# Review 146: Reinforced Self-Training (ReST) for Language Modeling, 14.09.23 https://arxiv.org/abs/2308.08998.pdf

**Paper: https://arxiv.org/abs/2308.08998v2**

שמעתם על RLHF? ראשי תיבות אלו הפכו להיות מאוד פופולריים לאור הצלחתו של ChatGPT שאומן בשיטה הזו. מאז לא מעט מודלי שפה מאומנים עם השיטה הזו למרות שיש טוענים שניתן להשיג את אותו הדבר עם SFT (self-supervised fine-tuning).   
  
אבל היום ב-#shorthebrewpapereviews נסקור מאמר שמציע שיטה לשיפור של RLHF. שיפור הביצועים המיוחל מושג על ידי שיפור איכות הדאטה on-the-fly. אבל קודם כל בואו נבין מה זה בעצם RLHF לאימון מודל שפה יסודי (foundational). הוא תהליך המורכב מ- 4 שלבים:

הוא אימון מקדים (pretraining) על קורפוס ענק

כיול מפוקח (SFT) או instruction tuning – אימון נוסף על דאטהסט מאוד איכותי (בד״כ שאלות ותשובות)

אימון מודל תגמול (reward) משערך את איכות הפלט(ציון סקלרי) של מודל שפה

הרצה של אלגוריתם RL (למשל PPO שהוא אלגוריתם on-policy אונליין) כדי לגרום למודל שפה להיות ״מיושר יותר״ עם הציפיות שלנו על ידי מקסום פונקצית התגמול ותוך שמירת המודל קרוב לזה בסוף שלב 2

כאמור המאמר הנסקר מציע שיטה לשיפור של RLHF אבל לעומת RLHF הקודם השלב האחרון מבוצע באמצעות אלגוריתם RL אופליין שמאפשר שימוש חוזר בדאטה. האלגוריתם מכיל את השלבים הבאים (הם ניסו זאת על משימת התרגום):

גנרוט דאטה ממדול שפה מאומן מהשלב הקודם וחישוב ציון (תגמול) עבור כל זוג של (טקסט, טקסט מתורגם)

בוחרים סף של הציון ומסננים את כל הנקודות שהציון שלהם קטן מהסף

מכיילים מודל על הדאטהסט הזה תוך שימוש באלגוריתם של אופליין RL (דאטה חדש לא מיוצר) – שזה סוג של למידת חיקוי (imitation learning)

מעלים את הסף ומבצעים את שלב 3 עד שהביצועים על סט אבלואציה נתון מפסיקים להשתפר

אם ביצענו שלב 1 למשך פחות מ G איטרציות חוזרים לשלב 1 אחרת מסיימים