Performance Comparison: RLHF and DPO

Introduction

In this report, we compare two fine-tuning models, RLHF and DPO, using a GPT-2 Medium model. We evaluate their performance in terms of sample efficiency (speed), response quality, and computation cost.

Performance Comparison

Sample Efficiency (speed)

Model	**Avg Inference Time per Question**
RLHF	0.4513s
DPO	0.4511s

Both methods have nearly identical inference times.

Response Quality

Metric	**RHLF**	**DPO**
BLEU Score	0.0880	0.0880
ROUGE-1	0.1300	0.1293
ROUGE-2	0.0280	0.0279
ROUGE-L	0.0856	0.0852

1. BLEU scores are the same for both RLHF and DPO. This means that the similarity between the output and reference texts is the same for both.
2. ROUGE scores are very slightly higher for RLHF. This means that recall is better for RLHF than for DPO.

Computation Cost

Factor	**RLHF**	**DPO**
Training	Uses reward model and PPO	Direct Optimization without reward model
Computational Cost	High (one epoch takes 10 hours)	Low (one epoch takes 1 hour)
Inference Cost	Same as DPO	Same as RLHF

Conclusion

RLHF and DPO perform almost identically in terms of response quality and inference speed. But RLHF is way more expensive and complex to train because it requires a reward model and RL. On the other hand, DPO is simpler and more efficient; it gives similar results without requiring a reward model. Since both models have nearly identical performance, **DPO is a better choice because it is easier to train without requiring a reward model and requires less computational power and time**.