Understanding R1-Zero-Like Training: A Critical Perspective

Zichen Liu[*†1,2], **Changyu Chen**[*1,3], **Wenjun Li**[*3], **Penghui Qi**[*1,2],
Tianyu Pang[1], **Chao Du**[1], **Wee Sun Lee**[2], **Min Lin**[1]
[1]Sea AI Lab
[2]National University of Singapore
[3]Singapore Management University

Abstract

DeepSeek-R1-Zero has shown that reinforcement learning (RL) at scale can directly enhance the reasoning capabilities of LLMs without supervised fine-tuning. In this work, we critically examine R1-Zero-like training by analyzing its two core components: *base models* and *RL*. We investigate a wide range of base models, including DeepSeek-V3-Base, to understand how pretraining characteristics influence RL performance. Our analysis reveals that **DeepSeek-V3-Base already exhibit "Aha moment"**, while **Qwen2.5 base models demonstrate strong reasoning capabilities even without prompt templates**, suggesting potential pretraining biases. Additionally, we identify an optimization bias in Group Relative Policy Optimization (GRPO), which artificially increases response length (especially for incorrect outputs) during training. To address this, we introduce Dr. GRPO, an unbiased optimization method that improves token efficiency while maintaining reasoning performance. Leveraging these insights, we present a minimalist R1-Zero recipe that achieves 43.3% accuracy on AIME 2024 with a 7B base model, establishing a new state-of-the-art.

`https://github.com/sail-sg/understand-r1-zero`[1]

GRPO

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{o}_i|}\sum_{t=1}^{|\mathbf{o}_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right]\right\},$$

$$\text{where } \hat{A}_{i,t}=\frac{R(\mathbf{q},\mathbf{o}_i)-\text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})}{\text{std}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})}.$$

Dr. GRPO

GRPO Done Right (without bias)

$$\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|\mathbf{o}_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right]\right\},$$

$$\text{where } \hat{A}_{i,t}=R(\mathbf{q},\mathbf{o}_i)-\text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\}).$$

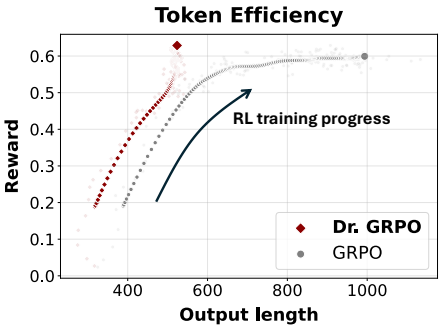

Figure 1: **Left**: Dr. GRPO introduces simple yet significant modifications to address the biases in GRPO (Shao et al., 2024), by removing the length and std normalization terms. **Right**: Our unbiased optimizer effectively prevents the model from generating progressively longer incorrect responses, thereby enhancing token efficiency.

1 Introduction

DeepSeek-R1-Zero (Guo et al., 2025) revolutionizes the pipeline of large language model (LLM) post-training by introducing the *R1-Zero-like training paradigm*: directly applying RL to base LLMs without relying on supervised fine-tuning (SFT) as a preliminary step. This new paradigm is appealing due to its simplicity and the demonstrated **RL scaling phenomenon**: the model reasoning capabilities improve along with a continual increase in

*Core Contributors.
†Project Lead.
[1]Developed with the LLM RL framework Oat: `https://github.com/sail-sg/oat`.

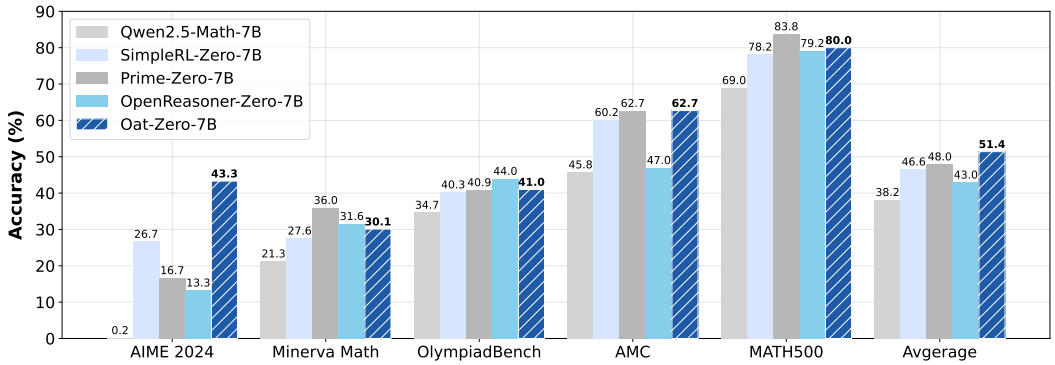

Figure 2: Model performance comparison. **Oat-Zero-7B** is RL-tuned with our minimalist recipe described in Sec. 1 (third paragraph). Please see Sec. B for more results.

model's response length. This phenomenon is also accompanied by the "Aha moment", at which the model learns emergent skills such as self-reflections.

In this paper, we aim to understand R1-Zero-like training by studying two essential components: *base models* and *RL*. In the first part, we investigate various attributes of base models, with the focus on the **Qwen2.5** model family (Yang et al., 2024a;b), which has been used in recent attempts to reproduce R1-Zero (Pan et al., 2025; Zeng et al., 2025; Liu et al., 2025b; Hu et al., 2025), as well as **DeepSeek-V3-Base** (Liu et al., 2024), from which the real R1-Zero model was RL-tuned. In the second part, we identify the **bias in optimization of GRPO** (Shao et al., 2024), which may lead to progressively longer *incorrect* responses. To this end, we propose a simple modification to eliminate the bias, i.e., to get GRPO Done Right (Dr. GRPO), which leads to **better token efficiency** (highlighted in Fig. 1).

Our analysis on base models and RL suggests a **minimalist recipe** for R1-Zero-like training: we RL-tune Qwen2.5-Math-7B using the (unbiased) Dr. GRPO algorithm on MATH (Hendrycks et al., 2021) level 3-5 questions with the Qwen-Math template, and achieve state-of-the-art performance (Fig. 2) with only 27 hours compute on 8× A100 GPUs. We hope our findings presented in this paper, models released, and the codebase open-sourced could benefit future research in the field. As an overview, we summarize the takeaways of this paper below:

Overview of takeaways

- (Sec. 2.1) Template is crucial to make base models **answer questions** instead of completing sentences. In addition, all base models already possess math-solving capability prior to RL.
- (Sec. 2.2) Intriguingly, Qwen-2.5 base models get an **immediate** $\sim$ 60% **improvement by not using template**, making us hypothesize that they may pretrain on concatenated question-answer texts when cooking the models.
- (Sec. 2.3) Nearly all base models already exhibit the "Aha moment", **including DeepSeek-V3-Base**.
- (Sec. 3.1, Sec. 3.2) Dr. GRPO effectively fixes GRPO's bias in optimization, achieving **better token efficiency**.
- (Sec. 3.3) Model-template **mismatch** can destroy reasoning capabilities before RL reconstructs it.
- (Sec. 3.4) **Math pretraining on Llama-3.2-3B** improves its RL ceiling.

2 Analysis on Base Models

In this section, we scrutinize a wide range of base models, including the Qwen-2.5 family (Yang et al., 2024a;b), Llama-3.1 (Grattafiori et al., 2024) and DeepSeek series (Liu et al., 2024; Shao et al., 2024; Guo et al., 2025), asking them 500 questions sampled from the MATH (Hendrycks et al., 2021) training set and analyzing their responses.

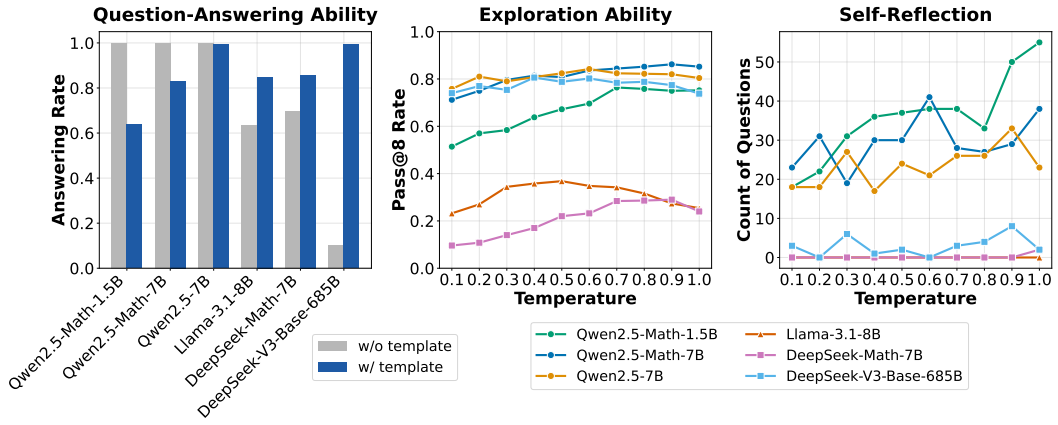

Figure 3: Model attributes across three aspects. **Question-Answering Ability**: the extent to which a pretrained language model provides a direct answer to a question rather than continuing or expanding upon it; **Exploration Ability**: pass@8 measures how well base models explore; **Self-Reflection**: counts are obtained through cross-validation between keyword-based detection and LLM-based detection, as detailed in Appendix D.

2.1 R1-Zero Trainability: Templates Construct Exploratory Base Policies

Since training from a base model is a fundamental setting of the R1-Zero-like paradigm, we first investigate whether widely used open-source base models, which are typically trained for sentence completion (i.e., $p_\theta(\mathbf{x})$), can have their question-answering capabilities effectively elicited through appropriate templates, thereby functioning as a question-answering base policy $\pi_\theta(\cdot|\mathbf{q})$. In addition to the *R1 template* (Template 1) in Guo et al. (2025), we consider the *Qwen-Math template* (Template 2) used by Zeng et al. (2025), as well as *No template* (Template 3):

Template 1 (***R1 template***). A conversation between User and Assistant. The User asks a question, and the Assistant solves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process is enclosed within <think> </think> and answer is enclosed within <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>.\nUser: {question}\nAssistant: <think>

Template 2 (***Qwen-Math template***). <|im_start|>system\nPlease reason step by step, and put your final answer within \\boxed{}.<|im_end|>\n<|im_start |>user\n{question} <|im_end|>\n<|im_start|>assistant\n

Template 3 (***No template***). {question}

Experimental settings. We include Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, Qwen2.5-7B, Llama-3.1-8B, DeepSeek-Math-7B and DeepSeek-V3-Base-685B for experiments. For each model, we first apply ***No template*** to get the model responses, then let GPT-4o-mini to judge whether the model responses are in an answering format (regardless of quality) or in a sentence-completion pattern. We record the percentage of responses that tend to answer the question as the metric. We then apply both ***R1 template*** and ***Qwen-Math template*** to obtain model responses, and determine the most suitable template for each model based on the metric. Finally, we evaluate the pass@8 accuracy of each model with the corresponding template to assess whether the base policies can explore rewarding trajectories for RL improvement.

Results. The left plot of Fig. 3 shows how well base models (with or without templates) answer the provided questions. We observe that Llama and DeepSeek models all improve the answering ability by employing the proper template (R1 template). However, Qwen2.5 models work best (with 100% answering rate) when no template is used. This intriguing property motivates further investigation which will be discussed in Sec. 2.2. Meanwhile, the lowest answering rate with no template suggests that DeepSeek-V3-Base is a nearly pure base model. This observation motivates us to explore whether a pure base model