המאמר היומי של מייק - 09.04.25  
O1-CODER: AN O1 REPLICATION FOR CODING

סוף סוף הגעתי לסקור את המאמר הזה שעשה לא מעט רעש בזמנו. המטרה המוצהרת של מחברי המאמר היא לחקות את o1 של OpenAI על משימות קידוד. המאמר השתמש בטכניקות RLHF בשילוב עם שיטת self-play שבה המאמר לומד על הדאטה שהוא עצמו מגנרט. המאמר מתחיל מדאטהסט של של שאלות קוד והשתובות על שאלות אלו (כלומר קוד :)).

הרעיון העיקרי של המאמר מכיל 6 שלבים עיקריים. בשלב הראשון המחברים בונים כלי(המאמר לא מרחיב על זה יותר מדי) לגנרוט טסטים מקיפים עבור שאלת קוד והקוד הנכון עבורה. בהמשך כלי זה(TTG) ישמש לשערוך של ה-reward עבור קוד שנבנה על ידי O1-CODER.

בשלב השני באמצעות MCTS שזה ראשי תיבות של Markov Chain Tree Search בונים את שרשראות הנמקה (reasoning) עבור הדוגמאות מהדאטהסט. MCTS הוא אלגוריתם לתכנון בקבלת החלטות שמבצע דגימה במרחב המצבים (טוקנים במקרה שלנו) כדי לשערך את הreward הפעולות האפשריות. האלגוריתם בונה עץ חיפוש באופן הדרגתי – בכל צעד הוא בוחר לפתח את הענף(סדרת טוקנים) שנראה הכי מבטיח, תוך איזון בין חקירה של אפשרויות חדשות לבין ניצול של מה שכבר נמצא כמוצלח. כל מסלול בעץ (שרשרת הנמקה הכוללת פתרון) מקבלת תגמול 0 או 1 עם TTG(עובר או לא עובר את כל הטסטים).

בשלב השלישי המודל עובר SFT על שרשראות ההנמקה שהובילו לפתרון הנכון (עם ציון 1). בשלב הרביעי מתחילים את אימון ה-self-play בצורה איטרטיבית כאשר דאטהסט האימון מועשר בכל איטרציה עם הדוגמאות הנוצרות על ידי המודל עצמו. בהתחלה מבצעים אימון SFT של המודל על הדאטהטס עם התשובות הנכונות בךבד(פרט לאיטרציה 0) או מבצעים אימון RLHF עם DPO (שזה Direct Preference Optimization) על הזוגות של דוגמאות חיוביות ושליליות.

לאחר מכן אנו מגנרטים שרשראות הנמקה עם המודל (פרט לתשובה הסופית) ומשתמשים במודל תגמול PRM(שזה Process Reward Model) למתן תגמול לשרשראות הנמקה אלו. אז בונים את התשובה על השאלה משרשרת ההנמקה ויוצרים טסטים לשאלה זו (ידועה לנו התשובה הנכונה לכל שאלה - כנראה השאלות הן חלק מדאטהסט גדול של שאלות פתורות). אחרי זה מחשבים את ה-reward על ידי הרצת טסטים על התשובות שגונרטו על ידי המודל (1 - הטסטים עברו, 0 - לא עברו) ומשלבים אותו עם התגמולים שהתקבלו במהלך הריזונינג (נקרא aggregation function). מאמנים את המודל במטרה למקסם את התגמול הזה (עם שיטת RL כלשהי) - כנראה שיש כאן איזושהי רגולריזציה אבל המאמר לא מרחיב על זה.

בסוף יוצרים דוגמאות עם המודל אחרי העדכון האחרון ומוסיפים אותם לדאטהסט ומתחילים מחדש את השלב הרביעי (self-play).

מאמר מאוד מעניין…

https://arxiv.org/abs/2412.00154