# Review 182: Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models

**Paper: https://arxiv.org/abs/2312.06585v4**

https://arxiv.org/abs/2312.06585

היום אנו מדברים על אימון של מודלי שפה. באחת הסקירות האחרונות הסברתי לכם איך מאמנים מודל שפה עם RLHF (למידה עם חיזוקים המשולבת עם משוב אנושי) ולמה זה נחוץ. אזכיר לכם כי צריך להשתמש בטכניקות של RL כדי לאמן מודל בתנועה כלומר לעדכן משקלים של המודל על הדאטה שנוצר אחרי העדכון האחרון של המודל. במונחי RL הדאטה נדגם לפי ה-policy הכי מעודכן (כלומר למידה on-policy).

אימון מודלי שפה עם RLHF בצורת on-policy הוא יקר (כל הזמן צריך ליצור דאטה) ולא תמיד יציב ולכן הוצעו מספר שיטות חלופיות פחות כבדות כמו DPO ו-REST. הרעיון ב-REST הוא לא ליצור דאטה חדש בצורה on-policy אלא:

ליצור דאטהסט באמצעות מודל התחלתי ולשלב אותו עם דאטה המתויג על ידי בני אדם

לבחור את הדאטה בעלת ערכי פונקציית תגמול גבוהים מעל סף התחלתי

לאמן את המודל עם הדאטה הזה

ליצור עוד דאטה עם המודל המעודכן (כל I איטרציות אחרת לדלג על השלב)

ליצור(לסנן) דאטה עם ערך(תגמול) גדול מערך סף גבוה יותר

לאמן מודל עם דאטה חדש…

המחברים לקחו את הרעיון הזה ושכללו אותו (מבחינת הביצועים) בהתבסס על הרעיון של (Expectation-Maximization(EM ולייתר שימוש בדאטה הנוצר על ידי בני אדם. הרעיון ב-EM הוא למקסם נראות מירבית של פונקציית הסתברות p ביחס לפרמטרים כאשר הדאטה נדגם מהתפלגות אחרת q. זה מורכב משני שלבים איטרטיביים:

E: מקרבים את q מבחינת KL Divergence (אופטימיזציה)ֿ

M: ממקסים את נראות מירבית (עם דאטה הנדגמת עם q) ביחס לפרמטרים.

אז המחברים לקחו את הרעיון הזה והפעילו אותו על RL בצורה הבא:

ליצור דאטה מהמודל

עד שהתגמול בסט ולידציה עולה:

לאמן מודל בסיס (תמיד מאמנים מודל בסיס להבדיל מ-REST) כאשר כל דגימה בדאטהסט ממושקלת עם ערך התגמול. מכיוון שפונקציית התגמול במאמר היא בינארית זה שקול לאי התחשבות בדוגמאות בעלי ערך פונקציית תגמול 0

חוזרים לשלב 1 מספר איטרציות