🚀המאמר היומי של מייק 26.09.24: ⚡️🚀  
RRM: ROBUST REWARD MODEL TRAINING MITIGATES REWARD HACKING

מאמר נחמד שמשך את עיניי עקב העובדה שהוא דן בנושא פונקציית תגמול (reward model או RM) של מודלי שפה. RM הנחוץ בתהליך היישור (alignment) של מודלי השפה המבוססים על RLHF שמטרתו מאוד בגדול לאמן מודל שפה להבחין בין תשובה טובה לתשובה רעה.

הנושא נחקר באינטנסיביות בשנים האחרונות והוצעו מספר שיטות לעשות רובן שכלולים שונים של (Proximal Policy Optimization (PPO כגון DPO, ORPO ועוד רבים שחלקם סקרתי. בדרך כללי לאימון RLHF נדרש דאטהסט המורכב משלישיות של שאלות ו-2 תשובות, אחת יותר מועדפת (המנצחת או w) והשנייה הפחות מועדפת (מפסידה או l). במהלך אימון RLHF המודל לומד להגדיל את הנראות של התשובה w להקטין את הנראות של תשובה l דרך מקסום של הפרש ה-reward שלהם (עם סיגמויד ולוג) תחת אילוצים כמו שמירה על הקרבה בין התפלגות הפלט של המודל המאומן למודל ההתחלתי.

המאמר מציע להתבונן באימון RLHF מזווית די מעניינת ושואל את השאלה הזה האם הצורה של תשובות משפיעות לנו בצורה לא מכוונות על תוצאת אימון בלי קשר לשאלה. כלומר המודל עושה "reward hacking" ומשתמש בתכונות של התשובות בלבד ללא קשר לשאלה כדי לאפטם את משקלי המודל. כלומר המודל יכול ללמוד לנצל דפוסים שונים כמו (sure, this is the response או n-grams מסוימים של התשובות) בלבד.

כדי להתגבר על הבעיה הזו המאמר מציע לערבב תשובות לשאלות שונות כלומר לעשות סוג של אוגמנטציה ולאמן את המודל כך שזה יקשה עליו לבצע reward hacking. למשל שתי תשובות לא רלוונטיות משאלות אחרות (w ו- l) לשאלה נתונה אמורות לקבל אותו התגמול ואילו תשובה w המתאימה לשאלה ותשובה l משאלה אחרי אמורה עדיין לתת reward גבוה ל-w ו-reward נמוך ל-l מהשאלה האחרת. יש כמובן צירופים נוספים שניתן להנדס ולאמן את המודל עליהם בצורת RLHF.

דרך אגב המאמר בונה פריימוורק סיבתי לבעיה הזו כולל DAG, סטים שהם d-separate וכדומה אבל אני לא בטוח שכל זה נחוץ להבנת המאמר . זה אמנם שגזל ממני זמן רענון המושגים האלו אבל כמה שיחות עם סונט עזרו לי מאוד.

https://arxiv.org/abs/2409.13156