המאמר היומי של מייק: 23.07.25

Reinforcement Pre-Training

חוזר מחופשה עם סקירה מאוד קצרה של רעיון מאוד מסקרן ודי אינטואיטיבי לאימון של מודל שפה. אנו רגילים שבשלב הראשון של אימון מודל שפה, הנקרא אימון מקדים, אנו מאמנים אותו על מה שנקרא next token prediction או NTP. כלומר בהינתן דאטהסט עצום ולא מתויג אנו ממקסמים את הנראות (likelihood) עבור כל טוקן בדאטהסט בהינתן ההקשר שלו כלומר כל הטוקנים לפניו. המטרה כאן היא למקסם את הנראות המשוערכת של הדאטהסט עם המודל המאומן (ניתן לראות זאת באמצעות שימוש פשוט בחוק בייס). ד״א ניתן לראות די בקלות שבאמצעות אימון מקדים כזה המודל מוסגל לרכוש מיומנויות רבות כלומר ידע במגוון תחומים, פתרון שאלות פשוטות וכדומה.

אחרי השלב הראשון באים השלבים של alignment כלומר SFT שזה Supervised Fine Tuning וגם RLHF (עם כל סדר ביניהם). המאמר שסוקרים היום שואל את השאלה הבא: למה לא ניתן לבצע אימון NTP על כל הדאטהסט עם למידה עם חיזוקים או RL. מתברר שזה אפשרי ויש לזה פוטנציאל לשיפור ביצועי המודל.

איך עושים זאת בפועל? עבור כל טוקן בטקסט אנו מבקשים מהמודל לעשות תהליך ריזונינג קצר כדי לנחש את הטוקן הבא. המודל מתבקש ליצור כמה מסלולי חשיבה כאלו - המסלול שמנחש את המילה בצורה נכונה מקבל תגמול 1 כאשר השאר מקבלים 0. לאחר מכן ניתן להשתמש בתגמולים אלו כדי לאמן את המודל בשיטה האהובה שלכם מעולם למידה עם חיזוקים (PPO, GRPO וכל שיטה אחרת). כלומר מקרה די קלאסי של RLVR או שזה RL עם verifiable rewards.

ההבדל העיקרי בין שיטת אימון מקדים זו ל-pretraining הרגיל של מודלי שפה הוא שימוש שונה בחיזוי הטוקן הבא - לא דרך סופטמקס אלא תגמול בינארי. המאמר כמובן מוכיח אמפירית שזה משפר את ביצועי המודל.

מאמר נחמד - קריאה קלילה לסופ״ש….

https://arxiv.org/abs/2506.08007