⚡️🚀המאמר היומי של מייק 15.09.24: ⚡️🚀  
Q\*: Improving Multi-step Reasoning for LLMs with Deliberative Planning

אחרי סערה החשיבה בזמן האינפרנס במודל החדש של openai התחלתי לבנור בפוסטים בנושא הזה ונתקלתי במאמר הדי מפורסם הנקרא Q\*. מתברר שהוא נמצא אי שם ברשימת המאמרים האינסופית שאני רוצה לסקור אך לא ב-20 הראשוניים אפילו. מכיוון שקיימות די הרבה סקירות של המאמר הזה ייתן סקירה יחסית קצרה בלי לרדת לפרטים יותר מדי.

המאמר מדבר על ״תהליך החשיבה או תכנון״ עבור מודלי שפה. למעשה זה סוג של CoT מנוהל על ידי פונקציית Q המשערך ערך של כל שלב במהלך ״החשיבה״ של המודל. כלומר עבור כל שלב ב-reasoning אנו רוצים להבין עד כמה מענה נתון של LLM יקרב אותנו לתשובה הסופית הנכונה. אתם מריחים כאן פונקציית Q ידוע מעולם למידה עם חיזוקים וזה הניחוש הנכון כאן.

כדי לפרמל את הבעיה במונח RL צריך להבין מה זה מצב (state) ופעולה (action). במקרה שלנו פעולה היא תשובה של LLM בשלב נתון של תהליך החשיבה שלו ומצב הוא סדרה של כל הפעולות עד השלב הזה כלומר כל התשובות (בסדר כרונולוגי) שהמודל נתן. והמטרה כאמור לבנות את פונקציית Q בהינתן מצב s\_t ופעולה a\_t נתונים בשלב t, כלומר לשערך את איכות תשובה a\_t עבור התשובה הקודמות a\_1, ….a\_t-1. ברגע שיש בידנו את Q אנו יכולים לבנות את ההמשך האופטימלי של שרשרת החשיבה a\_1, ….a\_t-1. כמובן היינו רוצים פונקציית Q אופטימלית כלומר כזו שמקיימת משוואת בלמן ובעלת תכונות טובות.

אבל איך נוכל לשערך את הפונקציה הזו אם יש בידינו רק מודל עם פרמטרים נתונים שלא מותאם (ישירות) לכל הסיפור של בחירת שרשרת חשיבה אופטימלית. כלומר אין לנו פוליסי אופטימלי שאותה אנו יכולים למנף ליצירת Q אופטימלי. המאמר מזכיר 3 אפשרויות.

בהינתן דאטהסט נתון של שרשראות חשיבה וציונים ניתן לשערך Q אופטימלי יחד עם השערוך שלו עבור הפוליסי המוקפא שלנו (כלומר מודל שפה) בצורה alternating (שערוך של של כל אחד באמצעות השני כל פעם).

מריצים את הפוליסי הקיים וכל פעמים בוחרים את הפעולה (תשובה) בעלת ערך Q מקסימלי, ומשפרים את שערוכה באמצעות חישוב של התגמול הכולל (עבור כל השלבים). דרך אגב קביעת מה זה התגמול המידי במצב s\_t לא נראה לי טריוויאלי

שימוש במודל שפה חזק אחר כדי ״לחקות״ את הפוליסי האופטימלי ובאמצעות הרצתו לשערך את Q האופטימלי.

כאמור ברגע שיש לנו שערוך טוב של Q האופטימלי אנו תמיד בוחרים את התשובה בעלת Q הגבוה ביותר מפול התשובות של LLM.

אז למה יש כן כוכבית בשם. האלגוריתם שהתקבל מאוד מזכיר את A\* המפורסם אך זה כבר נושא לסקירה אחרת…

https://arxiv.org/pdf/2406.14283