🚀המאמר היומי של מייק 21.09.24: ⚡️🚀  
REFT: Reasoning with REinforced Fine-Tuning

ממשיכים לסקור מאמרים שסללו לכאורה נתיב ל-o1. הפעם מאמר די בסיסי יחסית שהיה שווה לסקור אותה לפני יומיים אך התעצלתי לעבור על רשימת המאמרים שבניתי כדי להבין את זה. הרווח היחיד לאלו שעוקבים אחרי סקירותיי באופן יום יומי יתבטא בכך שיהיה לכם מאוד קל להבין את הסקירה הזו אם הצלחתם להבין (בערך) את 4 הקודמות.

המאמר מניח שיש בידינו דאטהסט של שאלות ושרשרת הריזונינג המובילה לתשובה (הנכונה). המאמר מציע לשפר את יכולת הריזונינג של מודל שפה בשני שלבים:

אימון רגיל (Self-Supervised Fine Tuning): על כל שרשראות הריזונינג מהדאטהסט. כלומר המודל לומד לשחזר את שרשרת הריזונינג של כל שאלה ברמת הטוקן כמו ש נעשה ב-SFT הסטנדרטי.

אימון של למידת פוליסי (שזה המודל עצמו) מעולם Reinforcement Learning: (מכאן נגזר שם המאמר) כאשר המודל מקבל פרס 1 אם המליח לגנרט שרשרת ריזונינג המובילה לתשובה הנכונה. תגמול צנוע הרבה יותר ניתן לתשובות מספריות לא נכונות עבור השאלות שהתשובות עליהן מספריות גם כן (כמו במאמר הקודם). תגמול 0 מתקבל בכל המקרים האחרים. אימון מתבצע עם PPO די סטנדרטי עם שערוך די סטנדרטי של פונקציית ערך V ופונקצית יתרון A (כמו במאמר המקורי של ג'ו שולמן מ-openai לשעבר)

https://arxiv.org/pdf/2401.08967