# Review 110: [Short] RLCD: REINFORCEMENT LEARNING FROM CONTRAST DISTILLATION FOR LANGUAGE MODEL ALIGNMENT, 29.07.23 https://huggingface.co/papers/2307.12950

**Paper: https://arxiv.org/abs/2307.12950v3**

איך מאמנים את LLM לתת תשובות ענייניות ולא מזיקות? מאמנים מודל תגמול לתת ציון גבוה לתשובה רצויה וציון נמוך לתשובה פחות רצויה. אז מכיילים LLM תוך כדי מקסום של ציון מודל התגמול המאומן (+ קוהרנטיות). אבל איך יוצרים דאטה לאימון מודל התגמול.

היום ב-#shorthebrewpapereviews נסביר לכם איך לעשות זאת ללא צורך במתייגים האנושיים. הגישה שהמאמר הנסקר מציע היא די פשוטה: קודם כל מבקשים ממודל לתת שתי תשובות לשאלה הנתונה – אחת עניינית והשניה לא קשורה או אחת לא מזיקה והשנייה ממש מעליבה. אז התשובה הראשונה (והרצויה) מקבלת ציון 1 ושנייה מקבלת ציון 0.לאחר מכן מאמנים מודל התגמול על הציונים האלה וכך נחסך הצורך במתייגים האנושיים.

השלב השני הוא די סטנדרטי: מכיילים מודל שפה עם אחת השיטות של למידה עם חיזוקים (RL) המחברים של המאמר בחרו ב-PPO (Proximal Policy Optimization) כאשר בניית פונקציה התגמול עבורה מתבצעת בצורה דומה לשיטות כיול #LLM הקודמות שהשתמשו ב-RLHF.  
  
השלב השני הוא די סטנדרטי: מכיילים מודל שפה עם אחת השיטות של למידה עם חיזוקים (RL) המחברים של המאמר בחרו ב-PPO (Proximal Policy Optimization) כאשר בניית פונקציה התגמול עבורה מתבצעת בצורה דומה לשיטות כיול #LLM הקודמות שהשתמשו ב-RLHF.