המאמר היומי של מייק - 03.02.25  
The Perfect Blend: Redefining RLHF with Mixture of Judges

אחרי יציאת המודל האחרון של DeepSeek העניין ל-RLHF או שיטת טיוב (fine-tuning) של מודלי שפה באמצעות שיטת Reinforcement Learning with Human Feedback. חוקרי DeepSeek הראה שניתן לאמן מודל שפה חזק לעשות הנמקה(reasoning) בעיקר עם RLHF (יש קצת SFT אבל עדיין הרוב). המאמר שנסקור היום יצא כמעט 4 חודשים לפני R1 של DeepSeek והוא מציע שיטה שמשפרת ביצועים של RLHF.

אחת הבעיות הגדולות של אימון RLHF הוא reward hacking שמתרחש כאשר המודל לומד למקסם את פונקציית התגמול (reward) אך כתוצאה מכך מתכנס למודל חלש או לא בטוח (למרות איבר הרגולריזציה שמנסה לשמור את המודל הסופי קרוב למודל שממנו מתחילים לעשות RLHF). המחברים מציעים להתמודד עם הבעיה הזו בשלוש דרכים. הראשונה היא סט של אילוצים על התשובה לפרומפט (שלמשל בודק האם הוא פוגעני) הנבדק על ידי ״השופט״ (judge) שתפקידו ממלא מודל שפה אחר. השיפור השני הוא שינוי של פונקציית תגמול המתבטא בחיסור ממנו תגמול בייסליין מסוים שתיכף אסביר מהו. השינוי השלישי בבניית הוא דאטהסט עליו מאומן RLHF.

החיסור הזה מזכיר לי שני דברים. קודם כל התגמול החדש (אחרי החיסור) נראה דומה לפונקציית יתרון (advantage) (רק דומה אבל היא לא) המוכרת לנו פונקציית לוס (יעד) משיטת PPO רק שהפעם היא לא מחושבת דרך פונקציית Value באמצעות שיטות GAE אלא בדרך אחרת. תגמול חדש זה מזכיר לנו מה שראינו בפונקציית יעד של המאמר של DeepSeek, שם הבייסליין חושב באמצעות תגמול ממוצע(מעל באץ') של המודל המטויב (המתקונן עם השונות). במאמר ההוא איבר זה שימש כאומדן של אותה פונקציית היתרון.

כאמור החידוש השני של המאמר (הראשון האילוצים שאנו מטילים על פלטי המודל) הוא הבייסליין המחוסר מהתגמול. המחברים מציעים לקחת את הבייסליין בתור התגמול עבור דוגמאות (תשובות) הזהב(= מועדפות) מדאטהסט של SFT (שאלות ותשובות) או מהשאלות עם התשובות המועדפות מדאטהסט של RHLF. כך התגמול שלנו הוא עד כמה התשובות של המודל המאומן נראות יחסית לתשובות המועדפות מבחינת תגמולן.

השינוי השלישי הוא בפונקציית יעד. בנוסף (המחברים מציעים 2 וריאנטים) למקסום של הנראות של התשובות המועדפות ומזעור הנראות לתשובות הפחות מועדפות (מבחינת התגמול), המאמר מציע רק למקסם את הנראות של התשובות המועדפות בלבד (באופן מפתיע זה עובד). המחברים גם ״מלבישים״ את הרעיונות הללו ששיטות קלאסית של RLHF כמו DPO ו-RAFT.

https://arxiv.org/abs/2409.20370