

Rate of Convergence

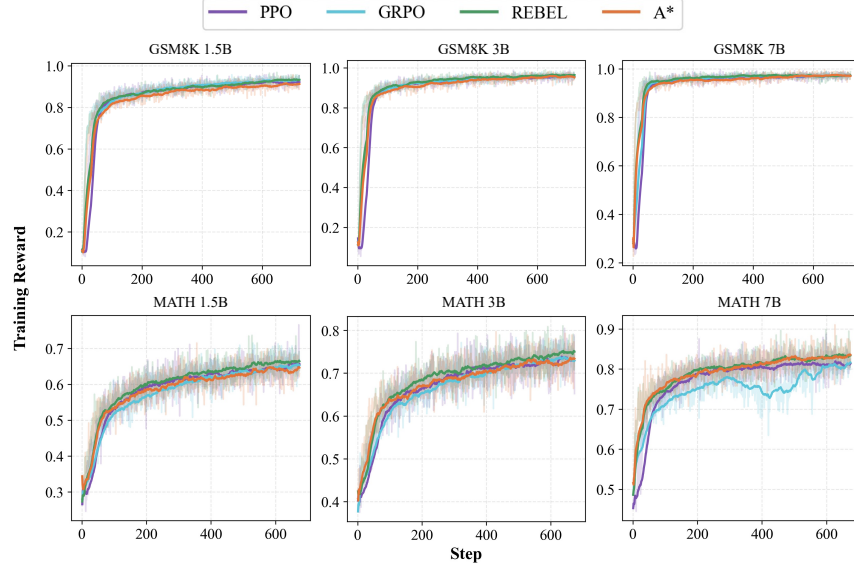


Figure 1: Training reward vs. steps over two datasets and three models.

Fig. 1 shows the training reward of PPO, GRPO, REBEL, and A^* -PO over MATH and GSM8K datasets at each step. We can see that all four methods have similar convergence behavior. Based on the plots, we can now safely conclude that A^* -PO is the fastest method as it requires the least amount of time to complete the same number of training steps, while reaching convergence at the same step count as the other methods. To illustrate this more clearly, we also include a plot of training and generation time versus training reward below. Note that the A^* -PO curves do not start at time zero, as they include the time required for data generation. Each line starts at the corresponding generation time for its dataset and model. While REBEL converges at a comparable rate in terms of time, it requires significantly more memory than A^* -PO.

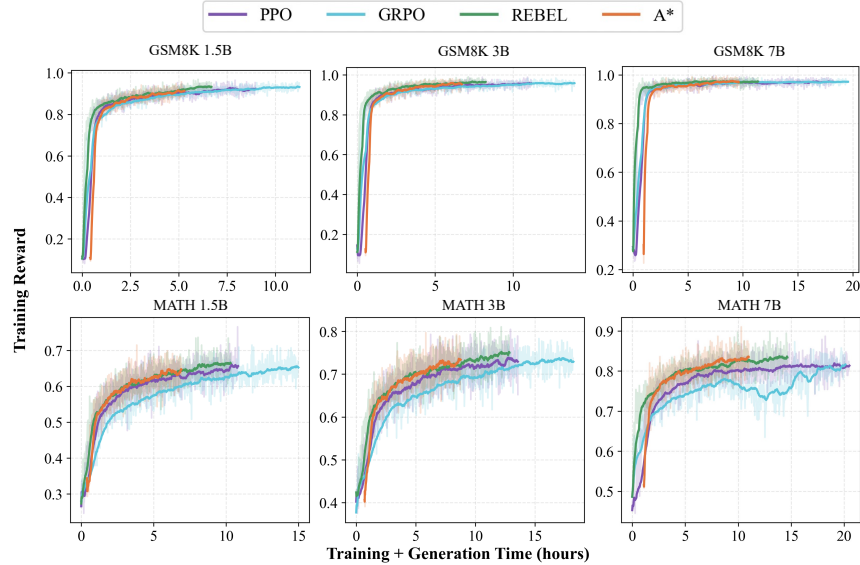


Figure 2: Training reward vs. training and generation time over two datasets and three models.