowledge injection. However, this paradigm suffers from severe catastrophic forgetting (Chen et al., 2025a), often degrading the model's inherent reasoning capabilities. More fundamentally, SFT encourages the model to merely memorize static knowledge rather than learning the generalized behavior of reasoning. Consequently, it fails to instill the critical strategy of "knowing what you know and what you do not know"—a dynamic judgment of Knowledge Boundaries. In contrast, Reinforcement Learning is uniquely suited to shape this behavioral strategy (Gandhi et al., 2025). Therefore, there is an urgent need for a RL framework that can inherently instill factual discipline into the model's reasoning behaviors without compromising its reasoning capabilities.

So, to bridge this gap, we propose Knowledge enhanced Reinforcement Learning (KnowRL), a framework that integrates factuality supervision directly into the RL reasoning loop. Unlike outcome-based approaches, KnowRL opens the "black box" of thinking. By decomposing the Chain of Thought into atomic facts and verifying them against the corresponding ground-truth knowledge, KnowRL provides dense, process level rewards. This design transforms the RL objective: instead of merely "getting the answer right," the model is incentivized to reason faithfully and, crucially, to recognize its knowledge boundaries—learning to say "I don't know" rather than fabricating a plausible sounding response when information is missing, as illustrated in Figure 1.

Experimental results validate the effectiveness of KnowRL. On hallucination benchmarks, KnowRL significantly reduces the error rate. For instance, dropping the Incorrect Rate on SimpleQA by 20.3% for 7B slow thinking model. Crucially, this gain in factuality does not come at the cost of reasoning ability; KnowRL maintains strong performance on complex reasoning datasets like GPQA and AIME 2025. Furthermore, our method demonstrates strong transferability and Out-Of-Distribution (OOD) generalization to knowledge domains outside the training distribution; it significantly improves performance on ChineseSimpleQA (He et al., 2024b) even when the primary knowledge source is purely English-based. These findings suggest that KnowRL helps the model internalize a universal verification behavior rather than merely memorizing language-specific facts. These results confirm that KnowRL successfully aligns the slow thinking process with factual accuracy, offering a robust path for building reliable reasoning models.

2 Knowledgeable Reinforcement Learning

To address the high hallucination rates in slow-thinking models, we propose a fundamental shift in supervision: focusing on the ***reasoning process*** rather than only on final outcomes. To this end, we introduce KnowRL, a Knowledge enhanced Reinforcement Learning framework designed to guide models toward verifiable and boundary-aware reasoning by integrating factual supervision directly into the RL loop. As illustrated in Figure 3, we first construct training data matched with knowledge, and then conduct RL training using a composite

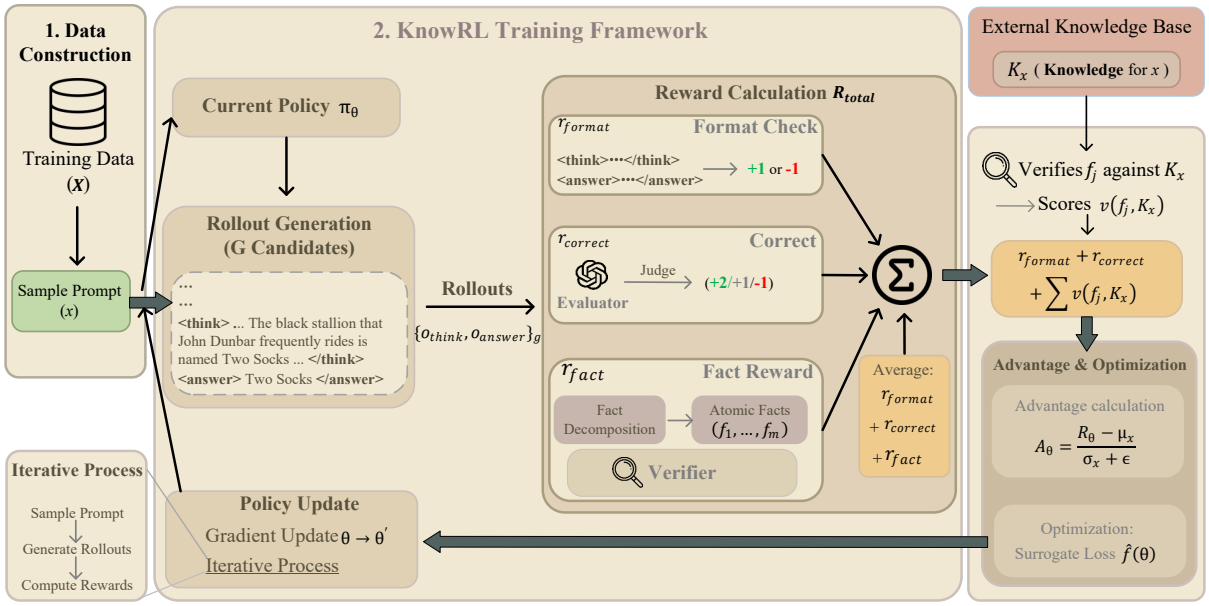


Figure 3: **Overview of the KnowRL framework.** We first construct training data matched with external knowledge. During RL training, we decompose the generated reasoning process into atomic facts and verify them against the knowledge base. Finally, we optimize the model using a composite reward that combines these factual scores with correctness and format checks.

reward that balances factual correctness with reasoning quality (details in Appendix A). By enforcing factual consistency at the step level, KnowRL effectively corrects incorrect reasoning paths and encourages models to recognize their ***knowledge boundaries***, ultimately leading to a more faithful and self-aware thinking process.

KnowRL builds upon the Group Relative Policy Optimization (GRPO) algorithm by introducing *factuality supervision* into the CoT reasoning process. This integration ensures that the model is guided not just by the final answer, but by the factual accuracy of its thinking steps. In the following, we detail how we design the composite Reward Function to evaluate these steps and how we implement the Factuality-Guided Policy Optimization.

Reward Function. Given a rollout $o = (o_{\text{think}}, o_{\text{answer}})$, inspired by FactScore (Min et al., 2023), we use GPT-4o-mini for the factual verification process. We decompose the reasoning trace into M atomic facts $\Phi(o_{\text{think}}) = \{f_1, \ldots, f_M\}$. Given an external knowledge base K, each atomic fact f_j is checked against its most relevant knowledge set $K_x \subseteq K$[2] to obtain a verification score $v(f_j, K_x) \in \{0, 1\}$. The factuality reward is then defined as the proportion of supported facts:

$$r_{\text{fact}}(o) = \begin{cases} \frac{1}{M} \sum_{j=1}^{M} v(f_j, K_x), & M > 0, \\ 0, & M = 0. \end{cases} \quad (1)$$

Additional components include a ***format reward*** $r_{\text{format}}(o)$ and a ***correct reward*** $r_{\text{correct}}(o)$. The format reward verifies whether the output follows the required `<think>...</think><answer>...</answer>` structure: if the format is valid, $r_{\text{format}}(o) = +1$; otherwise, $r_{\text{format}}(o) = -1$. The correctness reward evaluates the final answer o_{answer} using an evaluator model (GPT-4o-mini): if the answer is correct, $r_{\text{correct}}(o) = +2$; if the model explicitly refuses, $r_{\text{correct}}(o) = +1$; if the answer is incorrect, $r_{\text{correct}}(o) = -1$. The composite reward is then defined as

$$R_{\text{total}}(o) = r_{\text{format}}(o) + r_{\text{correct}}(o) + r_{\text{fact}}(o). \quad (2)$$

Factuality–Guided Policy Optimisation. For every prompt x the current policy $\pi_{\theta_{\text{old}}}$ generates a *group* of G candidate roll-outs $\{o^{(g)}\}_{g=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$, where G is the group size. Each trajectory is scored by the composite reward $R_{\text{total}}(\cdot)$, yielding a set of scalars $\mathcal{R}_x = \{R_g\}_{g=1}^{G}$.

Advantage construction. We summarise $\mathcal{R}_x$ with its sample mean μ_x and standard deviation σ_x. The

[2]Retrieved using `sentence-transformers/gtr-t5-large`.

credit assigned to trajectory g is the *group-relative advantage*

$$A_g = \frac{R_g - \mu_x}{\sigma_x + \varepsilon}, \tag{3}$$

where $\varepsilon(\ll 1)$ avoids division by zero. A_g is positive if $o^{(g)}$ attains above-average factual reward and negative otherwise, turning factual supervision into a signed learning signal.

Likelihood ratio. Define the trajectory-level importance ratio

$$\varrho_g = \frac{\pi_\theta\big(o^{(g)} \mid x\big)}{\pi_{\theta_{\text{old}}}\big(o^{(g)} \mid x\big)}.$$

The pair (ϱ_g, A_g) fully characterises how $o^{(g)}$ should influence the update: *increase* its probability if $\varrho_g A_g > 0$ and *decrease* otherwise.

Surrogate objective. Using these short symbols enables a compact, column-friendly surrogate:

$$\hat{\mathcal{J}}(\theta) = \frac{1}{G}\sum_{g=1}^{G} \min\big(\varrho_g A_g,\ \text{clip}(\varrho_g, 1-\epsilon, 1+\epsilon) A_g\big), \tag{4}$$

with a small clip threshold ϵ to bound the update. Because A_g encodes factual advantage, maximising $\hat{\mathcal{J}}$ explicitly transfers probability mass from hallucination-heavy traces ($A_g < 0$) to trajectories whose CoT is knowledge-supported ($A_g > 0$).

Regularised loss. To preserve exploration and limit divergence from the frozen reference policy π_{ref}, we add (i) an entropy bonus and (ii) a KL anchor. Let o denote a complete rollout and o_t its t-th token. The token-level Shannon entropy on prompt x is therefore $\mathcal{H}(\pi_\theta(\cdot|x)) = -\sum_t \pi_\theta(o_t|x) \log \pi_\theta(o_t|x)$. Its mini-batch expectation and the corresponding KL expectation are $\mathcal{E}_{\mathcal{H}} \triangleq \mathbb{E}_x[\mathcal{H}(\pi_\theta(\cdot|x))]$ and $\mathcal{E}_{\text{KL}} \triangleq \mathbb{E}_x[D_{\text{KL}}(\pi_\theta(\cdot|x) \| \pi_{\text{ref}}(\cdot|x)]$. With scalar weights $\beta_{\mathcal{H}}, \beta_{\text{KL}} > 0$, the final objective becomes

$$\mathcal{L}_{\text{KnowRL}} = -\hat{\mathcal{J}}(\theta) + \beta_{\mathcal{H}}\mathcal{E}_{\mathcal{H}} + \beta_{\text{KL}}\mathcal{E}_{\text{KL}}. \tag{5}$$

During training, each gradient step proceeds as follows. First, a mini-batch of prompts x is sampled and the old policy $\pi_{\theta_{\text{old}}}$ produces a group $\mathcal{O}_x$ of G roll-outs. Next, every trajectory $o^{(g)} \in \mathcal{O}_x$ undergoes the knowledge check that yields the composite reward $R_g = R_{\text{total}}(o^{(g)})$, whose factual component r_{fact} is computed against the external knowledge base. The set $\{R_g\}$ is then converted into group-relative advantages A_g via Eq. (3), so that a higher proportion of verified facts directly translates into a larger positive A_g, whereas hallucinated chains of thought receive negative credit. The new policy π_θ is updated by maximising the factual surrogate $\hat{\mathcal{J}}(\theta)$ in Eq. (4) while minimising the regularised loss $\mathcal{L}_{\text{KnowRL}}$ in Eq. (5); this couples the likelihood ratio ϱ_g with A_g so that gradients *increase* the probability of knowledge supported reasoning and *decrease* that of hallucination prone trajectories. Iterating this loop prompt sampling, factual verification, advantage normalisation, and loss-driven update—yields a policy that systematically suppresses hallucinations yet preserves answer accuracy.

Sub-task	DeepSeek-7B	Skywork-7B
Mathematics		
COMP.en	8.61% → 8.46%	4.75% → 7.12%
COMP.zh	7.35% → 7.60%	5.39% → 5.39%
CEE.zh	10.65% → 10.56%	8.79% → 9.35%
Physics		
COMP.en	1.69% → 0.00%	0.85% → 1.69%
CEE.zh	7.83% → 11.30%	3.48% → 5.22%
Average Acc.	**7.23% → 7.58%**	**4.65% → 5.75%**

Table 1: **Performance on OlympiadBench.** Results are shown as Base model → KnowRL-trained.

3 Experiments

3.1 Experimental Settings.

Datasets and Metrics. We use TruthfulQA (Lin et al., 2021), SimpleQA, and ChineseSimpleQA to evaluate hallucination, and GPQA (general domain) and AIME 2025 (mathematical reasoning) to evaluate reasoning ability. For the hallucination evaluation datasets, we randomly sample 300 examples from each of TruthfulQA, SimpleQA, and ChineseSimpleQA as test sets. Because slow-thinking models generate thousands of tokens per query, evaluating full datasets is computationally prohibitive. A 300-example subset remains highly challenging and provides a statistically valid evaluation, which is a standard practice for long-reasoning tasks. TruthfulQA is assessed with the ROUGE and BLEU metrics. For SimpleQA and ChineseSimpleQA, we evaluate performance using four metrics: (1) Incorrect Rate, (2) Refusal Rate, (3) Precision on Answered Questions (PAQ), and (4) F_1 score. PAQ measures the proportion of correctly answered questions among those that the model chooses to answer. For AIME 2025 and GPQA, we assess reasoning performance based on accuracy metric. All models are evaluated with the temperature set to 0.

Methods	Hallucination										Reasoning	
	TruthfulQA		SimpleQA				ChineseSimpleQA				GPQA Diamond	AIME
	Rouge	Bleu	PAQ	Incorrect	Refusal	F1	PAQ	Incorrect	Refusal	F1		
Skywork-OR1-7B-Preview												
Zero-shot	56.67	**55.33**	2.97	76.33	21.33	2.61	11.84	67.00	24.00	10.23	37.37	26.67
Self-Refine	58.00	54.00	3.90	74.00 (-2.33)	23.00	3.90	9.82	67.33 (+0.33)	25.33	8.40	46.67	36.67
FactTune-FS	**58.33**	50.33	0.76	43.33 (-33.0)	**56.33**	0.46	8.52	68.00 (+1.00)	25.67	7.26	40.91	26.67
DPO	52.67	49.33	4.81	85.66 (+9.33)	10.00	4.56	12.64	78.33 (+11.3)	10.33	11.95	34.85	30.00
SFT	57.67	51.67	**11.45**	77.33 (+1.00)	12.67	**10.68**	**19.70**	70.67 (+3.67)	12.00	**18.44**	34.85	23.33
TruthRL	57.33	50.60	3.78	56.00 (-20.3)	41.67	2.95	10.58	62.00 (-5.00)	30.67	8.66	39.39	26.67
KnowRL	57.67	54.33	3.21	60.33 (-16.0)	37.67	2.46	12.29	52.33 (-14.7)	**40.33**	9.19	42.42 ↑	36.67 ↑
DeepSeek-R1-Distill-Qwen-7B												
Zero-shot	53.33	51.00	2.09	78.00	20.33	1.86	8.07	68.33	25.67	6.88	40.91	30.00
Self-Refine	55.00	50.33	2.52	77.33 (-0.67)	20.67	2.23	8.11	68.00 (-0.33)	26.00	6.90	45.45	33.33
FactTune-FS	54.00	50.00	2.72	59.67 (-18.3)	38.67	2.07	10.24	76.00 (+7.67)	15.33	9.39	38.89	30.00
DPO	54.00	51.00	4.35	88.00 (+10.0)	8.00	4.16	13.14	79.33 (+11.0)	8.67	12.54	37.37	30.00
SFT	**57.67**	52.00	**8.42**	83.33 (+5.33)	9.00	**8.03**	**19.58**	76.67 (+8.34)	4.67	**19.11**	36.36	26.67
TruthRL	**57.67**	**54.67**	2.14	61.00 (-17.0)	37.67	1.64	5.76	60.00 (-8.33)	**36.33**	4.48	39.39	26.67
KnowRL	57.33	51.60	2.81	57.67 (-20.3)	**40.67**	2.09	10.26	58.33 (-10.0)	35.00	8.08	36.87 ↓	33.33 ↑

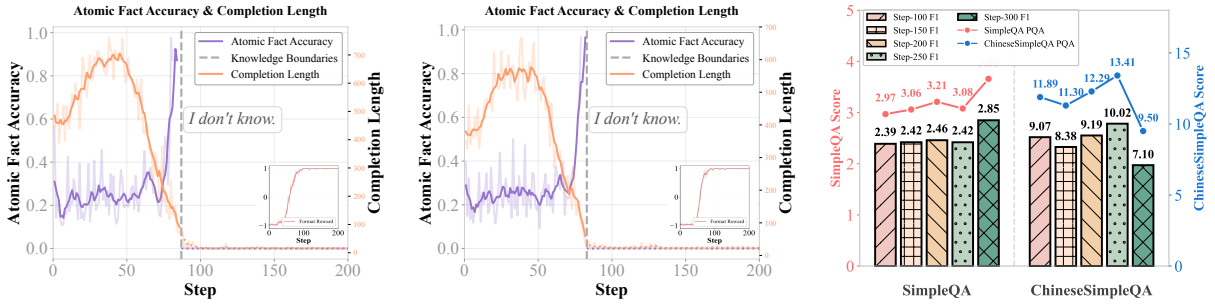
Table 2: **Main Experimental Results.** Performance of KnowRL compared against baselines on the Skywork-OR1-7B-Preview and DeepSeek-R1-Distill-Qwen-7B models. Zero-shot refers to the original model performance. The best results are marked in **bold**.

Models and Baselines. We select the Skywork-OR1-7B-Preview (He et al., 2025a) and the DeepSeek-R1-Distill-Qwen-7B for experiments. These models represent the two most popular slow thinking training methods: RL, represented by the Skywork-OR1-7B-Preview model, and distillation, represented by the DeepSeek-R1-Distill-Qwen-7B model. We select Self-Refine (Madaan et al., 2023) as the baseline for prompt engineering, and SFT, DPO, FactTune-FS (Tian et al., 2023) and TruthRL (Wei et al., 2025) as the baselines for post-training methods. For DPO, we use the distilled DeepSeek-R1 data as the chosen data; for FactTune-FS, the chosen data is composed of the original model's outputs, which are filtered for high factuality using FactScore. We use Low-Rank Adaptation (LoRA; Hu et al., 2022) to train all models. Further training details are provided in the Appendix B.

3.2 Main Results

KnowRL significantly mitigates hallucination while maintaining reasoning capabilities. As shown in Table 2, KnowRL consistently improves factual reliability across all datasets. For the DeepSeek-R1-Distill-Qwen-7B model, the Incorrect Rate on SimpleQA drops by over 20% (from 78.00% to 57.67%), with similar improvements observed on ChineseSimpleQA. Notably, since our training knowledge source is primarily English-based, the consistent gains on ChineseSimpleQA highlight strong cross-lingual transferability, suggesting that KnowRL helps the model internalize a language-agnostic verification behavior rather than merely memorizing language-specific factual patterns. Importantly, unlike baseline methods that often suffer from catastrophic forgetting, KnowRL preserves strong performance on complex reasoning benchmarks. For instance, it improves the GPQA score from 37.37% to 42.42% and maintains stability on AIME 2025, demonstrating that factual supervision does not conflict with deep reasoning abilities. This favorable trade-off can be attributed to KnowRL's process-level design: by rewarding the factual grounding of individual reasoning steps rather than solely adjusting outcome distributions, the model retains its capacity for extended logical deduction while learning to ground each step in verifiable knowledge. Furthermore, the dramatic drop in completion length (Figure 4(a) and (b)) reflects acquired boundary-awareness rather than a collapse of the slow-thinking process. The model efficiently truncates wasteful hallucinations for unknown facts while preserving extended reasoning chains for complex logical tasks.

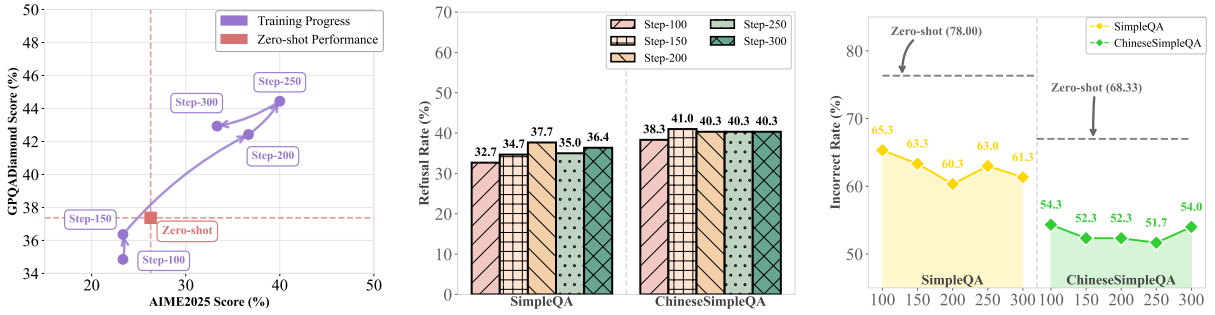
Training dynamics demonstrate that the model progressively learns to recognize knowledge boundaries. Figure 4(a) and (b) illustrates how KnowRL shapes the model's behavior. During training, the accuracy of atomic facts within the chain-of-thought steadily increases. As the model



(a) The Training Curves (Skywork-OR1-7B-Preview)

(b) The Training Curves (DeepSeek-R1-Distill-Qwen-7B)

(c) F1 and PAQ Across Training Steps on SimpleQA and ChineseSimpleQA



(d) Training Progress on GPQA and AIME 2025

(e) Refusal Rates Across Training Steps on SimpleQA and ChineseSimpleQA

(f) Incorrect Rates Across Training Steps vs. Zero-shot Performance

Figure 4: KnowRL training dynamics and reasoning behavior analysis. (a)–(b) present the training curves of two model architectures, showing improved factual alignment and stable reasoning length. (c)–(f) illustrate performance trends of Skywork-OR1-7B-Preview across different training steps, including F1 and PAQ (c), reasoning accuracy (d), refusal rate (e), and incorrect rate (f).

becomes more precise in its reasoning steps, it learns to distinguish known facts from unknown information. Consequently, instead of fabricating answers when uncertain, the model learns to abstain, reflected by a rise in the Refusal Rate and a sharp drop in the Incorrect Rate. This confirms that KnowRL effectively teaches the model to reason faithfully rather than merely optimizing for lucky guesses.

3.3 Extended Evaluation on Reasoning Benchmarks

We further evaluate KnowRL on OlympiadBench (He et al., 2024a) to examine its impact on complex reasoning in mathematics and physics. As shown in Table 1, both models maintain or improve their performance after KnowRL training. Specifically, the DeepSeek model increases its average accuracy from 7.23% to 7.58%, while the Skywork model improves from 4.65% to 5.75%. Both models achieved significant gains in the Chinese physics task (*physics_zh_CEE*), with DeepSeek-7B rising from 7.83% to 11.30%. Although there are minor fluctuations in individual sub-tasks, the overall trend confirms that factual supervision does not hinder structured reasoning.

Combined with the consistent results on GPQA and AIME 2025, these findings provide strong evidence that **KnowRL effectively preserves and enhances complex reasoning capabilities across diverse other domains, enabling models to reason more reliably under factual constraints**.

3.4 Training Step Analysis

To investigate how the number of KnowRL training steps affects the trade-off between factuality and reasoning, we evaluate models trained for 100, 150, 200, 250, and 300 reinforcement learning steps. The results are illustrated in Figure 4(c)–(f).

As shown in Figure 4(c), both F1 and PAQ scores improve rapidly in the early stage of training and reach a stable high level after moderate training, suggesting that factual precision and reliability are primarily acquired in the middle phase of optimization. Reasoning performance on GPQA and AIME 2025 (Figure 4(d)) remains stable or slightly improves within this range, indicating that factual supervision can refine reasoning without inducing degradation. Refusal rates (Figure 4(e)) increase and then plateau, reflecting that models progressively learn appropriate abstention behavior, while the incorrect rate (Figure 4(f)) decreases sharply

Method	Hallucination								Reasoning	
	SimpleQA				ChineseSimpleQA				GPQA Diamond	AIME
	PAQ	Incorrect	Refusal	F1	PAQ	Incorrect	Refusal	F1		
Ablation on Reward Combination (Base Model: DeepSeek-R1-Distill-Qwen-7B)										
$R = r_{\text{format}}$	2.63	74.00	24.00	2.27	7.08	74.33	20.00	6.29	39.39	30.00
$R = r_{\text{format}} + r_{\text{fact}}$	2.42	80.67 (+6.67)	17.33	2.19	8.87	75.33 (+1.00)	17.33	8.03	**47.47**	**40.00**
$R = r_{\text{format}} + r_{\text{correct}}$	**3.19**	60.67 (-13.33)	37.33	**2.46**	8.89	41.00 (-33.33)	**55.00**	5.52	38.89	**40.00**
$R = R_{total}$ (KnowRL)	2.81	57.67 (-16.33)	**40.67**	2.09	**10.26**	58.33 (-16.00)	35.00	**8.08**	36.87	33.33
KnowRL ($R_{refusal} = -1$)	1.26	78.67 (+4.67)	20.33	1.11	7.66	84.33 (+10.00)	8.67	7.32	34.85	30.00
Robustness Analysis (Using GRPO reward, $R = R_{total}$)										
Zero-shot	2.09	78.00	20.33	1.86	8.07	68.33	25.67	6.88	40.91	30.00
KnowRL (DAPO)	1.11	59.33 (-18.67)	40.00	0.83	9.14	59.67 (-8.66)	34.33	7.24	41.41	43.33
KnowRL (BNPO)	1.61	61.00 (-17.00)	38.00	1.23	11.48	61.67 (-6.66)	30.33	9.43	41.41	30.00
KnowRL (Dr.GRPO)	3.16	61.33 (-16.67)	36.67	2.45	11.33	60.00 (-8.33)	32.33	9.15	38.38	43.33

Table 3: **Ablation and Robustness Analysis.** Impact of reward components and different RL algorithms for the DeepSeek-R1-Distill-Qwen-7B. In reward ablation section, best results are in **bold** and second-best are underlined.

in early training and stabilizes thereafter. Beyond this stage, additional optimization provides limited factual improvement and introduces minor task-dependent variations rather than consistent gains.

These observations suggest that KnowRL's reinforcement learning dynamics follow a stable convergence pattern rather than the reversal effects sometimes observed in overfitted reasoning models (Berglund et al., 2023). Appropriate training duration is therefore crucial: insufficient updates hinder factual learning, whereas excessive optimization may overfit factual supervision signals and weaken generalization.

Overall, KnowRL achieves its best balance between factuality and reasoning stability when trained for an appropriate number of reinforcement learning steps, where the model fully internalizes factual supervision without over-optimization or reasoning drift.

3.5 Ablation Study

To understand how different reward signals contribute to the model's performance, we evaluate the DeepSeek-R1-Distill-Qwen-7B model under different reward combinations.

Factual reward (r_{fact}) enhances fact-grounded reasoning. As shown in Table 3, using the factual reward alone achieves the best performance on reasoning benchmarks such as GPQA and AIME 2025. This shows that r_{fact} encourages models to perform reasoning grounded in verifiable knowledge, helping them maintain accuracy while reducing random errors. This finding underscores the importance of factual verification in the reinforcement learning process, consistent with prior work showing that incorporating verifiable rewards can substantially improve the reliability and stability of RL training (RLVR; Yue et al., 2025).

Positive refusal incentives promote boundary awareness. Table 3 shows that giving a positive reward for refusals leads to higher refusal rates and fewer hallucination errors. This explicitly encourages the model to admit when it lacks knowledge. When we removed this positive incentive (setting $R_{\text{refusal}} = -1$), the incorrect rate increased significantly. This confirms that rewarding appropriate refusals is essential for learning knowledge boundaries.

4 Analysis

4.1 Robustness Analysis of Different RL Algorithms

To verify that the effectiveness of KnowRL is not limited to a specific optimization method, we extended our framework to three additional RL algorithms: BNPO (Xiao et al., 2025), DAPO (Yu et al., 2025), and Dr.GRPO (Liu et al., 2025b). We compare their performance using the DeepSeek-R1-Distill-Qwen-7B model.

KnowRL is robust across different algorithms. Table 3 demonstrates that every algorithm achieved the dual goal of our framework: they all significantly reduced the Incorrect Rate compared to the Zero-shot baseline, while successfully maintaining or even improving reasoning performance. This confirms that the KnowRL framework effectively reduces hallucinations without compromising reasoning capabilities, regardless of the specific RL algorithm used. Detailed analysis of response length

and atomic fact accuracy during training is provided in Appendix D.

4.2 Analysis of Factors Affecting Knowledge Boundaries

To examine whether the positive refusal reward in r_{correct} is the only factor driving boundary learning, we conducted an experiment on the DeepSeek-R1-Distill-Qwen-7B model using SimpleQA and ChineseSimpleQA. We modified the KnowRL setting by changing the refusal reward from positive to negative $r_{\text{refusal}=-1}$, while keeping other rewards unchanged. We evaluated the model at training steps 0, 100, 150, 200, and 250.

As shown in Figure 5, the refusal rate initially rises even without a positive refusal reward.

This confirms that the positive refusal reward is not the only factor affecting knowledge boundaries; the penalty for incorrect answers also encourages the model to be cautious in the early stages. However, this behavior is unstable. Later stages in Figure 5 show a significant drop in refusal rates accompanied by a spike in incorrect answers. This occurs because the model learns to guess to maximize scores rather than accepting refusal penalties—known as reward hacking. Thus, while correctness rewards trigger initial caution, positive refusal rewards are essential to maintain it and prevent guessing. **Overall, the results suggest that positive refusal rewards are not the only source of boundary learning; different reward combinations can jointly promote boundary-aware reasoning.**

We further investigate the scalability and evaluator sensitivity of our framework. Experiments on the larger DeepSeek-R1-Distill-Qwen-14B model show that KnowRL consistently reduces hallucinations while improving reasoning capabilities, confirming its scalability across model sizes(details in Appendix E). Moreover, we conduct an Evaluator Sensitivity Analysis by replacing the GPT-4o-mini used during training with different model. The results indicate that KnowRL maintains robust performance regardless of the specific evaluator used. Comprehensive results for these analyses are provided in Appendix F.

5 Robustness and Multi-run Evaluation

To address the potential sensitivity of single-run evaluations at zero temperature, especially on small-scale benchmarks like AIME, we conduct a robust multi-run evaluation. We re-evaluate the base models and KnowRL using a non-zero temperature of 0.6, generating 5 responses per prompt to report the average performance (Avg@5). This setup ensures that the observed improvements in factuality and reasoning are consistent and not artifacts of decoding variance.

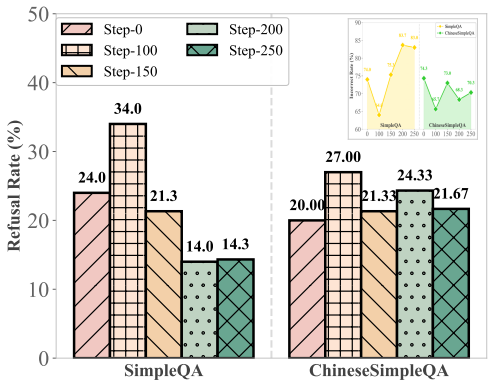


Figure 5: Trends of refusal and incorrect rates on SimpleQA and ChineseSimpleQA when training with the combination of format reward and negative refusal reward across different training steps.

As shown in Table 4, the multi-run averages consistently support our primary findings. Even under stochastic decoding, KnowRL significantly improves refusal rates on hallucination-prone queries (SimpleQA and ChineseSimpleQA) while simultaneously enhancing or maintaining performance on complex reasoning tasks (GPQA and AIME). For instance, KnowRL improves AIME accuracy from 29.33% to 34.00% for DeepSeek-7B, confirming that our framework effectively promotes reliable reasoning rather than simply collapsing to conservative refusals.

6 Related Work

Hallucination Mitigation. Extensive research focuses on mitigating hallucinations in LLMs through alignment strategies and knowledge-grounded methods (Cheng et al., 2025a; Kang et al., 2025; Wen et al., 2025; Xu et al., 2024a; Li and Ng; Lin et al., 2024; Sun et al., 2025; Liang et al., 2024; Ding et al., 2024; Xu et al., 2025; Tang et al., 2024; Han et al., 2024; Xu et al., 2024b; Yin et al., 2025; Yuan et al., 2023; Chen et al., 2024b; Zhang et al., 2024a; Dhuliawala et al., 2024; Feng et al., 2024; Liu et al., 2025a; Martino et al., 2023). This challenge is particularly acute in slow-thinking models, where complex reasoning can amplify factual errors into a snowball effect (Ji et al., 2023; Huang et al., 2025; Zhang et al., 2023; Cheng et al., 2025b;

Methods	Hallucination										Reasoning	
	TruthfulQA		SimpleQA				ChineseSimpleQA				GPQA Diamond	AIME
	Rouge	Bleu	PAQ	Incorrect	Refusal	F1	PAQ	Incorrect	Refusal	F1		
Skywork-OR1-7B-Preview												
Zero-shot	56.20	51.60	7.48	59.47	35.73	9.16	10.95	57.87	35.00	13.30	44.95	34.67
DPO	56.13	51.33	6.65	81.27 (+21.80)	12.93	10.93	11.83	77.47 (+19.60)	12.13	18.81	44.34	34.67
FactTune-FS	57.60	53.00	3.51	44.33 (-15.14)	54.07	2.20	10.84	40.60 (-17.27)	54.47	9.40	45.45	34.67
SFT	57.27	50.93	9.52	61.07 (+1.60)	32.40	10.71	8.21	78.93 (+21.06)	14.00	7.59	39.39	38.67
TruthRL	57.67	53.47	8.15	49.07 (-10.40)	46.60	8.30	12.33	32.60 (-25.27)	62.80	8.78	46.36	33.33
KnowRL	58.27	52.80	7.32	40.20 (-19.27)	**56.60**	6.17	10.30	28.40 (-29.47)	**68.33**	6.32	**48.89** ↑	**38.00** ↑
DeepSeek-R1-Distill-Qwen-7B												
Zero-shot	57.40	51.53	4.89	62.47	34.33	6.20	10.23	60.27	32.93	11.83	45.45	29.33
DPO	55.60	51.40	5.65	76.33 (+13.86)	19.07	8.74	9.54	67.20 (+6.93)	25.73	13.20	40.40	28.67
FactTune-FS	56.93	52.13	2.73	61.73 (-0.74)	36.53	2.12	8.90	56.60 (-3.67)	37.87	10.48	43.74	30.67
SFT	57.13	51.27	9.38	82.47 (+20.00)	9.00	15.72	9.86	82.07 (+21.80)	8.93	9.41	38.28	35.33
TruthRL	56.27	51.94	5.93	50.73 (-11.74)	46.07	6.19	9.37	30.47 (-29.80)	66.40	4.69	44.34	31.33
KnowRL	57.20	52.40	5.50	48.27 (-14.20)	**48.93**	5.43	8.38	28.07 (-32.20)	**69.40**	4.94	**46.97** ↑	**34.00** ↑

Table 4: **Multi-run Evaluation Results.** Robust multi-run performance (Avg@5, Temperature=0.6) of KnowRL compared against baselines on the Skywork-OR1-7B-Preview and DeepSeek-R1-Distill-Qwen-7B models. Zero-shot refers to the original model performance. The key improvements are marked in **bold**.

Yao et al., 2025; Zheng et al., 2025b; Zhang et al., 2025a).

Hallucinations stem from diverse factors, including data noise, RL side-effects, knowledge conflicts, and decoding uncertainties (Mündler et al., 2023; Song et al., 2025b; Ouyang et al., 2022; Zhang et al., 2024b, 2025d; Kuhn et al., 2023; Zhao et al., 2024). Fundamentally, these issues reflect the model's failure to recognize its own knowledge boundaries (Zhang, 2023; Tonmoy et al., 2024; Liang et al., 2024; Manakul et al., 2023; Kadavath et al., 2022), motivating research into knowledge-aware optimization and honesty alignment (Chen et al., 2022; Chen, 2023; Yang et al., 2024).

RL for Reasoning RL enhances LLM reasoning, enabling complex strategies like reflection and verification (Xie et al., 2025; Yeo et al., 2025; Mei et al., 2025; He et al., 2025b). To achieve more reliable reasoning, recent work increasingly focuses on improving RL algorithms (Hu et al., 2025; Nan et al., 2025; Chen et al., 2025c,b; Ichihara and Jinnai, 2025; Ding et al., 2025; Cai et al., 2025; Mroueh et al., 2025; Pang and Jin, 2025; Zhang et al., 2025b; Li et al., 2025b,d; Zheng et al., 2025a; Wang et al., 2025; Li et al., 2025a; Xi et al., 2025) and integrating RL with other components of the training and inference pipeline (Dong et al., 2024; Cen et al., 2024; Xie et al., 2024; Li and Yan, 2025; Li et al., 2025c; Zhu et al., 2024; Chen et al., 2025d; Jin et al., 2025b,a; Song et al., 2025a; Zhao et al., 2025). Fine-grained guidance methods, such as step-level value preference optimization (Chen et al., 2024a), tree search (Feng et al., 2023), and entropy-based preference clarification (Zhu et al., 2025), are proving valuable for enhancing model reliability.

7 Conclusion

This paper studies the high levels of hallucination in both slow-thinking models. We analyze how the current outcome-reward-driven reinforcement learning method, despite enhancing reasoning, fails to ensure fact-based thinking. To address this, we propose KnowRL, a knowledge-enhanced RL training method, and validate its effectiveness on multiple datasets. Our experiments demonstrate that directly supervising the model's thinking process with factual rewards is a more robust strategy than solely optimizing for final answers. This process-oriented supervision is crucial, as it fundamentally teaches models not just what the correct answer is, but how to reason reliably and recognize the boundaries of their own knowledge. We hope KnowRL offers the community an effective technical pathway to mitigate hallucinations in these slow-thinking models.

Limitations

Despite our best efforts, this work still has certain limitations that point to promising directions for future exploration and improvement.

Fundamental Mechanism Studies. Our experiments observe a significant multilingual hybrid reasoning phenomenon, where factual supervision in one language (e.g., English) improves performance in another (e.g., Chinese). While this demonstrates the strong transferability and robustness of KnowRL, the underlying theoretical mechanism, specifically how the model internalizes and transfers these knowledge boundaries across language distributions, warrants deeper investigation. Unraveling this mechanism could provide fundamental insights into the nature of chain-of-thought reasoning in large language models.

Extension to Multimodal Domains. The current KnowRL framework is tailored for textual reasoning and factuality. However, real world information often spans multiple modalities, such as interpreting charts in financial reports or analyzing diagrams in physics problems. Extending the atomic fact verification mechanism to verify information across visual or audio modalities represents a significant opportunity. Future work could explore adapting KnowRL to Vision-Language Models (VLMs), enabling grounded reasoning in broader, multisensory contexts.

Acknowledgement

We would like to express sincere gratitude to the reviewers for their thoughtful and constructive feedback. This work was supported by the National Natural Science Foundation of China (No. 62576307, No. NSFCU23B2055, No. NSFCU19B2027), the Fundamental Research Funds for the Central Universities (226-2023-00138), Yongjiang Talent Introduction Programme (2021A-156-G), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. This work was supported by Ant Group and Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph.

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*.

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, pages 2503–2514.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, et al. 2025. Training-free group relative policy optimization. *arXiv preprint arXiv:2510.08191*.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. 2024. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Step-level value preference optimization for mathematical reasoning. *arXiv preprint arXiv:2406.10858*.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025a. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*.

Huajun Chen. 2023. Large knowledge model: Perspectives and challenges. *arXiv preprint arXiv:2312.02706*.

Peter Chen, Xiaopeng Li, Ziniu Li, Xi Chen, and Tianyi Lin. 2025b. Spectral policy optimization: Coloring your incorrect reasoning in grpo. In *2nd AI for Math Workshop@ ICML 2025*.

Weizhe Chen, Sven Koenig, and Bistra Dilkina. 2025c. Lspo: Length-aware dynamic sampling for policy optimization in llm reasoning. *arXiv preprint arXiv:2510.01459*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788.

Xinxi Chen, Li Wang, Wei Wu, Qi Tang, and Yiyao Liu. 2024b. Honest ai: Fine-tuning" small" language models to say" i don't know", and reducing hallucination in rag. *arXiv preprint arXiv:2410.09699*.

Zhenfang Chen, Delin Chen, Rui Sun, Wenjun Liu, and Chuang Gan. 2025d. Scaling autonomous agents via automatic reward modeling and planning. *arXiv preprint arXiv:2502.12130*.

Ruoxi Cheng, Haoxuan Ma, Weixin Wang, Ranjie Duan, Jiexi Liu, Xiaoshuang Jia, Simeng Qin, Xiaochun Cao, Yang Liu, and Xiaojun Jia. 2025a. Inverse reinforcement learning with dynamic reward scaling for llm alignment. *arXiv preprint arXiv:2503.18991*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025b. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578.

Fei Ding, Baiqiao Wang, Zijian Zeng, and Youwei Wang. 2025. Multi-layer grpo: Enhancing reasoning and self-correction in large language models. *arXiv preprint arXiv:2506.04746*.

Mucong Ding, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, Amrit Bedi, and Furong Huang. 2024. Sail: Self-improving efficient online alignment of large language models. *arXiv preprint arXiv:2406.15567*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

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

Seungwook Han, Idan Shenfeld, Akash Srivastava, Yoon Kim, and Pulkit Agrawal. 2024. Value augmented sampling for language model alignment and personalization. *arXiv preprint arXiv:2405.06639*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024a. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Yang Liu, and Yahui Zhou. 2025a. Skywork open reasoner series. Notion Blog.

Qianxi He, Qingyu Ren, Shanzhe Lei, Xuhong Wang, and Yingchun Wang. 2025b. Beyond correctness: Confidence-aware reward modeling for enhancing large language model reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27215–27231.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. 2024b. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*.

Alex Heyman and Joel Zylberberg. 2025. Reasoning large language model errors arise from hallucinating critical problem features. *arXiv preprint arXiv:2505.12151*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yuki Ichihara and Yuu Jinnai. 2025. Auto-weighted group relative preference optimization for multi-objective text generation tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1134–1147.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O Arik, and Jiawei Han. 2025a. An empirical study on reinforcement learning for reasoning-search interleaved llm agents. *arXiv preprint arXiv:2505.15117*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025b. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2025. Unfamiliar finetuning examples control how language models hallucinate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3600–3612.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. 2025a. Disco: Reinforcing large reasoning models with discriminative constrained optimization. *arXiv preprint arXiv:2505.12366*.

Gen Li and Yuling Yan. 2025. Towards efficient online exploration for reinforcement learning with human feedback. *arXiv preprint arXiv:2509.22633*.

Junyi Li and Hwee Tou Ng. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. 2025b. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*.

Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. 2025c. Provably efficient online rlhf with one-pass reward modeling. *arXiv preprint arXiv:2502.07193*.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. 2025d. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *arXiv preprint arXiv:2401.15449*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Xukai Liu, Ye Liu, Shiwen Wu, Yanghai Zhang, Yihao Yuan, Kai Zhang, and Qi Liu. 2025a. Know3-rag: A knowledge-aware rag framework with adaptive retrieval, generation, and filtering. *arXiv preprint arXiv:2505.12662*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference*, pages 182–185. Springer.

Jianbiao Mei, Tao Hu, Daocheng Fu, Licheng Wen, Xuemeng Yang, Rong Wu, Pinlong Cai, Xing Gao, Yu Yang, Chengjun Xie, et al. 2025. o^2-searcher: A searching-based agent model for open-domain open-ended question answering. *arXiv preprint arXiv:2505.16582*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jerret Ross, and Jesus Rios. 2025. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. *arXiv preprint arXiv:2505.22257*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, et al. 2025. Ngrpo: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*.

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

Lei Pang and Ruinan Jin. 2025. On the theory and practice of grpo: A trajectory-corrected approach with fast convergence. *arXiv preprint arXiv:2508.02833*.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *arXiv preprint arXiv:2406.17169*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025a. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Linxin Song, Taiwei Shi, and Jieyu Zhao. 2025b. The hallucination tax of reinforcement finetuning. *arXiv preprint arXiv:2505.13988*.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. 2025. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. *arXiv preprint arXiv:2505.12886*.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. 2024. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. 2025. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

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

Xueru Wen, Jie Lou, Xinyu Lu, Yuqiu Ji, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, Debing Zhang, et al. 2025. On-policy self-alignment with fine-grained knowledge feedback for hallucination mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5215–5231.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, et al. 2025. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. 2025. Bnpo: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. 2024. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Erhan Xu, Kai Ye, Hongyi Zhou, Luhan Zhu, Francesco Quinzan, and Chengchun Shi. 2025. Doubly robust alignment for large language models. *arXiv preprint arXiv:2506.01183*.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *arXiv preprint arXiv:2403.18349*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024b. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*.

Chenlong Yin, Zeyang Sha, Shiwen Cui, and Changhua Meng. 2025. The reasoning trap: How enhancing llm reasoning amplifies tool hallucination. *arXiv preprint arXiv:2510.22977*.

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

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Chen Zhang. 2023. User-controlled knowledge fusion in large language models: Balancing creativity and hallucination. *arXiv preprint arXiv:2307.16139*.

Mike Zhang, Johannes Bjerva, and Russa Biswas. 2025a. Scaling reasoning can improve factuality in large language models. *arXiv preprint arXiv:2505.11140*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024a. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025b. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025c. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R Fung, Jing Li, Manling Li, and Heng Ji. 2024b. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, et al. 2025d. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. *arXiv preprint arXiv:2502.16143*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, et al. 2025. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*.

Suifeng Zhao, Tong Zhou, Zhuoran Jin, Hongbang Yuan, Yubo Chen, Kang Liu, and Sujian Li. 2024. Awecita: Generating answer with appropriate and well-grained citations using llms. *Data Intelligence*, 6(4):1134–1157.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. 2025a. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Hang Zheng, Hongshen Xu, Yuncong Liu, Lu Chen, Pascale Fung, and Kai Yu. 2025b. Enhancing llm reliability via explicit knowledge boundary modeling. *arXiv preprint arXiv:2503.02233*.

Banghua Zhu, Michael I Jordan, and Jiantao Jiao. 2024. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. *arXiv preprint arXiv:2401.16335*.

Nana Zhu, Zixian Feng, Hang Wang, Xing Gao, Xinyi Wang, and Yuanxing Liu. 2025. Clarifying user preference with maximum entropy based recommendation. *Data Intelligence*, 7(2):527–548.

A Data Construction

We use part of the data extracted from NqOpen (Kwiatkowski et al., 2019; Lee et al., 2019), as well as data from WebQuestions (Berant et al., 2013) and ComplexQuestions (Bao et al., 2016), as factual question data sources.

Initially, we filtered out trivial queries and applied semantic de-duplication to ensure data diversity. The remaining data underwent quality refinement and entity extraction via GPT-4o (see Appendix I for detailed prompts). To address the limitation of incomplete knowledge coverage in existing datasets (e.g., HotpotQA), we explicitly matched the extracted entities against the Wikipedia dump[3] to retrieve comprehensive factual contexts for each question. Finally, we applied length constraints to screen the samples, resulting in a robust dataset for stable training.

B Training Setups

We are training two 7B models, DeepSeek-R1-Distill-Qwen-7B and Skywork-OR1-7B-Preview, with Reinforcement Learning on 1×A800. Detailed training hyperparameters are listed in Table 5.

Parameter	Value
lora_rank	128
lora_alpha	256
torch_dtype	bfloat16
per_device_train_batch_size	20
gradient_accumulation_steps	4
learning_rate	1.0e-5
beta	0.001
lr_scheduler_type	cosine
warmup_ratio	0.03
vllm_gpu_memory_utilization	0.5
optim	adamw_8bit

Table 5: Hyperparameter settings for RL training. Both DeepSeek-R1-Distill-Qwen-7B and Skywork-OR1-7B-Preview share these identical configurations.

C Case Analysis

To explore KnowRL's impact on reasoning in factual tasks, we analyzed a reasoning case from the KnowRL training process, detailing the model's emergent reasoning behaviors. As shown in Figure 6, these include Mixed-Language Reasoning, Knowledge Anchoring, Cross-verification, and Key Information Extraction. For simple factual tasks, distinct from mathematical reasoning, the model typically first proposes an initial answer (knowledge anchoring) and subsequently verifies this initial answer through reasoning behaviors such as reflection and cross-verification. This observed process aligns with human cognitive approaches when facing factual tasks, which further suggests the suitability of an outcome-based, reward-driven reinforcement learning training paradigm for open-domain factual tasks.

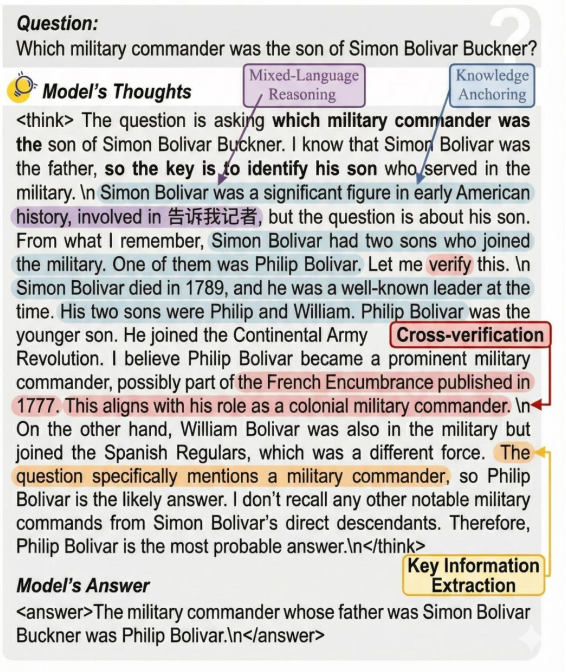
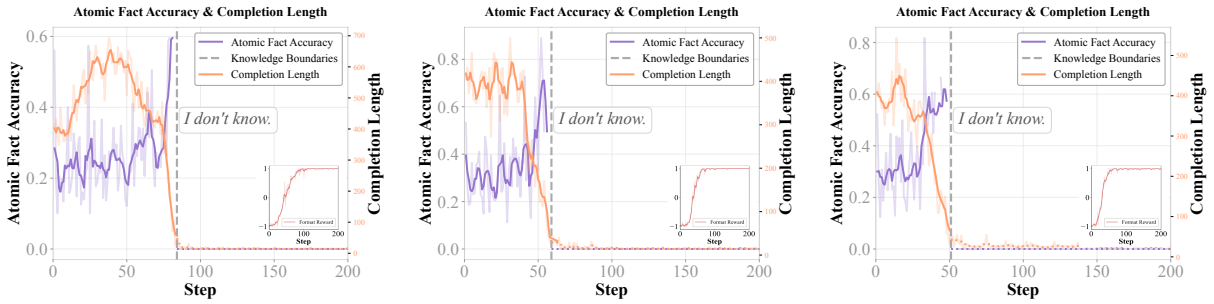


Figure 6: Case analysis of the KnowRL training process.

D Training Dynamics of Different RL Algorithms

As shown in Figure 8, the reward curves for all four algorithms rise and stabilize, indicating successful convergence. Figure 7 illustrates the training dynamics of the DeepSeek-R1-Distill-Qwen-7B model using BNPO, DAPO, and Dr.GRPO. A key observation is that all three algorithms effectively help the model establish clear knowledge boundaries to avoid hallucination. This effectiveness is evidenced by the rapid decline in completion length across all methods. The sharp drop indicates a behavioral shift where the model learns to provide concise refusals instead of generating long and potentially fabricated responses. This consistent pattern confirms that these RL approaches successfully guide the model toward reliable behavior and mitigate uncontrolled generation. Despite these shared trends, the algorithms differ in training efficiency and their impact on factual reasoning.

[3] 20231101.en, https://dumps.wikimedia.org/.



(a) **BNPO** (b) **DAPO** (c) **Dr.GRPO**

Figure 7: **Training Dynamics on DeepSeek-R1-Distill-Qwen-7B.** We visualize the training curves of different RL algorithms: (a) BNPO, (b) DAPO, and (c) Dr.GRPO.

E Scalability Analysis on Larger Models

We further evaluate the scalability of KnowRL by extending our experiments to the DeepSeek-R1-Distill-Qwen-14B model. As illustrated in Table 6, KnowRL maintains its efficacy in mitigating hallucinations while simultaneously enhancing reasoning capabilities. Notably, on SimpleQA, our method significantly reduces the Incorrect Rate from 83.00% to 68.33% while doubling the Refusal Rate (13.33% to 26.33%). This shift indicates that the model acquires a more precise awareness of knowledge boundaries on larger architectures. Furthermore, KnowRL preserves and even strengthens complex reasoning performance, evidenced by the improvement in GPQA Diamond accuracy from 46.97% to 51.01%. These results validate that the benefits of our approach are robust and scalable across different model sizes.

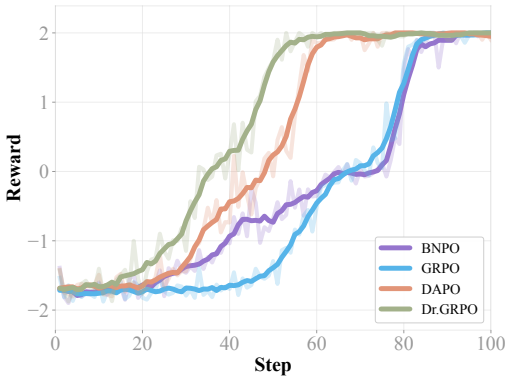


Figure 8: Comparison of training reward curves across different algorithms.

Metric	Zero-shot	KnowRL
TruthfulQA		
Rouge	53.33	54.67
Bleu	50.33	55.00
SimpleQA		
PAQ	4.23	6.17
Incorrect	83.00	68.33
Refusal	13.33	26.33
F1	3.93	6.14
ChineseSimpleQA		
PAQ	29.18	31.07
Incorrect	63.33	61.33
Refusal	6.33	11.00
F1	28.23	29.28
GPQA Diamond		
Accuracy	46.97	51.01
AIME		
Accuracy	40.00	36.67

Table 6: Performance comparison on DeepSeek-R1-Distill-Qwen-14B. The results demonstrate that KnowRL consistently improves hallucination mitigation (SimpleQA) and reasoning (GPQA) compared to the Zero-shot baseline.

F Evaluator Sensitivity Analysis

Metric	Qwen2.5-72B-Instruct	GPT-4o-mini
TruthfulQA		
Rouge	57.00	57.33
Bleu	53.67	51.60
SimpleQA		
PAQ	1.53	2.81
Incorrect	64.33	57.67
Refusal	34.67	40.67
F1	1.21	2.09
ChineseSimpleQA		
PAQ	9.74	10.26
Incorrect	58.67	58.33
Refusal	35.00	35.00
F1	7.68	8.08
GPQA Diamond		
Accuracy	38.38	36.87
AIME		
Accuracy	33.33	33.33

Table 7: Robustness analysis of KnowRL using different evaluator models during training.

To confirm that the effectiveness of KnowRL is not dependent on a specific reward model used during training (e.g., GPT-4o-mini), we conducted an ablation study by replacing it with Qwen2.5-72B-Instruct (Team, 2024). As presented in Table 7, the results demonstrate that KnowRL maintains comparable performance across different evaluators. Specifically, we observe that the model trained with GPT-4o-mini exhibits more conservative behavior, achieving a higher Refusal Rate and a lower Incorrect Rate on SimpleQA. In contrast, using Qwen2.5-72B-Instruct yields slightly higher performance on reasoning-heavy benchmarks like GPQA Diamond, suggesting a subtle trade-off between strict factuality enforcement and reasoning preservation depending on the evaluator's characteristics. This confirms that the efficacy of our method stems from the intrinsic KnowRL framework rather than reliance on a specific external judge.

G Evaluation of Generative Diversity

To investigate whether KnowRL's boundary-aware training inadvertently leads to over-conservatism or a loss of generative diversity, we evaluate our models using NoveltyBench (Zhang et al., 2025c). This benchmark measures the ability of LLMs to produce diverse outputs for the same prompt. We conduct evaluations on the *nb-curated* (100 prompts) and *nb-wildchat* (1,000 prompts) subsets, sampling 10 responses per prompt at a temperature of 0.6. We report the *distinct* metric, which averages the number of semantically unique equivalence classes across generations.

Model	nb-curated	nb-wildchat
DeepSeek-7B	1.61	1.71
DeepSeek-7B + KnowRL	1.54	1.74
Skywork-7B	1.63	1.65
Skywork-7B + KnowRL	1.68	1.64

Table 8: Generative diversity (*distinct* scores) on NoveltyBench subsets.

As shown in Table 8, the variation in *distinct* scores after KnowRL training is extremely marginal (± 0.01 to ± 0.07), indicating that the model's generative diversity remains intact. We attribute this preservation to the use of Low-Rank Adaptation (LoRA), which allows the model to learn new boundary-aware behaviors while mitigating the catastrophic forgetting of its original generative capabilities. While KnowRL increases refusal rates for out-of-knowledge queries to reduce hallucinations, it does not compromise the model's creative breadth in open-ended scenarios. We note that establishing an absolute "False Refusal Rate" remains challenging due to the difficulty of probing a model's precise internal knowledge, which we leave for future investigation.

H Implementation Details for Baselines

To ensure fair comparison and reproducibility, all trainable baselines are fine-tuned using Low-Rank Adaptation (LoRA) and utilize the exact same QA training dataset as KnowRL. The specific implementation details for each baseline are described as follows:

- **Self-Refine (Prompt Engineering):** The model generates an initial response and then performs self-critique to provide feedback. It uses this self-feedback to refine its output iteratively. This "FEEDBACK $\rightarrow$ REFINE" loop is repeated until a stopping condition is met, specifically when it reaches a maximum of 5 iterations or when the model determines that no further improvement is needed.

- **SFT (Supervised Fine-Tuning):** We train the model using the correct reasoning processes and final answers distilled from the DeepSeek-R1 model as the target outputs.

- **DPO (Direct Preference Optimization):** We construct the preference pairs by using the correct answer distilled from DeepSeek-R1 as the "chosen" response, and the model's own incorrect generation as the "rejected" response.

- **FactTune-FS:** Following the methodology of the original paper (**?**), we sample multiple responses from the model for a given prompt and evaluate them using FactScorer. We then construct DPO training pairs by selecting two responses that have a FactScore difference greater than 0.8.

- **TruthRL:** We adopt the exact same GRPO training setup as KnowRL to ensure a strictly fair comparison. We only change the reward function to match TruthRL's original design: $+1$ for a correct answer, 0 for a refusal, and -1 for an incorrect answer.

I Prompts

Prompt Used by the GPT-4o for Data Filtering

You are an entity extraction assistant that identifies key entities in questions.
TASK:
1. First normalize the query by properly capitalizing names, titles, and other named entities
2. Determine if the query has sufficient context to be answered meaningfully
3. Extract only the most important entities from the query that are essential for answering it
RULES:
1. Extract a MAXIMUM of 2 specific entities (people, places, objects, works, etc.)
2. Output the MOST important entity first, then the secondary entity (if any)
3. Extract precise named entities, not general concepts or phrases
4. Keep related entities together as a single entity (e.g., character names with their roles)
5. Return individual entities rather than relationships or possessive forms
6. Only extract truly representative entities - ignore generic terms that don't specifically define the query
7. Only REJECT queries that meet the rejection criteria below
ONLY reject queries in these specific cases:
1. When the entity in the query is completely ambiguous (e.g., "Who is that person?")
2. When the query lacks necessary qualifying information (e.g., "Who will win?" with no mention of what contest)
3. When the query is too vague to determine its intent (e.g., "What happened to him?")
4. When the query is time-sensitive and contains temporal references like "now", "current", "latest", "recent", etc.
5. When the query lacks sufficient information to determine a single definitive answer, potentially leading to multiple correct interpretations or answers
6. Be careful not to extract purely numerical information such as a year as an entity

Note: Queries with historical context, pop culture references, geographical locations, or other well-defined entities should be ACCEPTED.
EXAMPLES:
Example 1:
Original Query: "who played barbara gordon batgirl?"
Normalized Query: "Who played Barbara Gordon Batgirl?"
Output: Normalized Query: "Who played Barbara Gordon Batgirl?"
Entities: ["Barbara Gordon Batgirl"]

NOT: ["Barbara Gordon", "Batgirl"] - This is incorrect because "Barbara Gordon Batgirl" is a single character entity.
Example 2:
Original Query: "what continent does armenia belong to?"
Normalized Query: "What continent does Armenia belong to?"
Output: Normalized Query: "What continent does Armenia belong to?"
Entities: ["Armenia"]

NOT: ["Armenia", "continent"] - The term "continent" is a generic category, not a specific entity representative of this query.
Example 3:
Original Query: "who is niall ferguson's wife?"
Normalized Query: "Who is Niall Ferguson's wife?"
Output: Normalized Query: "Who is Niall Ferguson's wife?"
Entities: ["Niall Ferguson"]

Example 4:
Original Query: "who was the italian leader in ww1?"
Normalized Query: "Who was the Italian leader in WW1?"
Output: Normalized Query: "Who was the Italian leader in WW1?"
Entities: ["Italian leader", "WW1"]
Example 5:
Original Query: "who will play mr gray in the film?"
Normalized Query: "Who will play Mr.

Gray in the film?"
Output: Normalized Query: "Who will play Mr. Gray in the film?"
REJECT (insufficient context - which film?)
Example 6:
Original Query: "who is in charge of libya now?"
Normalized Query: "Who is in charge of Libya now?"
Output: Normalized Query: "Who is in charge of Libya now?"
REJECT (time-sensitive query with temporal reference "now")

Example 7:
Original Query: "what did werner heisenberg discover?"
Normalized Query: "What did Werner Heisenberg discover?"
Output: Normalized Query: "What did Werner Heisenberg discover?"
REJECT (lacks sufficient specificity - Heisenberg made multiple discoveries)
Please try to output in this format:
Normalized Query: "The normalized version of the query"
Entities: ["entity1", "entity2"]
If you need to reject, still include the normalized query:
Normalized Query: "The normalized version of the query"
REJECT (reason for rejection)
Extract key entities from this query: "query"