⚡️🚀המאמר היומי של מייק 18.06.24:⚡️🚀

Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate REWARD HACKING

 הסקירה הזו ממשיכה את קו הסקירות האחרונות שכתבתי בנושא RLHF. כמו שאתם זוכרים פונקציית לוס ב-RLHF מכילה שני איברים: האיבר שמנסה למקסם את פונקציית התגמול (reward) והאיבר השני מנסה לשמור את מודל השפה אחרי טיוב (פוליסי סופי) קרוב למודל שמתחילים את ה-RLHF ממנו. בעבר יצאו מאמרים שהציעו לאמן כמה מודלי reward ואז למצע (או לקחת מקסימום) של כל ה-rewards של מודלים אלו עבור שאלה ותשובה נתונות של מודל שפה. זה לטענתם מקטין את הסיכוי שהמודל שפה ב-RLHF יבצע reward hacking כלומר יתכנס לפוליסי הממקסם תגמול אך בפעול מגנרט תשובות באיכות גרועה.

המאמר שנסקור היום טוען שגישה זו אינה אופטימלית כי פונקציית לוס שאיתה מאומנים מודלי reward (כלומר Bradley-Terry) גורמת לכך שכל שני מודלי reward שונים רק בקבוע שתלוי רק בשאילתה x יקבלו את אותו הערך של פונקציית לוס. בפועל זה אומר כי לכל לערכי ה- rewardֿ, המופקים על ידי המודלי, יכול להיות ממוצעים ובפועל הבחירה של המקסימלי או הממוצע מכמה מודלי כאלו עשויה להיות לא אופטימלית (כמו ממוצע של תפוזים ועגבניה). אז המאמר מציע לאמן פונקציית תגמול עם רגולריזציה שבאה ״לרסן״ את הקבוע זה שתלוי רק בשאילתה ובכך ״לסכנרן״ מודלי reward שונים.

https://arxiv.org/abs/2312.09244