🚀המאמר היומי של מייק -01.10.24: ⚡️🚀  
Larger and more instructable language models become less reliable

שנה טובה, מתוקה ושקטה לעוקביי היקרים! אני חושד שהמאזן הקלורי של רובכם הופר בבוקר אז אני מביא לכם סקירה קלילה (פורמלית של אתמול). ודרך אגב הסקירה של היום תהיה אוסף של כל הסקירות עד עכשיו ואני אפרסם את זה מחר בבוקר.

המאמר שנסקור היום הוא לא מתמטי והוא דן ביכולות של מודלי שפה. המדד מתבונן ביכולות של מודלי שפה לפתור בעיות דרך הפריזמה של 3 מדדים שונים. השניים מהם הם די סטנדרטיים וברורים והם אחוז נכונות/אי נכונות של התשובה אך השלישי הוא אחוז הימנעות של מודל שפה מהתשובה. אכן בלא מעט מקרים מודלי שפה בוחרים להגיד לנו שלא יודעים את התשובה ולפעמים זה די מעצבן (אבל לפעמים ממש לא).

המחברים מצאו כי LLMs נכשלים ביצירת "אזורי פעולה אמינים לבעיות קלות": אפילו במשימות הנתפסות כפשוטות על ידי בני אדם, LLMs ממשיכים לעשות טעויות. כלומר אין "מקלט בטוח" ברור של באיזור קושי נמוך שבו המודלים מבצעים באופן עקבי ללא שגיאות.

שיפורי ביצועים (הנובעים מאימון דאטה יותר טוב, אימון משופר ויישור) מתרחשים בעיקר עבור בעיות מורכבות, בעוד lLLMs ממשיכים לטעות במקרים קלים: כלומר LLMS יותר חזקים מראים ביצועים משופרים במשימות מאתגרות. עם זאת, שיפור זה אינו מתרחב באופן אחיד למשימות פשוטות יותר, מה שיוצר חוסר התאמה בין ציפיות אנושיות לביצועי המודל.

אימון יעיל (המאמר קורא לזה shape-up) מפחיתים הימנעות אך מגבירים אי-נכונות של התשובות: המאמר מראה שמודלים חדשים וחזקים יותר פחות נוטים להימנע ממתן תשובות. עם זאת, הפחתה זו בהימנעות מלווה לעתים קרובות בעלייה בתשובות לא נכונות במקום תשובות נכונות.

בנוסף אחוז הימנעות לא עולה עם רמת הקושי של הבעיה: היינו רוצים כי Prob(הימנעות|קושי) יהיה קבוע, כלומר מודלים היו נמנעים מלענות לעתים קרובות יותר ככל שקושי המשימה עולה. אולם המחברים מראים ששיעורי ההימנעות נשארים יחסית קבועים בכל רמות הקושי.

המחברים גם בדקו את יציבות תשובות המודל לניסוחים שונים של הבעיה ומצאו כי מודלים חזקים יותר מפגינים יציבות גבוהה יותר לניסוח המשימה (פרומפט). כלומר תשובתם פחות תלויה בניסוח הבעיה. למרות שיפורים ביציבות, עדיין יש אזורים (של בעיות) שבהם הביצועים יכולים להשתנות משמעותית בהתאם לניסוח שנעשה בו שימוש, אפילו עבור מודלים מעוצבים.

בנוסף השיפורים ביציבות התשובה לא מונוטוניים (מבחינת קושי הבעיה): חלק מהניסוחים (של הבעיה) מבוצעים טוב יותר במקרים מורכבים אך גרוע יותר במקרים קלים: הקשר בין יעילות הניסוח וקושי המשימה אינו תמיד פשוט. חלק מהניסוחים שעובדים היטב למשימות מאתגרות עשויים לבצע באופן גרוע במשימות קלות יותר, מה שמסבך את תהליך בחירת הניסוח.

עוד תוצאות מעניינות רבות במאמר הזה - ממליץ בחום להעיף מבט…

https://www.nature.com/articles/s41586-024-07930-y