AUGMENTING LARGE REASONING MODELS WITH CONTRASTIVE GOAL-CONDITIONED RL

DEVAN SHAH, KEVIN WANG, DAVID YAN

MOTIVATION

As the paradigm in artificial intelligence shifts from pre-training scaling laws toward test-time training with RL, reasoning models have emerged as the next frontier. Ever since the release of DeepSeek-R1, a growing trend involves training on complex reasoning tasks with verifiable outcomes (i.e. reward of 1 if the solution is correct and 0 otherwise, with simple format rewards. We make the observation that this outcome-oriented reward paradigm is effectively a goal-conditioned setup.

Meanwhile, in the broader RL community, recent self-supervised RL algorithms have shown strong success on classical goal-conditioned settings, where sparse reward only provides a single bit of reward feedback for each trajectory. A core question thus arises: given this quasi-goal-conditioned paradigm in NLP, can these same goal-conditioned self-supervised RL methods be used to advance LLM reasoning?

INTRODUCTION

Contrastive RL: Contrastive RL [1] is a goal-conditioned method learns representations of state-action pairs ($\varphi(s, a)$) and future states ($\psi(s_f)$) such that the representations of future states are closer than the representations of random states. Formally, we maximize the InfoNCE loss:

$$\mathcal{L}(s,a,s_f^+,s_f^-) \triangleq \log\sigma(\underbrace{f(s,a,s_f^+)}_{\phi(s,a)^T\psi(s_f^+)}) + \log(1-\sigma(\underbrace{f(s,a,s_f^-)}_{\phi(s,a)^T\psi(s_f^-)}))$$

The final reward model captures the expected reward of achieving a goal state at intermediate reasoning steps.

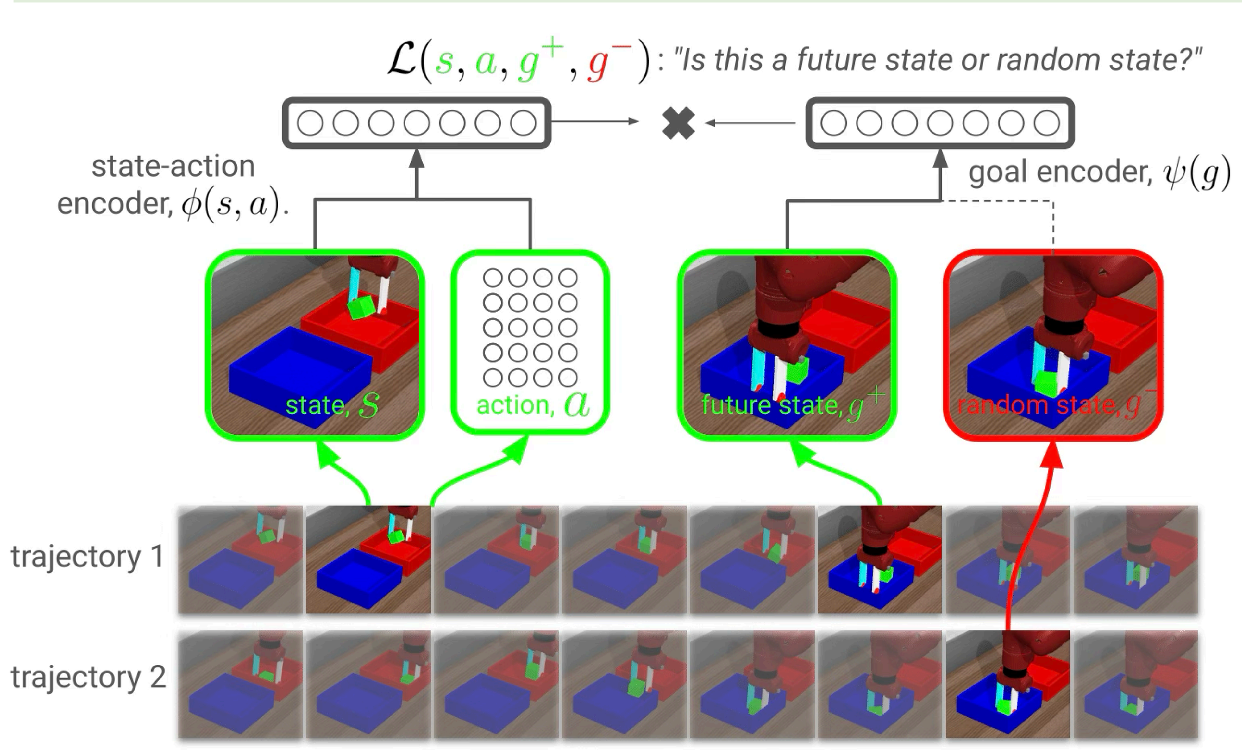

Figure 1: Classic example of GCRL for a robotics task [2].

Related Work: We build off LongProc's [3] countdown benchmark prompt and validation pipeline to train our reward model. ScaleAI presented a method to train reward models using contrastive learning [4], but their method requires a human-labeled dataset of positive and negative examples, limiting generalizability. Our method requires no labels and works fully unsupervised.

APPROACH

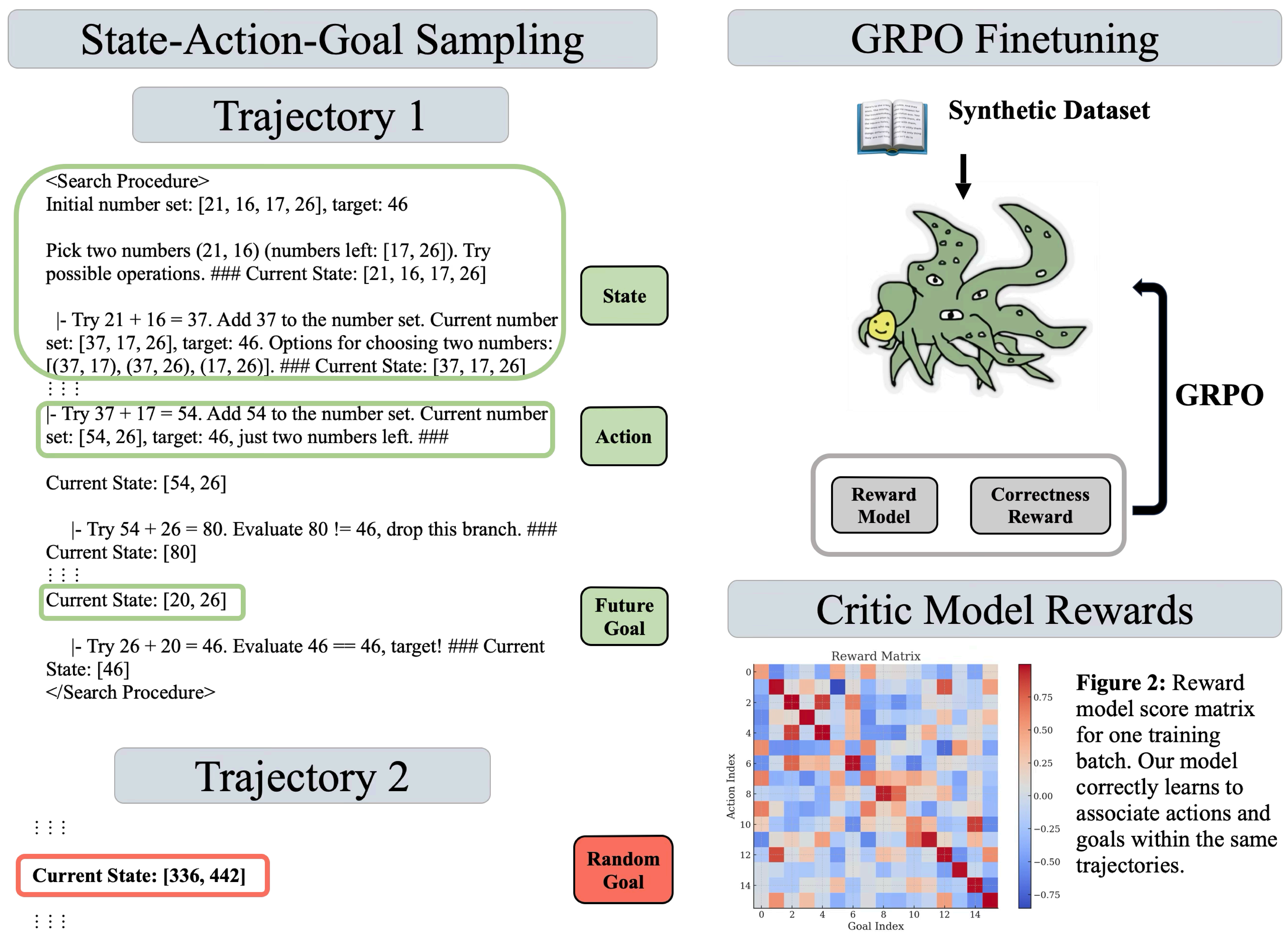

Figure 2. Reasoning traces are split into state–action pairs, each matched with a later goal from the same trace. A pretrained LM is fine-tuned on the resulting triples via GRPO, while a contrastive critic (heat-map) learns dense rewards that align actions with future goals. The improved policy then generates new traces, which are resampled into fresh triples, closing a self-supervised actor-critic loop for goal-conditioned reasoning.

Critic Model Architectures

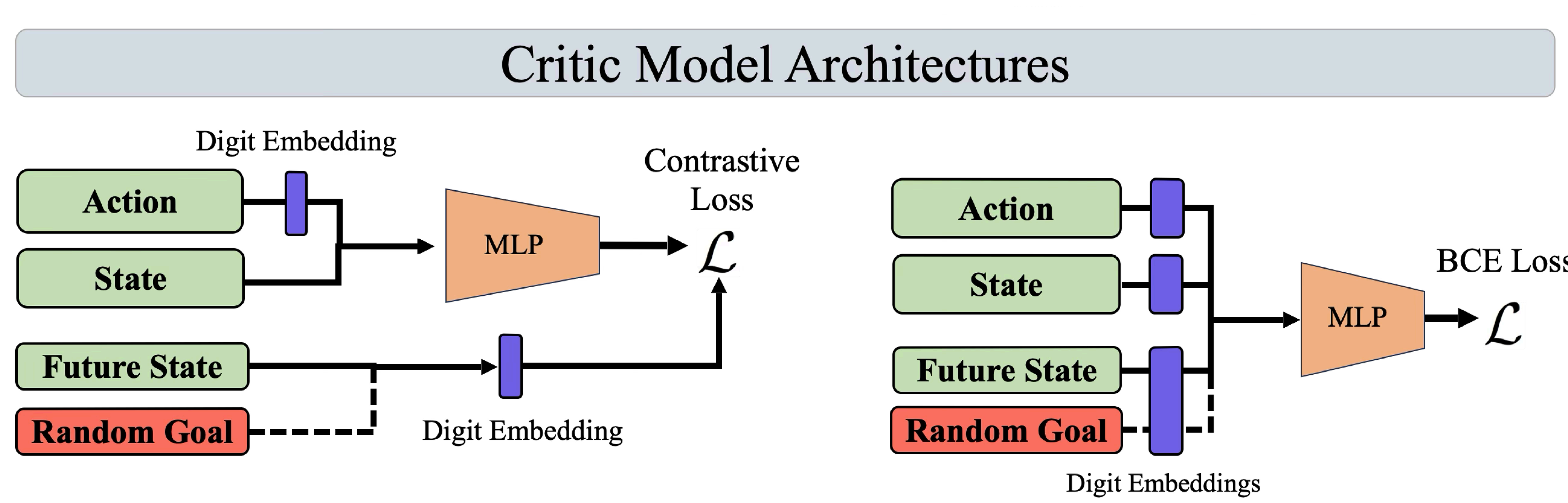

Figure 3: Architectures experimented on for the critic model. We found that the left architecture using a contrastive loss was insufficiently expressive because it relies on dot-product similarity to compare the state-action and goal embeddings.

CONTRASTIVE CRITIC

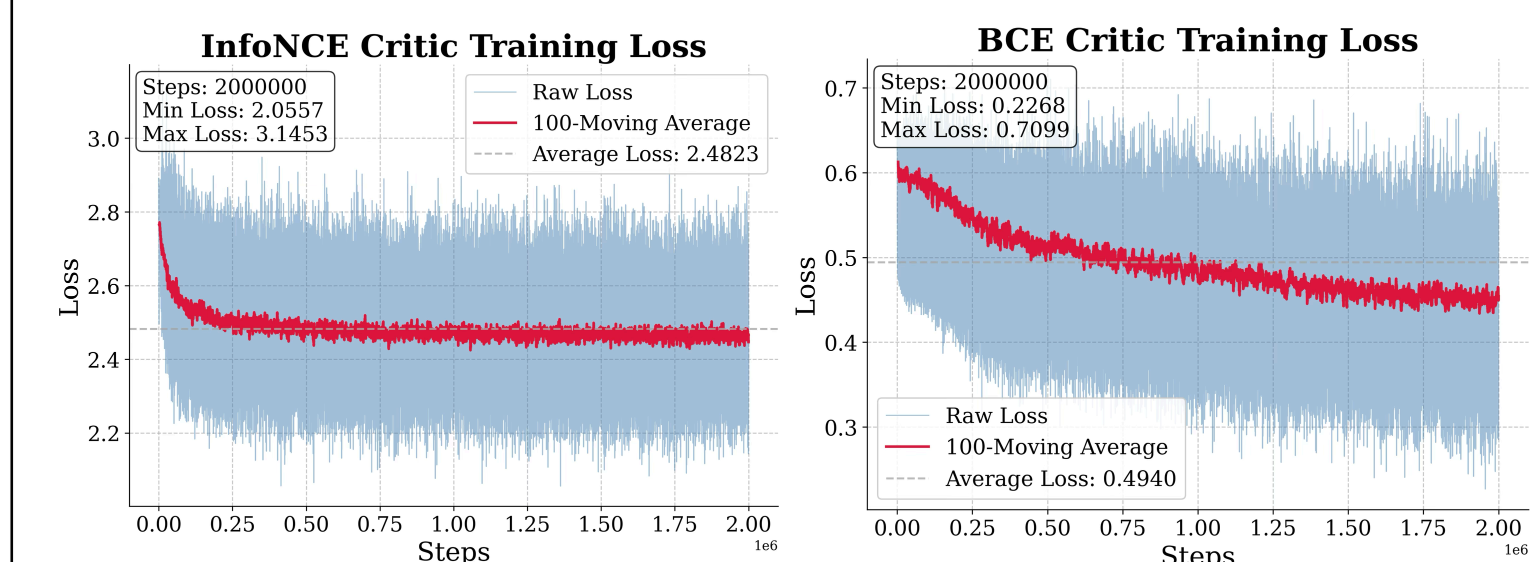

Figure 4: We train our critic models on 2,000,000 synthetic examples extracted from 100,000 procedurally generated trajectories to associate goals with trajectories.

GRPO RESULTS

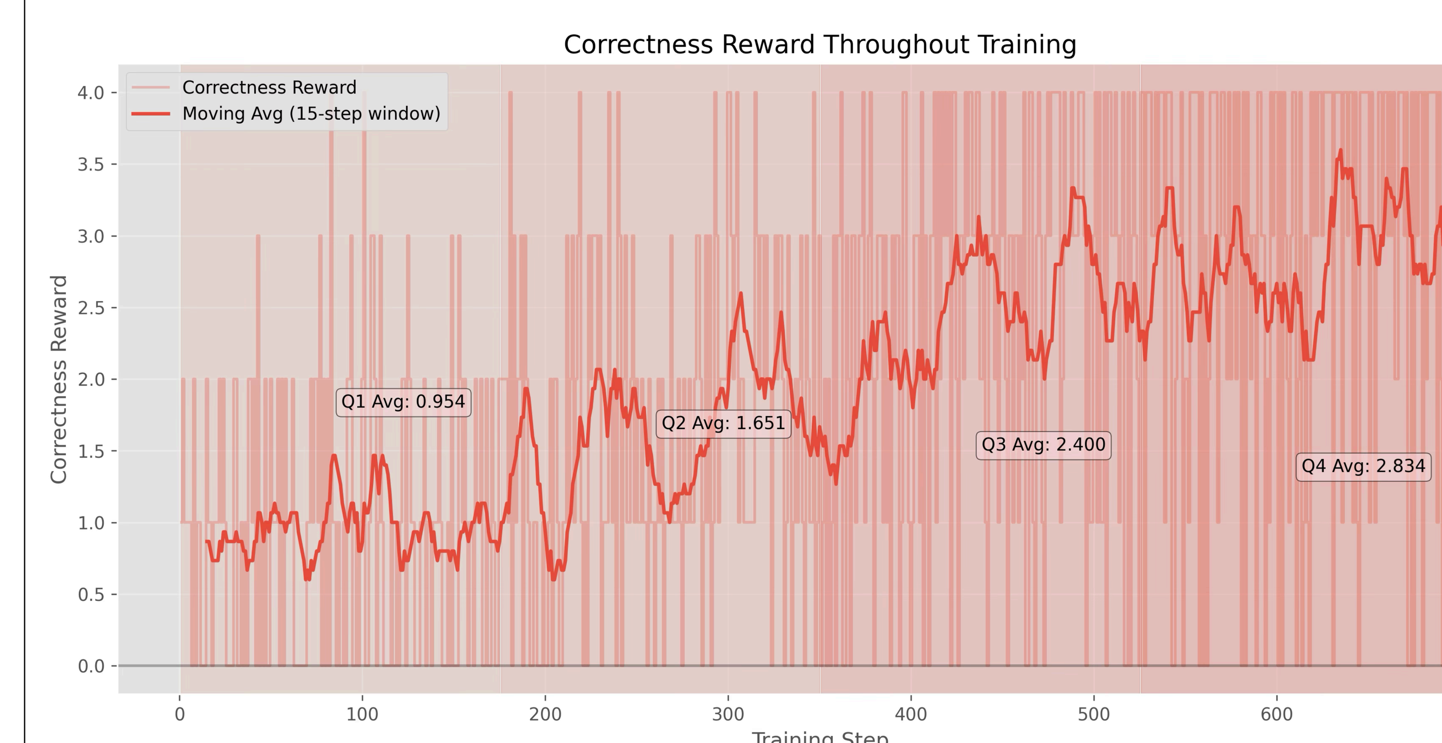

Figure 5: Baseline success rate over time during GRPO without contrastive loss.

REFERENCES

[1] DeepSeek-AI, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://arxiv.org/abs/2501.12948

[2] Benjamin Eysenbach, Tianjun Zhang, Ruslan Salakhutdinov and Sergey Levine. Contrastive Learning As a Reinforcement Learning Algorithm. https://arxiv.org/abs/2206.07568

[3] Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett and Danqi Chen. LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation. https://arxiv.org/abs/2501.05414

[4] Vaskar Nath, Dylan Slack, Jeff Da, Yuntao Ma, Hugh Zhang, Spencer Whitehead, Sean Hendryx. Learning Goal-Conditioned Representations for Language Reward Models. https://arxiv.org/abs/2407.13887