# Review 169: Safe RLHF: Safe Reinforcement Learning from Human Feedback

**Paper: https://arxiv.org/abs/2310.12773v1**

https://huggingface.co/papers/2310.12773

אחד הדברים שקידמו מודלי שפה ענקיים (LLMs) בתקופה האחרונה הוא טכניקה הלקוחה מ-RL (למידה עם חיזוקים) הנקראת RLHF (למידת RL עם משוב אנושי). טכניקה זו שימשה את לטיוב(finetune) של LLMs וגרמה להם לעקוב להוראות בצורה מדויקת יותר.

היום ב-#shorthebrewpapereview נסקור מאמר שמציע שיפור לשיטה הנפלאה הזו. דרך אגב בפעם הראשונה RLHF שימשה לאימון של InstructGPT המודל שקדם ל-ChatGPT שאנחנו כה אוהבים. דרך אגב ניתן להשתמש בשיטות RLHF לא רק למודלי שפה אלא למגוון מטלות מדומיינים שונים (נעשה בהצלחה רבה ב-20 השנה האחרונות).

אז מה שזה RLHF למודלי שפה ומה כזה נפלא בו? RLHF מאפשר אימון מודל on-the-fly כלומר המודל מתאמן על דאטה שהוא בעצמו יוצר כמו בבעיות האחרות של RL. זה שונה מלמידה מפוקחת שמתבצעת על דאטהסט סגור שלא משתנה במהלך הלמידה. כדי להפעיל טכניקות RLHF לפיין-טיון של LLMs אנו חייבים מודל תגמול (reward) שנותן ציון גבוה לתשובות טובות וציון נמוך לתשובות פחות טובות.

בנייה של מודל תגמול נעשית באמצעות שימוש בדאטה מתויג על ידי אנשים כאשר כל מתייג מבצע בחירה של תשובה טובה יותר בין שתי תשובות. לאחר מכן משתמשים בשיטת PPO (Proximal Policy Optimization) כאשר בכל איטרציה המודל מתעדכן ויצר דאטה חדש (זוגות (שאלה, תשובה) חדשים). המאמר מציע לשדרג את הגישה הזו על ידי אימון (על דאטהסט שונה מזה שמודל התגמול מאומן עליו) עוד מודל עלות (cost model) שמודד עד כמה התשובה שניתנה בטוחה (לא רוצים שמודל שפה יסביר איך לשדוד בנק ולא להיתפס).

אז התשובות לא הבטוחות יתויגו עם 1 והתשובות הבטוחות יקבלו ציון -1. בגדול מאמנים את המודל העלות באמצעות מקסום של הסיכוי (המחושב לפי מודל Bradley-Terry (BT)) שהתשובה המנצחת (מבחינת הבטיחות) מקבלת ציון גבוה יותר מהתשובה המפסידה (יש עוד איבר נוסף מנסה למזער את העלות של התשובה המנצחת).

בסוף משתמשים במודל הזה יחד עם מודל התגמול כאשר המטרה היא למקסם את התגמול תוך כדי העדפה של תשובות בטוחות כלומר כאלו שיש להם ציון בטיחות גבוה. פורמלית הבעיה מוגדרת כמקסום התגמול תוך שמירה של העלות שלילית או הבטיחת חיובית (פותרים עם מכפילי לגרנז'). מאמר קצת כבד מתמטית אבל הרעיון די ברור.