המאמר היומי של מייק - 18.02.25  
THINKING LLMS: GENERAL INSTRUCTION FOLLOWING WITH THOUGHT GENERATION

סקירה מספר 400 - כדי לא להכביד עליכם יותר מדי בחרתי מאמר קליל יחסית והסקירה הולכת להיות בלי נוסחאות ודי קצרה. המאמר מציע שיטה קצת במהות דומה Group Relative Preference Optimization או GRPO בקצרה שעשתה הרבה כותרות לאחרונה. ותיכף אני הולך להסביר למה אני מתכוון כאן. רק אציין שהמאמר מציע שיטה להגברת יכולת הנמקה כללית של מודל ולא מתמקד רק בשאלות תכנות ובעיות מתמטיות.

המאמר מציע שיטת טיוב (fine-tune) למודלי שפה המתקדמת בהקנייתם יכולת הנמקה (reasoning) למודלי שפה ללא צורך בדאטה מתויג. המאמר מציע לבצע אימון בסגנון RLHF אבל להבדיל מהדיפסיק (הממציאה של GRPO), המחברים הציעו להשתמש בשיטת DPO שלא משתמשת בפונקציית התגמול כלל. אציין ש-GRPO לא מאמנת מודל תגמול (reward) כמו ש-PPO עושה אלא משתמשת בנכונות התשובה והפורמט שלה כפונקצית תגמול.

אז מה משותף בין GRPO לבין השיטה המוצעת במאמר? שניהם למעשה מציעים לא לקנוס את המודל על תהליך החשיבה (שעלול להיות לא נכון אך להוביל לתשובה הנכונה) אלא לשפוט אותו רק על בסיס נכונות התשובה של המודל (כאמור GRPO גם קונס על אי עמידה בפורמט של התשובה). אחרי שהבנו את הקשרים המהותיים של השיטה המוצעות עם השיטות המפורסמות בואו נצלול למה שהמאמר מציע.

כאמור המאמר מציע לטייב יכולת הנמקה של מודל שפה ללא שימוש בדאטה מתיוג עם RLHF. כמו שאתם זוכרים RLHF עם DPO דורש זוגות של תשובות מועדפות ופחות מועדפות. מכיוון שאמרנו שהשיטה לא דורשת דאטה מתויג אז אתם יכולים לנחש שבניית הזוגות נעשית על ידי מודל שפה שופט שבוחר תשובות טובות ורעות בדומה לשיטת RLAIF שזה קיצור של Reinforcement Learning from AI Feedback. מודל השופט מופעל על תשובות (ולא שרשרת הנמקה!) של המודל המאומן ומחליט מה בין תשובות היא הטובה והגרועה ביותר. זוגות אלו משמשים לאימון המודל בצורת DPO. כמובן שיש פה גם הנדסת של מטה-פרומפט הגורם למודל ״לחשוב״ אבל שרשרת חשיבה זו לא משתתפת באימון המודל.

https://arxiv.org/abs/2410.10630