Generated Model Performance Report

Simon Green

January 29, 2025

Results and Model Comparison

This report presents the performance evaluation of six reinforcement learning models: EnTRPO, GenTRPO, TRPO, PPO, and TRPOR. The comparison is based on reward statistics across different environments. We present the data but do not draw any conclusions from it in this report.

Model Performance Table

The table below summarizes the performance of different models in terms of mean and standard deviation of rewards, along with maximum and minimum rewards recorded during training. A higher mean reward indicates better average performance, while a lower standard deviation suggests more stability.

Environment Model	Ant-v5	Humanoid-v5
entrpo	$4600.26M$ $1885.15\mu \pm 1425.33\sigma$ $2484E, 8R$	$914.93M$ $320.13\mu \pm 396.83\sigma$ $3943E, 4R$
trpo	$10.27M$ $-40.80\mu \pm 42.11\sigma$ $15E, 16R$	N/A

Performance Analysis Through Plots

To gain deeper insights into the models' behavior, the following plots provide visualizations of various performance aspects.

Learning Stability

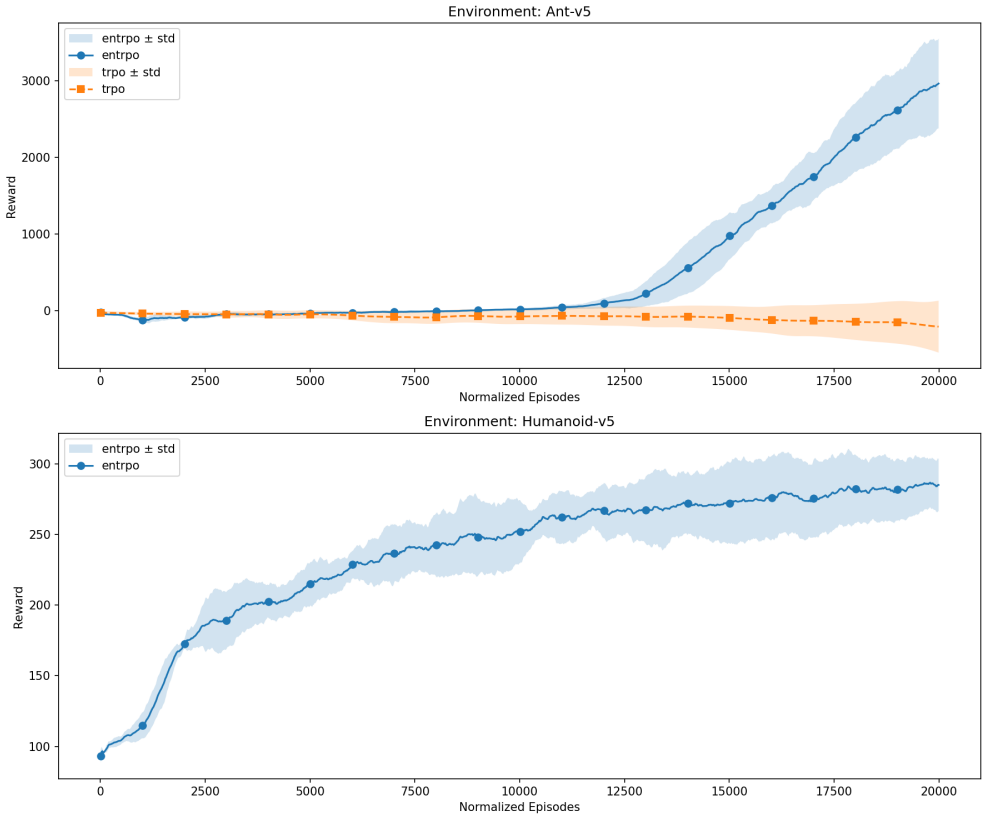

Figure 1: Learning Stability for Different Models

Learning stability measures how consistent a model's performance is over time. A smoother and more steadily increasing reward curve indicates that the model is learning in a reliable and predictable manner. Models with high variance in their learning curves may be struggling with instability or sensitivity to hyperparameters.

Learning Stability (Coefficient of Variation)

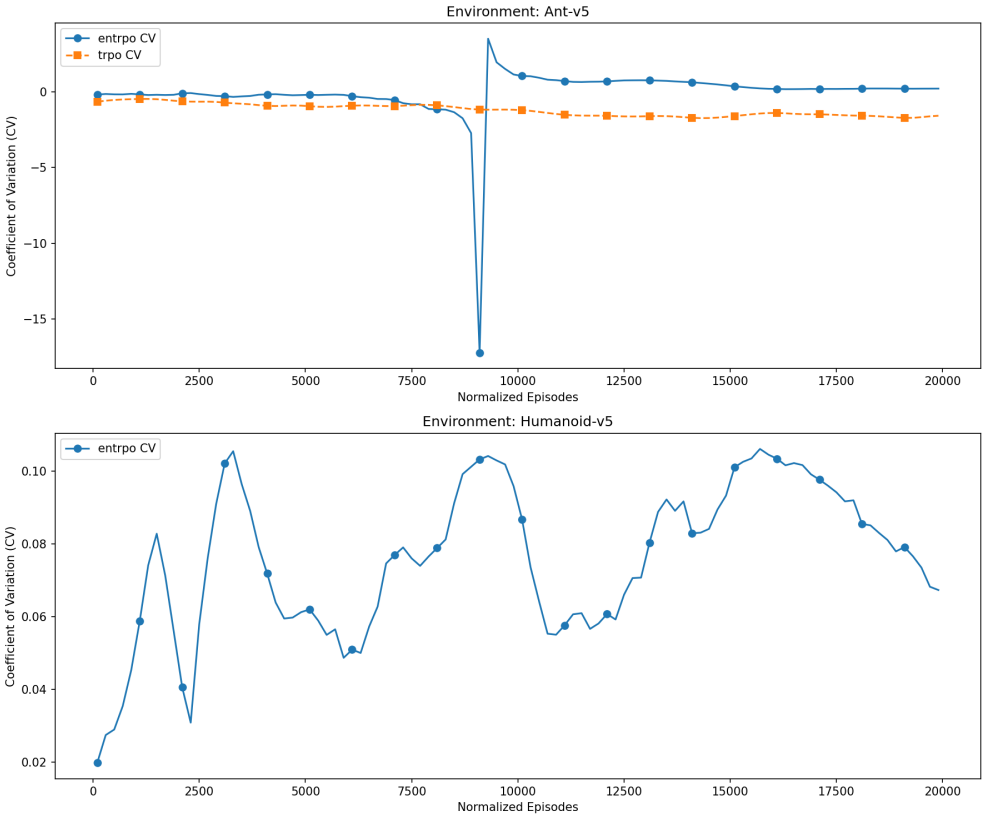

Figure 2: Learning Stability (Coefficient of Variation)

The coefficient of variation (CV) provides a normalized measure of learning stability. A lower CV indicates that the model's performance is less volatile, meaning it generalizes better across different training runs. High CV values suggest inconsistent learning, which could be due to stochasticity in training or sensitivity to random seeds.

Sample Efficiency

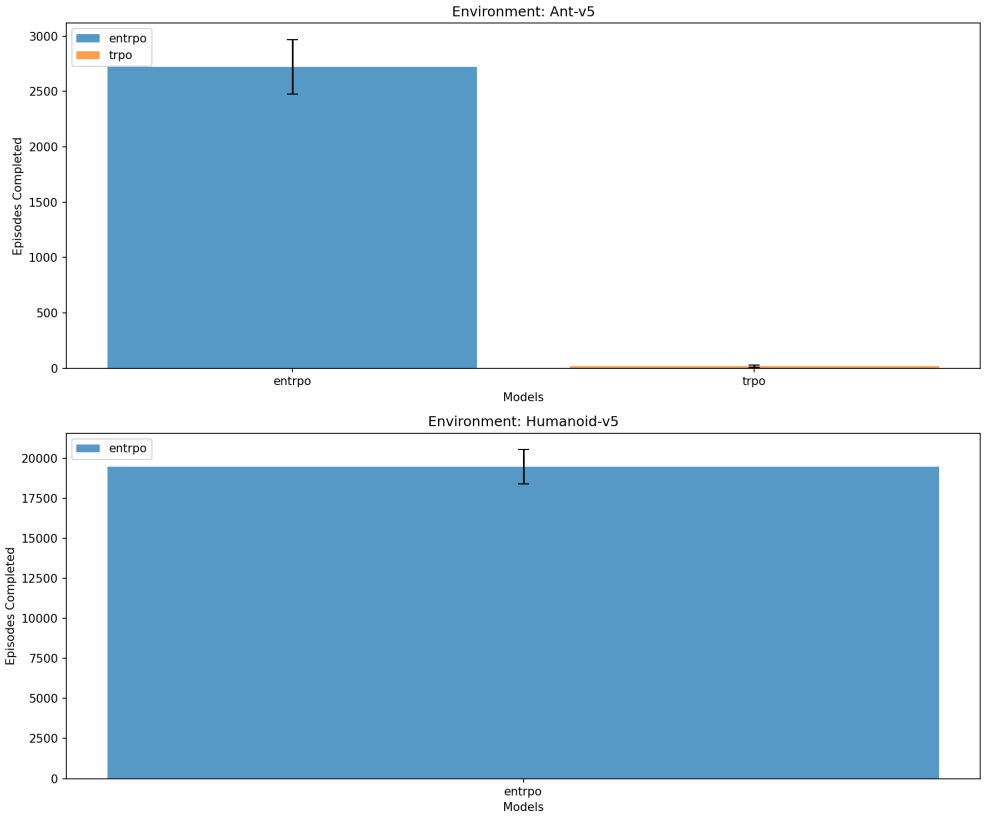

Figure 3: Sample Efficiency Across Models

Sample efficiency refers to how quickly a model improves given a limited number of training episodes. Models that reach higher rewards with fewer episodes are considered more sample-efficient. This metric is crucial in real-world applications where data collection is expensive or time-consuming.

Combined Sample Efficiency Results

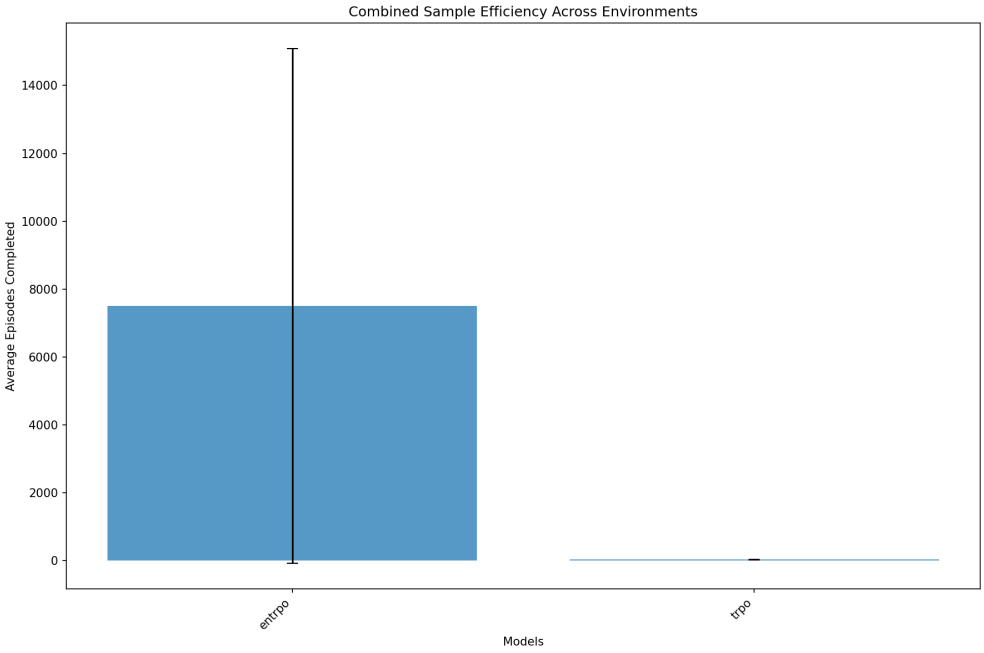

Figure 4: Combined Sample Efficiency Results

This combined sample efficiency plot compares models across all environments, showing which algorithms consistently require fewer interactions to achieve optimal performance. A model with consistently high sample efficiency is preferable for scenarios where training costs are high.

Resampled Rewards and Outlier Removal

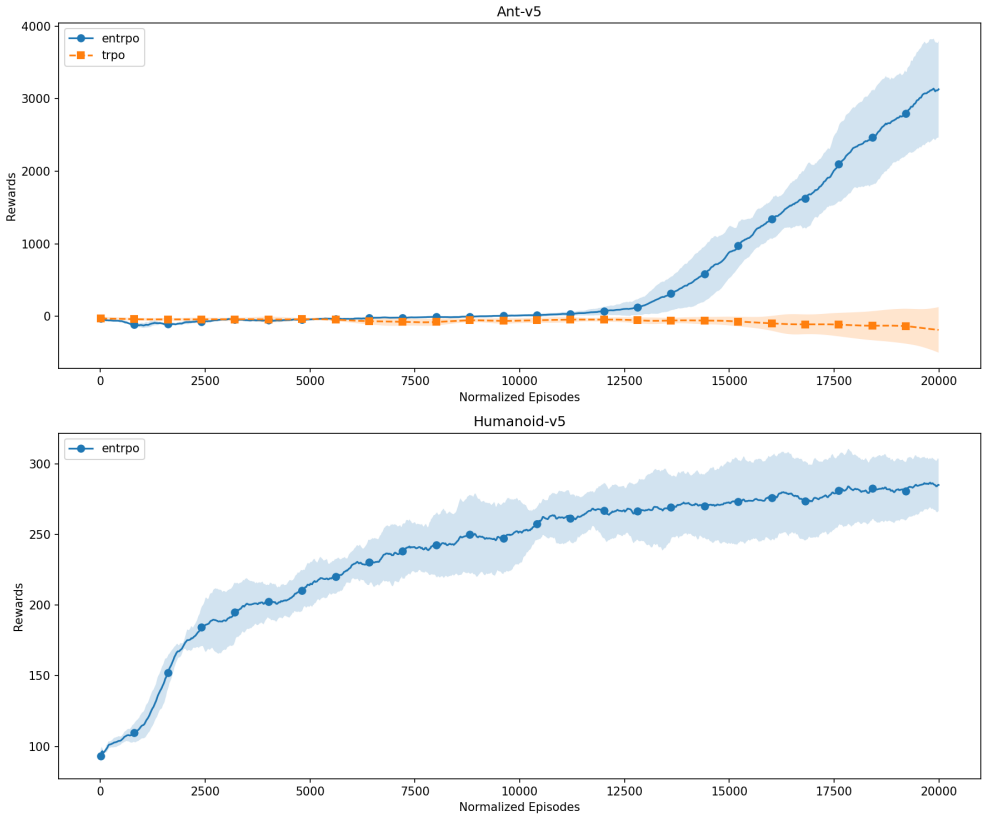

Figure 5: Resampled Rewards with Outlier Removal

This plot visualizes the reward distributions after applying smoothing and outlier removal techniques. The presence of large spikes or dips in rewards can indicate unstable learning or catastrophic forgetting. Cleaning the data helps highlight true performance trends and removes misleading fluctuations.

Raw Data

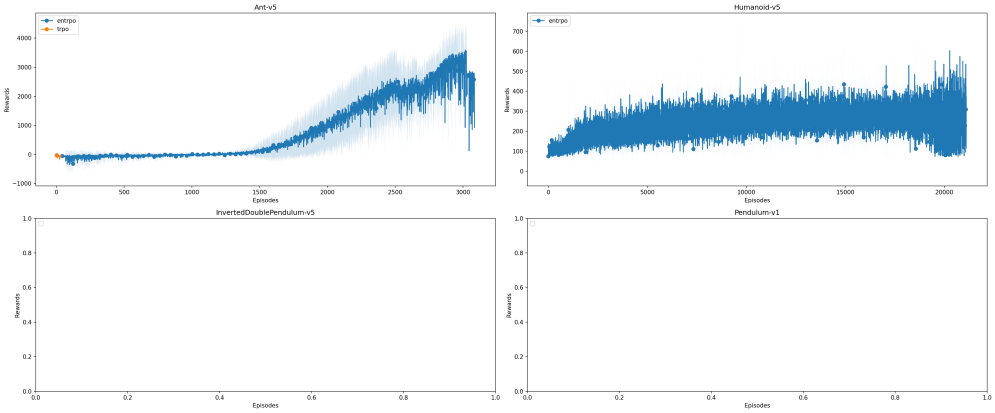

Figure 6: Raw Reward Data for Different Models

Since the plots apply resampling and smoothing to compare the models on the same episode scale and make the plots less noisy with data points distribution, the raw data plot shows the actual reward values recorded during training. We can observe that the trajectories will less episodes consumed more timesteps per episodes.