Additional Rebuttal Material

We investigate how the performance of our best model (PPO training TÜLU 2 13B with the 70B UltraFeedback RM and UltraFeedback prompts), and continue training beyond the 1 epoch result reported in the main text of the paper. We show these evaluations in Figure 1. We find that how performance changes over training is quite distinct between each evaluation: while some evaluations continuously improve (e.g., AlpacaEval 2), others improve and then degrade (IFEval, GSM8k), or remain relatively unchanged over training (MMLU). We show the AlpacaEval 2, IFEval, and GSM8k performances individually in Figure 2. This highlights the need to measure a *broad variety of benchmarks*: while just examining AlpacaEval 2 would suggest that training for 3 epochs (or longer) is best, examining IFEval we would find that the model after 500 steps (0.5 epochs) is best, and GSM8k peaks at the 1 epoch mark.

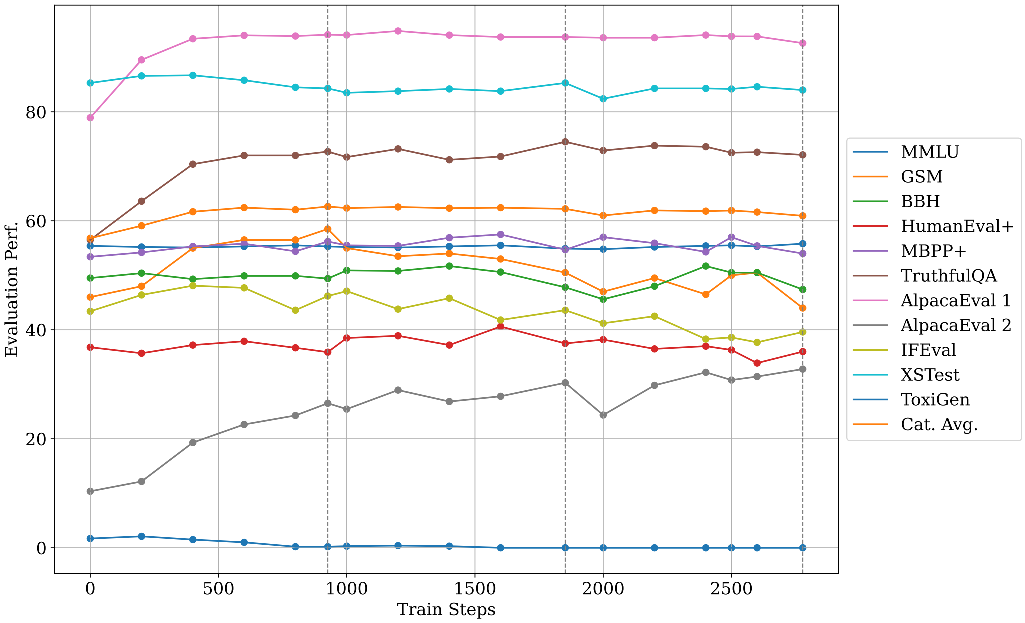

Figure 1: Performance of all evaluations over PPO training steps when training using the 70B UltraFeedback RM and UltraFeedback prompts for 3 epochs. Grey dashed lines indicate epoch boundaries.

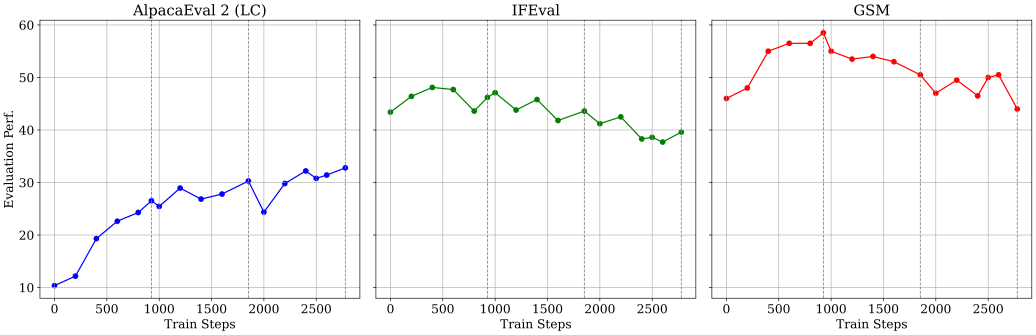

Figure 2: Performance on (left) AlpacaEval 2, (middle) IFEval, (right) GSM8k over PPO training steps when training using the 70B UltraFeedback RM and UltraFeedback prompts for 3 epochs. Grey dashed lines indicate epoch boundaries.