Week 9 Report

Mu Junrong

Abstract

This report investigates the methodologies and findings presented in the paper DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (Team (2024)) and Understanding R1-Zero-Like Training: A Critical Perspective (Liu et al. (2025)), which proposes a reinforcement learning (RL)-based approach to enhance reasoning capabilities in large language models (LLMs). The report also examines the models proposed in the papers, including the DeepSeek-R1-Zero model and DeepSeek-R1 model. Some key algorithms, including Group Relative Policy Optimisation (GRPO) and Debiased GRPO are also discussed in the paper. Overall, DeepSeek-R1 achieves better performance comparable to proprietary models like OpenAI's o1-1217, especially for reasoning ability.

1 Summary of the papers

This report focuses on two closely related papers, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning" and "Understanding R1-Zero-Like Training: A Critical Perspective". The first paper proposes an approach to improve large language models' (LLMs) reasoning ability through reinforcement learning (RL), by introducing two core models(Team (2024)).The second paper critically evaluates this approach, while offering refinements into its underlying mechanisms, through the newly proposed algorithm (Liu et al. (2025)).

1.1 DeepSeek-R1

The DeepSeek-R1 paper introduces two core models, DeepSeek-R1-Zero and DeepSeek-R1, respectively. The model DeepSeek-R1-Zero is trained purely through RL from a pretrained base (DeepSeek-V3-Base) without any supervised fine-tuning (SFT). On the other hand, the model DeepSeek-R1 is a more robust model that incorporates a small set of curated Chain-of-Thought (CoT) examples as a cold start before undergoing RL, followed by reasoning-focused RL, rejection sampling, and a final RL stage across general prompts.

The paper also introduces the algorithm Group Relative Policy Optimization (GRPO), which estimates relative advantages across multiple sampled responses, enabling stable and efficient RL without the need for a value network. A reward system composed of accuracy-based and format-based signals drives the models to generate verifiable, structured reasoning steps.

1.2 Understanding R1-Zero-Like Training

The paper challenges the assumptions and mechanisms behind DeepSeek-R1-Zero. The paper argues that many of the observed reasoning behaviors may not be purely induced by RL, but instead emerge from biases in the pretrained base model. Through a series of diagnostic experiments, they show that "Aha-moments" and long Chain of Thought (CoT) can already be elicited from base models like Qwen2.5 when prompted appropriately ()Wei et al. (2022).

They also challenge GRPO for its optimization bias, particularly its tendency to reward longer outputs regardless of accuracy—resulting in inefficiency and verbosity. To address this, they propose Dr. GRPO (Debiased GRPO), which removes length normalization and improves token efficiency.

2 DeepSeek-R1-Zero

DeepSeek-R1-Zero is a large language model (LLM) that learns to perform complex reasoning through pure reinforcement learning (RL), without any supervised fine-tuning (SFT). It is trained from a pretrained base model, DeepSeek-V3-Base, using the reinforcement learning (RL) algorithm named Group Relative Policy Optimization (GRPO) during the fine-tuning stage. This work is significant because it demonstrates that a large model can acquire advanced reasoning abilities (e.g., reflection, self-verification, long chain-of-thought) solely from reinforcement learning guided by rule-based rewards.

2.1 Model architecture and training pipeline

The Deep-R1-Zero model starts from the base model DeepSeek-V3-Base (Team (2023)), which is a dense transformer with around 37 billion activated parameters per forward pass (i.e., only 37 billion parameters are used when generating a token since only 37 billion are activated (used) during each forward pass) and total model size is around 671 billion parameters (due to Mixture-of-Experts architecture). The training objective is to modify the behavior of this pretrained model through RL to prefer reasoning-style answers. The outputs will follow the format of

```
<think> reasoning steps </think>
<answer> final answer </answer>
```

The training pipeline is as follows:

- Input Prompt: A zero-shot prompt for a reasoning task (e.g., math question).
- Multiple Completions: The model samples multiple outputs per input prompt.
- Reward Evaluation:
  - Correctness-based reward
  - Format-based reward: `<think> and <answer>`
  - Optional language consistency reward
- Group Advantage Estimation: The completions are compared within a group to compute relative advantages.
- Policy Update: Use GRPO to improve the model towards higher-reward completions.

2.2 Algorithm: Group Relative Policy Optimization (GRPO)

The reward function is deterministic and rule-based, and no learned reward model is used. Key reward components include:

- Accuracy Reward: for math tasks, the boxed final answer (i.e., the output) is compared against ground truth.
- Format Reward: encourages structured responses where full credit is given if it follows the format `<think> and <answer>` and penalize otherwise.

The algorithm Group Relative Policy Optimization (GRPO) is a variant of Proximal Policy Optimization (PPO) designed to avoid training a critic or value function and Use relative ranking of outputs instead of absolute value estimation (Schulman et al. (2017)). Thus it Works efficiently with grouped responses ()Team (2024).

Given a prompt q, the model samples a group of responses $\{o_1, o_2, \ldots, o_G\}$ with corresponding rewards $\{r_1, \ldots, r_G\}$.

The normalized advantage for each sample i is calculated by

$$A_i = \frac{r_i - \mu_r}{\sigma_r}$$

where$\mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j$ (i.e., the mean of all rewards) and $\sigma_r = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_r)^2 + \epsilon}$ (i.e., the standard deviation of all rewards)

The objective function of GRPO is

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}A_i, \text{clip}(\cdot)\right) - \beta D_{KL}(\pi_\theta||\pi_{\text{ref}})\right]$$

Where:

- $\pi_\theta(o_i|q)$: the new policy's probability for response o_i
- π_{old}: old policy (used for importance sampling)
- D_{KL}: KL-divergence penalty between new policy and reference model (stabilization)
- β: KL coefficient (regularization strength)

The reward function ensures that higher-reward completions are more likely and high deviation from the base model is penalized.

For example, the prompt is
is the sum of 123 and 456? and the outputs are

```
        o1: <think>123 + 456 = 579</think><answer>579</answer>
        o2: <think>123 + 456 = 589</think><answer>589</answer>
o3: <think>Let's estimate: 120 + 450 is 570</think><answer>570</answer>
o4: <think>123 + 456 is between 500 and 600</think><answer>540</answer>
```

The reward calculated by the reward function will give $R(o1) = 1.0, R(o2) = 0.5, R(o3) = 0.4, R(o4) = 0.3$, which is based on correct final answer and output format. Next the normalize rewards are computed, and the model is updated using GRPO.

3 DeepSeek-R1

DeepSeek-R1 is a Reinforcement Learning-enhanced LLM built on top of DeepSeek-R1-Zero with several key improvements. It includes a small cold-start supervised fine-tuning (SFT) stage and a more robust RL training pipeline. It also rejects sampling to ensure higher-quality training data and includes a final RL stage across general-purpose prompts.

It not only performs well on math and logic tasks but also generates cleaner, more readable Chain-of-Thought (CoT) reasoning. It outperforms many proprietary LLMs (e.g., OpenAI o1-1217) on several benchmarks.

The pipeline of DeepSeek-R1 is

Pretrained Model (DeepSeek-V3-Base) $\rightarrow$ Cold-Start SFT (Small curated CoT) $\rightarrow$ RL with GRPO (Reasoning-focused) $\rightarrow$ Rejection Sampling $\rightarrow$ Final RL (General prompts + reward model)

while the pipeline of DeepSeek-R1-Zero is

Pretrained Model (DeepSeek-V3-Base) $\rightarrow$ RL with GRPO

3.1 Model architecture

The DeepSeek-R1 training process consists of four main stages: Cold-Start supervised fine-tuning (SFT), Reasoning-Focused RL by GRPO, rejection sampling and supervised fine-tuning (SFT), and RL on general prompts.

During Cold-Start SFT, the base model is fine-tuned on a small dataset (around thousands of samples). The dataset consists of human-written CoT-style answers (long, markdown-like)

and questions from math, logic, and reasoning tasks. The goal is to give the model a clean and structured reasoning format.

4 Dr. GRPO and Minimal R1-Zero

Dr. GRPO (Debiased GRPO) is a modified reinforcement learning algorithm designed to fix specific biases in GRPO:

- strong preference for longer outputs, regardless of correctness.
- over-sensitivity to variance in rewards, which can destabilize training.

4.1 The improved algorithm: Dr. GRPO

In GRPO, the normalized advantage for each sample i is calculated by

$$A_i = \frac{r_i - \mu_r}{\sigma_r}$$

. However, if σ_r is small (i.e. all rewards are similar), the advantages become exaggerated. The reward function of GRPO has length bias, where longer completions tend to get higher rewards ()Liu et al. (2025).

Dr. GRPO keeps the core idea of ranking outputs within a group, like GRPO, but removes the parts that introduce undesirable optimization bias.

In Dr. GRPO, it has the following improvements:

- The normalized advantage for each sample i is calculated by
$$A_i = r_i - \mu_r \quad \text{(Dr. GRPO)}$$
- The optional length-normalized reward is given by
$$r_i = \frac{r_{\text{raw}}(o_i)}{\text{len}(o_i)}$$
This prevents the model from being rewarded just for outputting more tokens.

For example, the prompt is
is the sum of 123 and 456? and the outputs are

```
o1: <think>123 + 456 = 579</think><answer>579</answer> (Correct)
o2: <think>123 + 456 = 589</think><answer>589</answer> (Wrong)
o3: <think>Estimate: 120 + 460 = 580</think><answer>580</answer> (Wrong)
o4: <think>This is a trick question, let me think
again...</think><answer>579</answer> (Correct but long)
```

If length-normalized rewards is used, $o1$ would have higher advantage than $o4$, thus discouraging verbosity.

5 Conclusion

This report examined the training methodologies, algorithms, and model behaviors introduced in DeepSeek-R1 and Understanding R1-Zero-Like Training. Through detailed analysis of the DeepSeek-R1-Zero and DeepSeek-R1 models, we observed that reinforcement learning, particularly the Group Relative Policy Optimization (GRPO) algorithm can effectively elicit complex reasoning behavior in large language models (LLMs), even in the absence of supervised fine-tuning (SFT). The staged pipeline in DeepSeek-R1, combining a cold-start SFT, structured reward shaping, and multi-domain RL, demonstrated state-of-the-art performance across multiple reasoning benchmarks, such as AIME, MATH-500, and GPQA.

However, there are limitations in the original formulation, notably GRPO's bias toward verbosity and its sensitivity to reward variance. To address this, the Dr. GRPO algorithms is designed to remove standard deviation normalization and optionally normalizes rewards by output length.

Together, these findings demonstrate the viability of reinforcement learning as a standalone training signal for reasoning in LLMs. At the same time, they caution against over-attributing emergent behaviors to reinforcement signals alone, as many such capabilities may already reside in pretrained models and be sensitive to prompting or optimization artifacts. Going forward, integrating principled RL algorithms like Dr. GRPO, efficient distillation pipelines, and well-calibrated reward functions offers a promising path toward accessible and interpretable reasoning-capable LLMs.

References

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

DeepSeek Team. Deepseek-v3: Towards scalable and capable multilingual language models. *arXiv preprint arXiv:2312.10301*, 2023. URL https://arxiv.org/abs/2312.10301.

DeepSeek Team. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2024. URL https://arxiv.org/abs/2501.12948.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.11903.