🚀המאמר היומי של מייק 27.09.24: ⚡️🚀  
REWARD-ROBUST RLHF IN LLMS

הסקירה של היום הינה בנושא שהוא די דומה לסקירה של אתמול (26.09.24). נושא של הסקירה הוא שיפור של יישור (alignment) של מודלי שפה במהלך אימון RLHF. גם המאמר הזה מציע שיטה שבאה ״לתקן״ את פונקציית התגמול (reward) אבל מזווית טיפה שונה מאשר המאמר שסקרנו קודם.

המחברים מצביעים על כך ששימוש בפונקציית תגמול יחידה במהלך אימון RLHF אינו אופטימלי מכמה סיבות. הסיבה הראשונה היא חוסר עקביות בין המתייגים במהלך תיוג הדאטה המשמש לאימון RLHF (כלומר תשובות מועדפות ולא מועדפות לשאלות מהדאטהסט) שעלול לגרום לתשובות ״מבולבלות״ של המודל לאחר האימון. הבעיה השניה היא reward hacking של המודל המתבטא בכך שהמודל לומד להחזיר תשובות הממקסמות את פונקציית התגמול תוך מתן תשובות לא ״מיושרות״ עם העדפות המתייגים או לא הגיוניות.

המאמר ניגש לסוגיה זו מנקודת מבט בייסיאנית. אם נניח שקיימת פונקציית תגמול אידאלית שאין לנו גישה אליה אז ניתן להתבונן בכל פונקציית תגמול שנבנה איזה דגימה ממרחב ״פונקציות תגמול רועשות״. המחברים מציעים לכמת את אי וודאות שיש לנו בפונקציית התגמול על ידי אימון של כמה פונקציות תגמול.

אז איך כל הסיפור הזה עובד? קודם כל מאמנים פונקציית תגמול רגילה דרך נוסחת Bradley-Terry הסטנדרטי.

לאחר מכן מאמנים כמה פונקציות תגמול שימדלו לנו את אי הוודאות. בשביל זה לוקחים backbone רגיל (מודל שפה) ומוסיפים אליו כמה ראשים (heads) שכל אחד הוא למעשה פוקנצית תגמול. כל ראש מאומן לפלוט את התוחלת ואת השונות של ערך התגמול והתגמול עצמו מוגרל מהתפלגות גאוסית המוגדרת על ידיהם.

פונקציית לוס שהם משתמשים לאימון הראשים היא די לא טריוויאלית אך בגדול ממזערת את השגיאה הריבועית של שערוך התגמול (וזה קצת מורכב ומסתמך על פונקציית תגמול סטנדרטית מהשלב הראשון בנוסף לגישת Bradley Terry). במהלך האימון כל דוגמא מוגרלת (מנווטת) לראש שלו וכך אנו מקבלים כמה פונקציות תגמול.

המחברים אומרים שהם ״היו רוצים״ (והם השתמשו בה על דוגמאות הצעצוע שלהם) לבנות את הלוס עבור אימון RLHF בתור צירוף לינארי של פונקצית התגמול הרגילה התגמול המינימלי בין כל פונקציות התגמול. כאן האיבר השני למעשה מהווה שערוך של אי הוודאות שדנו בה למעלה. באופן פרקטי במהלך אימון RLHF הם בוחרים ערך התגמול המתקבל בפונקציית התגמול בעלת שונות הנמוכה ביותר.

https://www.arxiv.org/abs/2409.15360