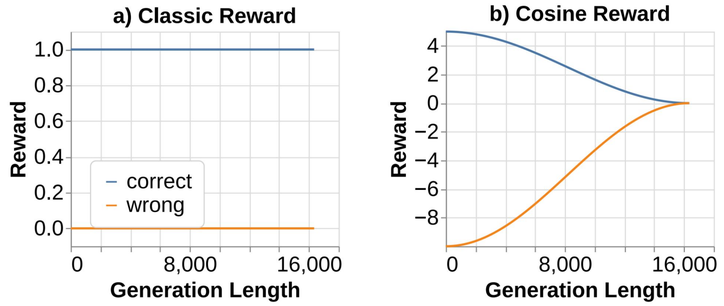
Simpler is Better: Finding the Best Reward Function in Long Chain-of-Thought Reinforcement Learning for Small Language Models

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Junkuan Liu, Zichen Zhang, Luning Wang
{junkuan, zhangzzc, lnwang}@umich.edu

Cosine Reward

- For correct answers, shorter completions are preferred, with rewards decreasing from +2.0 (shortest) to +1.0 (longest).
- For incorrect answers, longer completions are less punished, with rewards increasing from −10.0 (shortest) to 0.0

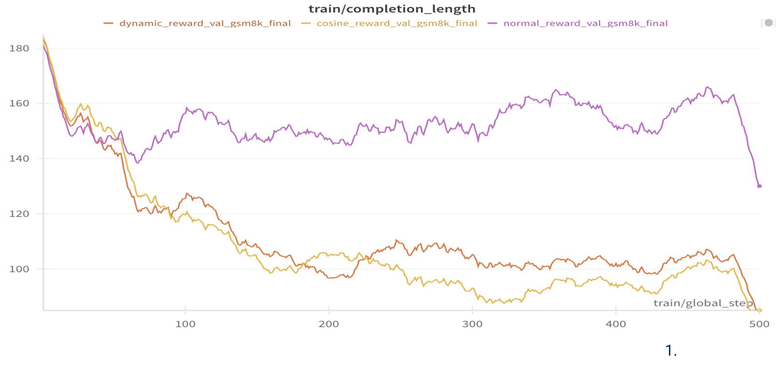


Dynamic Reward

- Start: when accuracy is low, the Cosine Reward incentivizes longer reasoning, leading to reward hacking through repetition, demanding a stronger repetition penalty.
- Rest of the training: when accuracy becomes higher, reward hacking becomes less likely, requiring a weaker repetition penalty.

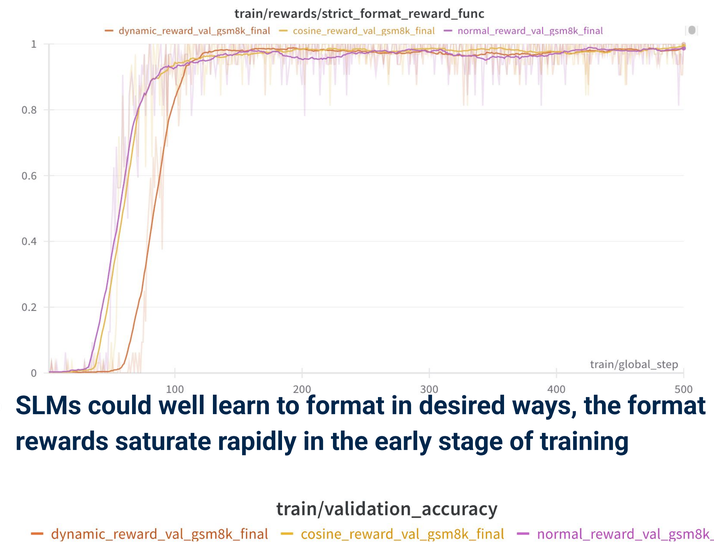
$$R_{\text{dynamic}} = (1 - w_{\text{repetition}})R_{\text{cosine}} - w_{\text{repetition}} \sum_{t=1}^{T} P_{\text{repetition}}(x_t)$$

$$w_{\text{repetition}} = \sigma(\alpha f_{\text{rep}}(X) + \beta),$$

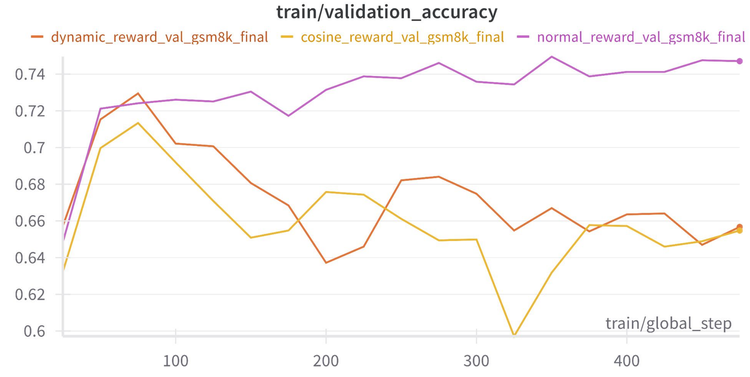


1.

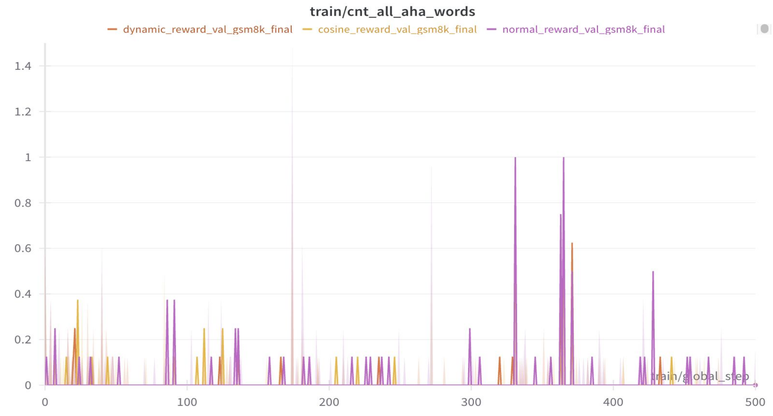
- Cosine reward and dynamic reward could have a regularization effect on completion length.



- SLMs could well learn to format in desired ways, the format rewards saturate rapidly in the early stage of training



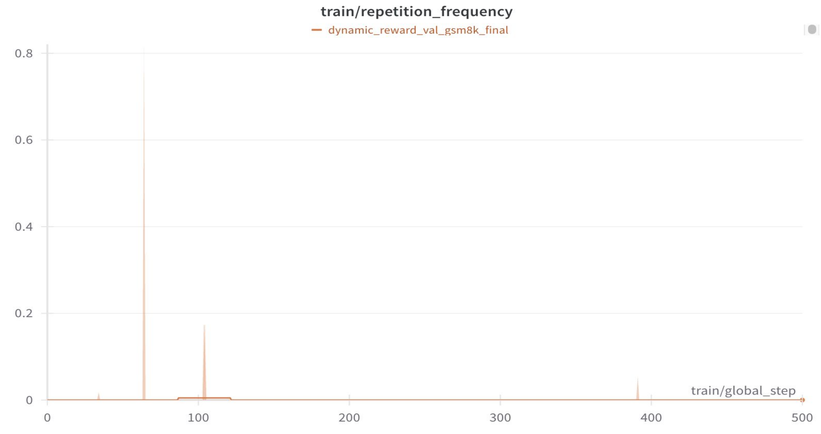
- Normal reward could consistently improve the model's performance, while the cosine reward and dynamic reward show negative impact in the long term



- 'Aha' words are observed in SLM reasoning, but their occurrences don't show an explicit relationship with reward functions.

Key Findings & Takeaways

- Simpler is better, at least for SMALL language models!
- The classic reward function of a constant number for correct answers works the best
- Different from some previous works on larger models, we do not observe any significant reward hacking in our training for all three different reward functions
- The reward function that leads to the longest reasoning path, in our case the classic reward, leads to best performance in small language models. Our hypothesis is that small models tend to require longer reasoning path.
- Dynamic reward and cosine reward does not have a significant difference on small language models
- All three rewards lead the model to easily attain perfect format rewards, such as XML count reward, int reward and other format rewards



- Our repetition penalty of n-gram tokens (n=20) were rarely activated in the training

Limitations

- We only experiment on one SLM (Qwen2.5-3B-Instruct), without observing the effects on other SLMs and larger models.
- Our case study is limited, without investigating into more samples for deeper understanding of the phenomena.