🚀המאמר היומי של מייק 23.09.24: ⚡️🚀  
Training Language Models to Self-Correct via Reinforcement Learning

 ממשיכים בקו הסקירות על שיפור יכולת הריזונינג של מודלי שפה (מסדרת ״כל הדרך ל o1"). המאמר הזה של דיפמיינד שיצא לפני כמה ימים משך את עיניי מרגע ששמתי לב עליו (לראשונה ראיתי אותו בלינקדאין נראה לי). לקח לי לא מאוד זמן להבין את העיקר של המאמר הזה כי הוא מכיל הסברים מאוד מפורטים ומעמיקים והיתה לי הרגשה ש״מרוב עצים לא רואים את היער״.

אוקיי כמו שאתם כבר מבינים מהשם המאמר מציע שיטה לשיפור של יכולות תיקון עצמית (self-correction) של מודלי שפה. הנושא נחקר רבות בשנתיים האחרונות (וגם לפני) והוצעו מספר שיטות לטיפול בבעיה. אלא, כמו שמחברי המאמר מציינים שיטות אלו אינן מובילות לשיפור ביצועים משמעתי עקב העובדה שהן מאומנים על התפלגות מוטעית של התפלגות התשובה הראשונה (שאותה מתקנים) של ה-LLM (זה מה שלקח לי לא מעט זמן לזקק מהמאמר).

המאמר מתבונן בשתי שיטות לתיקון עצמי (הם עשו SFT על הדאטהסטים המגונרטים על ידיהם): Star (שסקרתי לפני כמה ימים) ומהמאמר הזה (נקרא Pair-SFT במאמר). בגישה בנו דאטהסט על שלישיות המכילות שאלה, תשובה לא נכונה (כלומר שרשרת ריזונינג המוביל אליה) ותשובה נכונה (גם הריזונינג שהוביל אליה) כאשר ניתנה על ידי המודל אחרי התיקון העצמי (עם פרופמט מסוים). במקרה השני השלישיה הורכבה מהשאלה, תשובה לא נכונה ותשובה נכונה אקראית (לא אחרי התיקון עצמי) לשאלה הזו.

בשני המקרים המחברים ראו שאין שיפור משמעותי אחרי התיקון העצמי ואחרי אנליזה די רצינית הגיעו למסקנה כי זה נובע מאי ״התאמה של התפלגות התשובה הראשונה״ להתפלגות ההתחלתית של המודל. הרציאונל כאן הוא שאנו מאמנים מודל לתקן לא בדיוק מה שהמודל יוצר אלא משהו קצת אחר.

המחברים מציעים שיטה דו שלבית לפתרון בעיה זו. בשלב הראשון אנו מנסים לגרום למקסם את תגמול(מניחים שיש פונקצית reward נתונה) עבור תשובה נכונה אחרי תיקון עצמי (כלומר שרשרת ריזונינג שמובילה לתשובה הזו תוך שמירה של התפלגות הפלט המאומנת (או פוליסי בשפת RL) של LLM קרובה להתפלגות התחלתית שלו. כלומר עושים סוג של PPO כאשר היעד הוא מקסום של ההפרש בין התגמול עבור התשובה הנכונה לבין מחקר KL בין ההתפלגות המאומנת (כלומר פוליסי) להתפלגות ההתחלתית.

בשלב השני ממקסמים את סכום התגמולים אחרי שתי התשובות (לפני ואחרי התיקון) תוך שמירה של הקירבה של ההתפלגויות שלהם להתפלגות ההתחלתית של LLM.

מקווה שהסברתי פחות או יותר מובן…

https://arxiv.org/pdf/2409.12917