🚀המאמר היומי של מייק 20.09.24: ⚡️🚀   
Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning

ממשיכים בסקירות מאמרים ״החשודים״ בסלילת נתיב למודל o1 (שרבים כבר התאכזבו ממנו אמנם אך אותי הוא מסקרן מבחינת חידוש הפרדיגמה). המאמר שנסקור היום פחות מתמטי מזה של אתמול (הכל פורסם בערוץ הטלגרם שלי) ובתקווה הסקירה תהיה יחסית קצרה וקולעת.

מזכיר שהמאמר מציע שיטה לשיפור הרוזונינג של מודלי שפה כאשר יש לנו דאטהסט D גדול יחסית של שאלות ותשובות ודאטהסט קטן בהרבה של שאלות ותשובות עם שרשרת ריזונינג. המאמר מציע שיטה בסגנון של למידת curriculum די נפוצה בלמידה עמוקה - כמה מודלי שפה הכי טריים אומנו עם השיטה הזו (בשילוב עם עוד שיטות כמובן). בלמידת curriculum מאמנים מודל החל מדוגמאות קלות ובמהלך הלמידה מעלים את קושי הדוגמאות.

אבל איך קשורה למידת curriculum לשיפור יכולת ריזונינג של מודל שפה. וזה בדיוק היופי של המאמר דרך אגב. המחברים שמו לב שאם נספק למודל את כל שרשרת הריזונינג מהתחלה ועד השלב די קרוב לתשובה הסופית אז יהיה לא יותר קל לשחזר את השלבים החסרים בשרשרת. וזה בדיוק מה שהמאמר עושה. כלומר המאמר מאמן את מודל (בשיטת RL דומה ל-STaR שסקרתי ב 17.09, למידת פוליסי די סטנדרטית) אבל הפעם המודל לומד לשחזר את שלבי הריזונינג מנקודות שונות בשרשרת.

המאמר טוען ששיטת למידת curriculum הסטנדרטית פחות מתאימה למקרה הזה כי המודל שלמד להשלים שלבי ריזונינג אחרונים מתקשה ללמוד לעשות את מההתחלה ו״מאבד״ את הידע שצבר. בעקבות כך המחברים מאמנים משימות ריזונינג ברמות קושי שונות (בהקשר המדובר) יחד עם איזושהי אסטרטגיה חכמה מעולם ה-multi-tasking.

שני דברים אחרונים לגבי המאמר הזה. קודם כל פונקצית תגמול (reward) הינה די סטנדרטית כאן עם חידוש קטן שעבור משימות עם תשובה מספרית המודל מקבל פרס קטן (ולא אפס) אם הוא נותן תשובה מספרים לא נכונה (ו-1 במקרה של תשובה נכונה). המאמר משתמש ב-PPO שהיא שיטה די סטנדרטית לפיינטיון של LLM אם אתם לא רוצים שהוא ישכח את כל מה שהוא למד לפני הפיינטיון.

https://arxiv.org/pdf/2402.05808