Comparative Analysis of RLHF and DPO in Handling Harmful or Stereotype-Triggering Questions

1. Introduction

Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) are two prevalent fine-tuning techniques for aligning language models. While RLHF incorporates reinforcement learning to align responses with human preferences, DPO directly optimizes for preference data without the complexities of reward modeling. This report compares these methods based on sample efficiency, response quality, and computational cost.

2. Methodology

The evaluation was conducted using fine-tuned versions of `gpt2-medium`, trained separately using RLHF and DPO. The models were tested on harmful or stereotype-triggering questions from the test split, generating responses for comparative analysis. The responses were quantitatively assessed using BLEU and ROUGE scores against the "more_preferred" column from the dataset.

3. Performance Metrics

Metric	RLHF	DPO
BLEU Score	0.0565	0.0559
ROUGE-1	0.1571	0.1894
ROUGE-2	0.0306	0.0307
ROUGE-L	0.1030	0.1197

4. Analysis

4.1 Response Quality

- **BLEU Score:** RLHF slightly outperforms DPO in BLEU, indicating better word-for-word similarity to the reference responses.
- **ROUGE Scores:** DPO outperforms RLHF in all ROUGE metrics, especially ROUGE-1 and ROUGE-L, suggesting superior recall and phrase overlap with reference responses.

- **Conclusion:** While RLHF ensures closer alignment with exact phrasing (higher BLEU), DPO captures more comprehensive contextual similarities (higher ROUGE scores).

4.2 Sample Efficiency

- RLHF typically requires significant reinforcement learning iterations, involving reward modeling and policy gradient optimization.
- DPO, being a simpler direct optimization approach, achieves similar results without requiring complex reward modeling.
- **Conclusion:** DPO is more sample-efficient since it reaches competitive performance without extensive iterative tuning.

4.3 Computational Cost

- RLHF involves multiple steps: reward modeling, policy optimization, and reinforcement learning, leading to increased computational demand.
- DPO directly optimizes for human preferences, eliminating the need for a separate reward model.
- **Conclusion:** DPO is computationally more efficient due to its streamlined training process.

5. Conclusion

Which is better?

- **For precision in response alignment (BLEU):** RLHF is slightly better.
- **For capturing contextual relevance (ROUGE):** DPO performs better.
- **For efficiency (both sample and computational):** DPO is superior due to its direct optimization approach.

Thoughts: If computational efficiency and scalability are priorities, DPO is preferable. However, if exact alignment with human-preferred phrasing is the goal, RLHF may be more suitable despite higher computational costs.