המאמר היומי של מייק - 31.01.25  
Law of the Weakest Link: Cross Capabilities of Large Language Models

מבוא והגדרת הבעיה:

המחברים מדגישים פער קריטי במחקר ה-LLM הקיים - הנטייה להתמקד בהערכת יכולות מבודדות תוך התעלמות ממשימות מהעולם האמיתי הדורשות מיומנויות מרובות(aka AGI :)), המכונות יכולות צולבות (cross-cap). המאמר ממסגר בעיה זו באמצעות טקסונומיה מקיפה של 7 יכולות בודדות ושבע יכולות צולבות, כגון קידוד וחשיבה ושימוש בכלים וקידוד. כדי להתמודד עם המורכבות הטבועה בהערכת הצמתים הללו, המחברים מציעים את CrossEval, מדד המורכב מ-1,400 הנחיות מתויגות על ידי בני אדם המיועדות לבדוק את ביצועי ה-LLM במשימות רב-ממדיות.

דוגמאות ליכולות צולבות:

קידוד וחשיבה: פרומפט בקטגוריה זו עשוי לבקש מהמודל לנתח קטע קוד ולקבוע אם הוא מיישם נכון פונקציה מתמטית מורכבת. משימה זו דורשת לא רק ידע בקידוד אלא גם חשיבה לוגית כדי לאמת את נכונות הפונקציה.

שימוש בכלים וחשיבה: בדוגמה אחרת, הנחיה עשויה לדרוש מהמודל להשתמש בכלי אחזור מידע מבוססי אינטרנט כדי לענות על שאלה לגבי מגמות מזג אוויר היסטוריות, ולאחר מכן לספק הסבר אנליטי שלב-אחר-שלב של הדפוסים הנצפים. משימה זו דורשת הן יכולות חשיבה והן שימוש בכלים חיצוניים.

מתודולוגיה:

הגדרות יכולת מקיפות: הם בונים טקסונומיה מפורטת של יכולות בודדות וצולבות, המסווגת משימות לקטגוריות רחבות ותתי-קטגוריות מדויקות.  
מדד CrossEval: מסגרת הערכה חדשנית זו מורכבת מ-1,400 הנחיות, 4,200 תגובות מודל, ו-8,400 דירוגים אנושיים. מערך ההנחיות כולל משימות ברמות קושי שונות, החל משאלות עובדתיות פשוטות ועד למשימות מורכבות הדורשות יכולות צולבות.

הערכה מבוססת LLM: המחקר מציג מסגרת הערכה מרובת-התייחסויות שבה מעריכים מומחים מעריכים את איכות התגובות המרובות של המודל בסולם ליקרט. המחברים גם מפתחים אסטרטגיית הערכה מבוססת הפחתת נקודות לדיוק משופר.

ניתוח דינמיקת יכולות צולבות: המחברים מוצאים שביצועי יכולות צולבות לעתים קרובות מצייתים ל"חוק החוליה החלשה ביותר" — שבו הביצועים מוגבלים על ידי היכולת האינדיבידואלית החלשה ביותר.

ממצאים ניסיוניים:

הממצאים חושפים מספר תובנות מפתח המדגישות את המגבלות והחוזקות של ה-LLM הנוכחיים כאשר הם מתמודדים עם פונקציות יכולת צולבות.

חוק החוליה החלשה ביותר:

התצפית הבולטת ביותר היא שביצועי היכולות הצולבות מוגבלים על ידי היכולת האינדיבידואלית החלשה ביותר, בהתאם ל"חוק החוליה החלשה ביותר". מתוך 58 תרחישי יכולת צולבת שנבדקו ב-17 מודלי LLM, 38 הראו ביצועים נמוכים יותר מכל אחת מהיכולות האינדיבידואליות המעורבות, בעוד ש-20 ציונים נמצאו בין היכולות החזקות והחלשות אך היו קרובים הרבה יותר לחלשה יותר. למשל, במשימות המשלבות שימוש בכלים וחשיבה, אם המודל הציג כישורי חשיבה חלשים, זה פגע משמעותית בביצועים גם כאשר יכולת המודל להשתמש בכלים הייתה מיומנת. אפקט זה נצפה ללא קשר למורכבות או לאופי המשימה.

אפקט "חוק החוליה החלשה ביותר" נשמר ללא קשר לאיזה מעריך מבוסס LLM שימש. בין אם GPT-4o או Claude 3.5 Sonnet שימשו כשופטים, התוצאות באופן עקבי התקבצו ליד היכולת האינדיבידואלית החלשה יותר. עקביות זו מחזקת את חוסנם של ממצאי המדד ומרמזת שהמגבלות הנוכחיות של LLM הן מבניות עמוקות ולא ספציפיות למתודולוגיות הערכה.

חסרונות בשימוש בכלים:

שימוש בכלים התגלה כיכולת החלשה ביותר בכל ה-LLM שנבדקו. משימות הדורשות גלישה באינטרנט, אחזור נתונים דינמי, או הרצת קוד חיצוני הוכחו כמאתגרות במיוחד. הציונים הגבוהים ביותר למשימות הכוללות שימוש בכלים מעולם לא עלו על 50 בסולם של 1-100 לאורך המדד. באופן בולט, אפילו מודלים עם פונקציונליות מפרש קוד, כמו Gemini Pro Exp, התקשו לשמור על ביצועים שווים למשימות חשיבה פשוטות יותר.

חולשה זו קריטית מכיוון ששימוש בכלים הוא יסודי ליישומים רבים בעולם האמיתי, כגון סיוע במחקר, ניתוח נתונים, וסוכני AI. המחברים מדגישים שמודלים המסתמכים אך ורק על מקורות נתונים סטטיים ביצעו באופן גרוע בהשוואה למשימות שבהן מידע מפורש יותר היה זמין ישירות בתוך ההנחיה.

פער ביצועים ביכולות צולבות:

בממוצע, מודלים השיגו 65.72 למשימות יכולת בודדות אך רק 58.67 למשימות יכולת צולבות, פער של 7.05 נקודות. זה מדגיש את הקושי שמודלים נתקלים בו בעת שילוב מיומנויות מרובות. משימות ״תרגום מספרדית וחשיבה״ ו״הקשר ארוך (long context) וקידוד״ הדגימו פערים גדולים במיוחד, המרמזים שנדרש אופטימיזציה נוספת בתרחישי עיבוד רב-לשוני והקשר ארוך.

יעילות CrossEval בהבחנה:

CrossEval הוכח כיעיל בהבחנה בין הבדלים עדינים אפילו בין LLM מתקדמים ביותר. למשל, מודל Claude 3.5 Sonnet עקב בעקביות על קודמיו (המודלים הקודמים של אנטרופיק) במשימות הכוללות זיהוי תמונות וחשיבה וספרדית וזיהוי תמונות. התקדמות זו משקפת את ההתפתחות של מודלי Claude מתוחכמים יותר ומדגישה את הערך של CrossEval במדידת השיפורים העדינים ביכולות LLM.

שיפור מדדי קורלציה:

המדד הדגים שיפור במדדי קורלציה להערכות מבוססות LLM במקרה שמספקים ל-LLM המבצע אבלואציה דוגמאות מתויגות. קורלציית פירסון השתפרה מ-0.578 ללא דוגמאות מתויגות ל-0.697 עם שתי דוגמאות, המצביע על כך שהכללת התייחסויות מתויגות היטב שיפרה משמעותית את אמינות ההערכה.

סיכום:

הניסויים מגלים שבעוד ש-LLM משתפרים במהירות, הם נשארים מוגבלים מאוד על ידי הרכיבים החלשים ביותר שלהם. טיפול במגבלות אלו חיוני להשגת מערכות AI חסונות יותר, רב-תפקודיות המסוגלות לפתור בעיות מורכבות מהעולם האמיתי.

https://arxiv.org/abs/2409.19951