How Far Can Unsupervised RLVR Scale LLM Training?

Bingxiang He*, Yuxin Zuo*, Zeyuan Liu*, Shangziqi Zhao*, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, Xiusi Chen, Youbang Sun, Xingtai Lv, Xuekai Zhu, Li Sheng, Ran Li, Huan-ang Gao, Yuchen Zhang, Lifan Yuan, Bowen Zhou, Zhiyuan Liu, Ning Ding

Tsinghua University · Shanghai Artificial Intelligence Laboratory

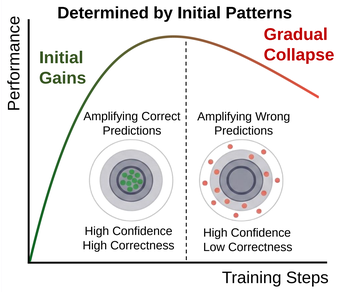
Full Paper · Code · Homepage

FOR MORE DETAILS

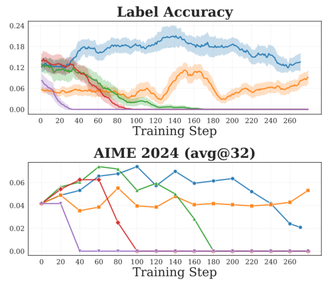
❖When Does Intrinsic Unsupervised RLVR Work?

Intrinsic URLVR universally follows a **rise-then-fall** pattern across all methods. Early gains reflect **confidence-correctness alignment** in the model's prior, while eventual collapse is inevitable when this alignment breaks down.

Success of Intrinsic URLVR depends on if initial confidence aligns with correctness

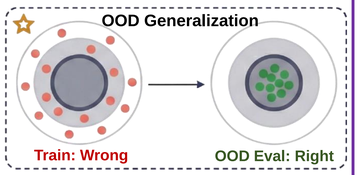


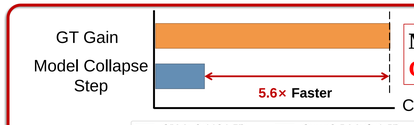
Rise-Then-Fall Pattern



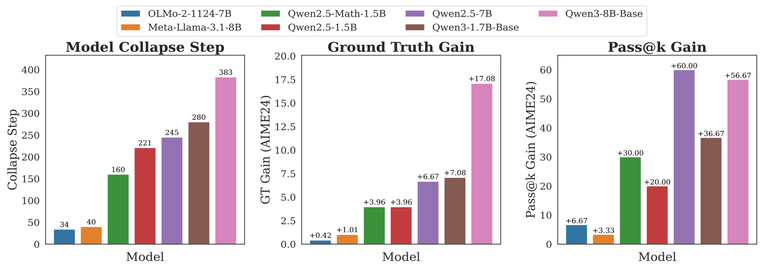
Different Intrinsic Methods Different Failures

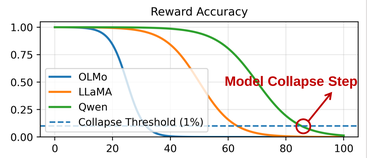
Training **amplifies initial preferences** rather than correcting errors within the same problem. Surprisingly, this amplification can still **generalize to unseen problems** and improve performance.





Model Collapse Step correlates strongly with GT Gain, better than Pass@k Gain as indicator.





Definition of Model Collapse Step

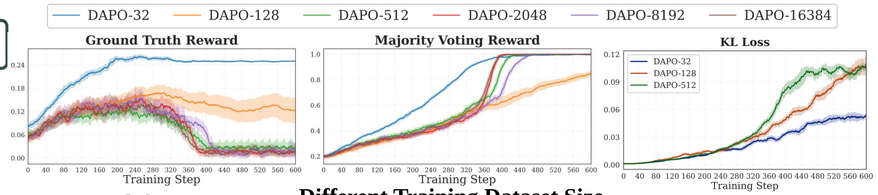
Model Collapse Step of Intrinsic URLVR can serve as indicator for RL trainability

❖How Can Sharpening from Intrinsic URLVR Be Applied Safely?

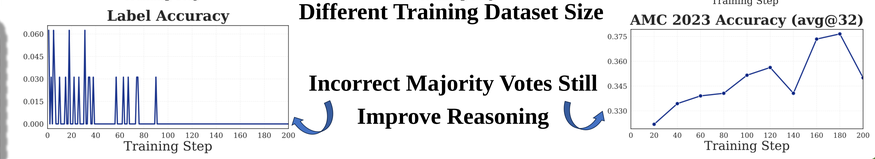
Small datasets induce localized rather than systematic policy shift, even training on wrong problems can yield gains, making **test-time training a safe and practical application.**

Intrinsic URLVR can be safely applied in Test-Time Training without collapse

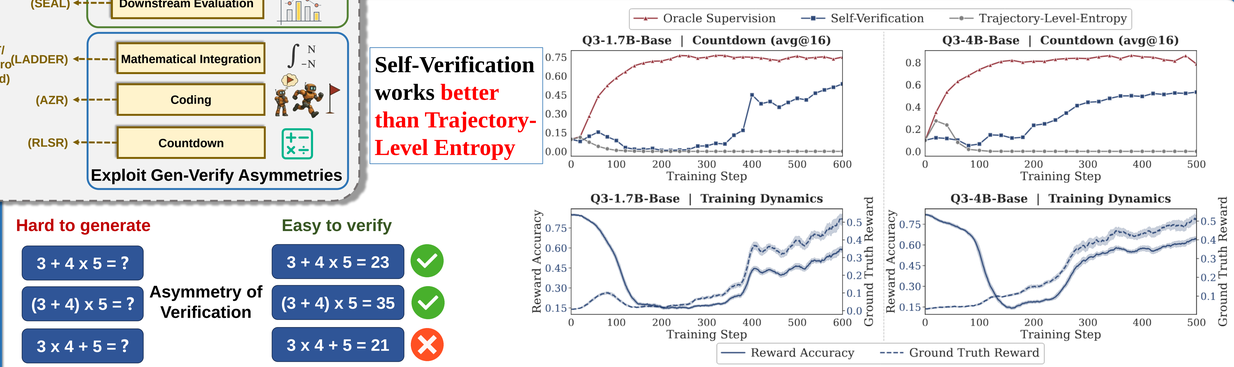
Model collapse can be prevented when training data is **sufficiently small and domain-specific**.



Different Training Dataset Size



Incorrect Majority Votes Still Improve Reasoning



Self-Verification works better than Trajectory-Level Entropy

Hard to generate

3 + 4 x 5 = ?

(3 + 4) x 5 = ?

3 x 4 + 5 = ?

Asymmetry of Verification

Easy to verify

3 + 4 x 5 = 23

(3 + 4) x 5 = 35

3 x 4 + 5 = 21

Self-Verification move towards scalable URLVR beyond intrinsic methods

We propose the **Model Collapse Step** as a novel indicator of model priors, which measures standard **RL trainability** by tracking reward accuracy collapses during intrinsic URLVR. This indicator achieves accuracy in assessing trainability **on par with running standard RL itself**, but with **higher efficiency and outperforming pass@k.**

Intrinsic rewards are fundamentally bounded by what the model already knows. **External rewards** grounded in **unlabeled data or generation-verification asymmetry** provide signals that **scale with data and computation** rather than saturating with model capacity, offering **a more promising path towards scalable URLVR.**

❖How Can We Measure Model Prior?

❖Can We Move Towards Scalable URLVR?