Mathematical Statistics
Stockholm University
Master Thesis **2026:5**
http://www.math.su.se

Fine-Tuning Language Models with Preferences for Text Summarization Tasks: A Comparative Study of Reinforcement Learning from Human Feedback and Direct Preference Optimization

Fanny Walldén*

June 2026

Abstract

The work investigates reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), two approaches used to fine-tune language models using human preference. The approaches are trained and evaluated for the text summarization task, using the ROUGE and BERTScore metrics. In addition, the language models are fine-tuned using AI preference obtained by an off-the-shelf language model to investigate their robustness to different sources of feedback. The results demonstrate the main disadvantage of RLHF, namely, the approach is prone to suffer from reward hacking. Reward hacking is a common problem in reinforcement learning that occurs because the objective diverges from the actual goal. In addition, the results demonstrate that the main advantage of DPO lies in its robustness and stability. Lastly, the results show that the approaches are robust to different sources of feedback, enabling efficient feedback generation without loss of quality.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: fawa8135@student.su.se. Supervisor: Chun-Biu Li.