המאמר היומי של מייק: 30.07.25

Forget What You Know about LLMs Evaluations - LLMs are Like a Chameleon

שכחו כל מה שחשבתם על הערכת LLMs –מודלי שפה גדולים הם כמו זיקית (לפחות נכון ללפני 5 חודשים על ידי חוקרים ישראלים).

אנו התרגלנו למדוד התקדמות בבינה מלאכותית דרך המספרים בטבלאות הדירוג. אבל המאמר שנסקור היום מציע תזה מטרידה: ייתכן (אני נוטה להאמין להם) שהציונים המרשימים של המודלים הם לא עדות להבנה אמיתית, אלא להסוואה מושלמת. המודלים המובילים שלנו אולי לא "מבינים"(בלי להיכנס עמוק להגדרה המדויקת לכך), אלא פשוט לומדים לחקות בצורה יפה(זוכרים תוכים סטוכסטיים) את התבניות השטחיות של מבחני ההערכה.

התובנה המרכזית של המאמר: LLMs, שמצטיינים בהתאמת סדרות מילים, יכולים להגיע לביצועים גבוהים בשתי דרכים שונות מאוד, או דרך הבנה אמיתית, או דרך חיקוי סגנוני מתוחכם. החיקוי הזה הוא צורה מסוימת של (overfitting), שבה המודל לא באמת מבין את התוכן, אלא את ה"מרקם הסטטיסטי" של המבחן עצמו. הוא לומד לזהות "שאלה בסגנון MMLU", בלי להבין באמת היסטוריה או פיזיקה.

כדי להבחין בין שתי הדרכים האלה, החוקרים פיתחו כלי חדש: C-BOD) Chameleon Benchmark Overfit Detector). זה לא עוד בנצ'מרק, אלא משהו רובסטי יותר. הפיצ'ר העיקרי שלו הוא בגיאומטריה הלשונית שהוא מייצר: הוא לוקח שאלה קיימת, ומשנה את הניסוח, המבנה והסגנון שלה אבל משאיר בדיוק את אותה המשמעות. הוא זז במרחב השפה לאורך וקטור שהוא אורתוגונלי למשמעות. שאלה שמנוסחת אחרת, מילות אחרות אבל בדיוק אותה כוונה.

המרחק הסגנוני הזה נשלט על ידי פרמטר עיוות (μ), והתוצאה על ידי שינוי בביצועים (Δμ) והיא לא רק ירידה בציון התשובה, אלא מדד של ״שיפוע״ של הידע של המודל. אם הידע יציב ואמיתי, אין בעיה לשנות ניסוח. אבל אם מדובר בזיקית, שינוי קל בסגנון, והביצועים מתרסקים. זוהי חתימה מובהקת של התאמת-יתר.

כשבדקו כך 26 מודלים מובילים: התוצאות היו מדאיגות:

שבריריות כסטנדרט: רוב המודלים, ובעיקר אלו שבטופ של טבלאות הדירוג, חיים על "פסגות מחודדות". הציונים הגבוהים שלהם תלויים ישירות בניסוח המדויק של שאלות ההערכה, מה שמרמז על התאמת-יתר ל-benchmark.

קללת הגודל: דווקא המודלים הגדולים יותר היו שבירים במיוחד. לא רק שהם "חכמים יותר", אלא יש להם מספיק פרמטרים כדי לזכור תבניות ברמת דיוק קיצונית מה שמייצר גבולות החלטה חדים אך שבריריים.

אנומליית LLaMA (הם בדקו llama3): משפחת המודלים של Meta הציגה עמידות גבוהה יותר – מישור ביצועים חלק יותר. הסיבות לא ברורות, אך ייתכן שמדובר בסט נתונים מגוון יותר, או בשיטת אימון שמעודדת הכללה אמיתית ולא שינון.

החידוש האמיתי של המאמר איננו רק בכלי החדש, אלא בתפיסת ההערכה שהוא מציע: הוא קורא לנו לנטוש את הגישה הסטטית של "מה הציון?" ולעבור לשאלה הדינמית: "עד כמה יציב הידע של המודל?". זה מעבר ממכניקה קלאסית ל"מכניקה סטטיסטית" של הערכת אינטיליגנציה מלאכותית.

לדעתי C-BOD הוא קריאה לפתח סט עקרונות וכלים להבנת הדינמיקה הפנימית, כשלי הידע ונוף ההבנה של מודלים מורכבים.

https://arxiv.org/abs/2502.07445