המאמר היומי של מייק: 19.07.25  
GENARM: Reward Guided Generation with Autoregressive Reward Model for Test-Time Alignment

עבר כבר שבוע מהסקירה האחרונה והרגשתי צורך דחוף לסקור איזה מאמר. האמת די הרבה זמן לא היתה לי הפסקה כזו גדולה לצערי גם רוחב הפס שלי אינו אינסופי. טוב, יאללה מתחילים לסקור.

המאמר מדבר על גנרוט דאטה באמצעות מודל שפה תוך התחשבות(כוונון) במודל reward (תגמול) חיצוני האומד את איכות הטקסט המגונרט. האיכות נאמדת רק כאשר הגנרוט נגמר בסוף הטקסט כלומר עבור התשובה המלאה. נציין כי ניתן להשתמש בטריק שידוע לנו משיטת DPO שזה Direct Preference Optimization כדי להתחשב בציון ממודל התגמול עבור התשובה כדי להכווין את התפלגות הגנרוט של המודל.

שיטת תיקון ממודל התגמול נובעת מהנוסחה עבור פונקציית לוס של אימון מודל שפה עם LHF כאשר המטרה (של האימון) היא למקסם את הציונים של התשובות של המודל עם רגולריזציה שמנסה לשמור את התפלגות המודל המאומן קרובה להתפלגות ההתחלתית של המודל במונחי מרחק KL. בד"כ מבצעים אימון כזה על דאטהסט של שאלות עם תשובות רצויות ולא רצויות שכאמור למקסם את יחס הציונים ביניהם. התיקון מתבצע ללוג של הסתברות של משוב המגונרט המלא y (בהינתן ההקשר x) על ידי החיבור של הציון (r(x, y (ממושקל) ופונקציית נרמול התלויה ב x בלבד (המאמר לא מרחיב על אופן שערכו).

אז איך כל הסיפור הזה (גנרוט עם פונקציית תגמול ללא אימון RLHF) עבד לפני המאמר הזה? בזמן הגנרוט בשביל לגנרט טוקו הבא בהינתן הטוקנים שכבר גונרטו אנו דוגמים כמה המשכים עד סוף התשובה ואז ניתן להשתמש במודל תגמול בשביל לשערך את איכותה. אז הטוקן שהוא נמצא ההשלמה בעלת הנראות המתוקנת הגבוהה ביותר. הסיבה לכך היא העובדה שלא ניתן לשערך רק את התשובה המלאה ולא חלקית שלא מאפשרת חישוב התיקון עבור כל טוקו מגנט בצורה ישירה. יש עוד שיטות לעשות את זה אבל הם או לא יעילות או ביצועיהן לא כאלו טובות.

המאמר המסוקר מציע לאמן מודל שמטרתו היא לשערך (r(x, y עבור תשובות חלקיות בהתבסס על הדאטהסט של שאלות עם תשובות רצויות ולא רצויות. המאמר ממקסם את היחס סכום התגמולים עבור כל הטוקנים של התשובות הרצויות לאלו של לא רצויות. מודל זה כמובן מתבסס על מודל שפה עם ראש מאומן כמו שנעשה עבור אימון מודל תגמול רגיל עבור תשובות מלאות. המחברים טוענים כי מודל תגמול קטן יחסית למשל 7B מסוגל לשפר את איכות הגנרוט בהתאם ל alignment הרצוי עבור גנרוט למודל הרבה יותר גדול עם 70B פרמטרים.

בצורה כזו ניתן לבצע גנרוט בהתבסס על כמה מדיניות הalignment שכל אחת מיוצגת על ידי דאטהסט משלה. אחרי אימון של מודל התגמול עבור כל אחת מהן ניתן לבנות את התיקון ללוג של נראות עבור טוקו הבא על ידי סכום ממושקל של התגמולים עבור כל אחת מהן כאשר המשקול תלוי במידת התחשבות בכל אחת מהמדינות alignment אלו.

מאמר קליל אך עם זאת די מעניין

https://arxiv.org/abs/2410.0819