המאמר היומי של מייק: 02.09.25

A Survey on Large Language Model Benchmarks

סקירה מס' 499:

נתחיל מהעובדה שאנו פשוט מוצפים בבנצ'מרקים שמטרתם לאמוד את ביצועי המודלים שלנו. כל מודל חדש מגיע עם סט מבחנים חדש כדי להוכיח את יכולותיו, מה שיוצר מצב קרוב לכאוטי שבו קשה לדעת מהי באמת פריצת דרך מה סתם cherry-picking. הדבר מקשה מאוד על השוואת מודלים ומעקב אחר התקדמות אמיתית, במיוחד כאשר מערכות אלו נפרסות בתחומים בעלי סיכון גבוה כמו רפואה ופיננסים.

סקירה זו מכניסה טיפה של סדר לבלבול הזה. על ידי טקסונומיה שיטתית של 283 (!!) בנצ'מרקים, המאמר מספק את המיפוי הראשון של התחום כולו (ככה נטען שם). החידוש המרכזי שלו הוא מערכת אינטואיטיבית לסיווג מבחנים אלו, שעוזרת לנו להבין את העבר, ההווה והעתיד של אופן המדידה של AI. שפה משותפת זו חיונית לחוקרים כדי לזהות פערים ולבנות הערכות טובות ומשמעותיות יותר.

#### התרומה הגדולה ביותר של המאמר היא מיון כל מבחני הביצועים של LLMs ל-3 קטגוריות ברורות, החל מיכולות בסיסיות ועד למשימות מתמחות בעלות סיכון גבוה.

1.מבחני ביצועים ליכולות כלליות (General Capabilities): אלו הם המבחנים הבסיסיים לכל מודל שפה, המכסים את יכולות הליבה שלו בבלשנות, ידע והסקת מסקנות. המאמר מראה כיצד אלו התפתחו ממבחנים מוקדמים כמו GLUE, שנועדו לאחד את ההערכה , למבחנים אדברסריים קשוחים יותר שנועדו לחשוף הסתמכות של מודלים על "רמזים סטטיסטיים מטעים" במקום על הבנה אמיתית. כעת, התחום מתקדם לעבר "בנצ'מרקים חיים" כמו HELM, שמתעדכנים כל הזמן כדי להישאר צעד אחד לפני היכולות הגדלות של המודלים.

2.מבחני ביצועים לתחומים ספציפיים (Domain-Specific): קטגוריה זו עוקבת אחר התפתחות מודלי השפה מכלים כלליים למומחים בתחומים כמו מדע, משפטים והנדסה. הסקר מראה כיצד מבחני הביצועים חייבים להתאים את עצמם לכל תחום. בהנדסה, למשל, המבחנים עברו מיצירת קוד פשוטה ברמת הפונקציה (HumanEval) לבעיות מציאותיות ברמת המערכת, שמקורן בבעיות אמיתיות מ- (GitHub (SWE-bench. במשפטים, בנצ'מרקים כמו LawBench משתמשים כעת במסגרות חינוכיות מוכרות כמו הטקסונומיה של בלום כדי להעריך רמות שונות של חשיבה משפטית.

3.מבחני ביצועים ממוקדי-מטרה (Target-Specific): זוהי הקטגוריה הצופה פני עתיד והחשובה ביותר, המתמקדת לא במה שהמודל יודע, אלא באיך שהוא מתנהג. היא מכסה את שני התחומים שיגדירו את הדור הבא של AI:

סיכון ואמינות (Risk & Reliability): אזור זה מתמודד עם הבעיות הגדולות ביותר של מודלי שפה, כמו המצאת דברים (הזיות), הפגנת הטיה והדלפת נתונים פרטיים. הסקר מפרט את המרוץ המתמשך בין טכניקות "פריצה" (jailbreak) – שבהן משתמשים מרמים בעדינות את המודל כדי לעקוף את כללי הבטיחות שלו – לבין מבחני בטיחות חדשים המשתמשים בצוותי תקיפה אוטומטיים (red-teaming) כדי למצוא נקודות תורפה.

סוכנים (Agents): זהו הגבול החדש, שבו מודלי שפה פועלים כמערכות אוטונומיות שיכולות לתכנן, להשתמש בכלים ולקיים אינטראקציה עם תוכנות כדי להשיג מטרות. המאמר מארגן את מבחני הביצועים המתקדמים הללו לפי מה שהם מודדים: יכולות ספציפיות כמו שימוש בכלים, ביצועים כוללים במשימות מורכבות, מומחיות בתחום מקצועי, ובטיחות בתרחישים מסוכנים.

#### יותר ממפה: מבט מפוכח על מה שלא עובד

המאמר מספק גם ביקורת נוקבת ומפוכחת על הבעיות המרכזיות באופן שבו אנו מעריכים כיום מודלי שפה. הוא חורג מעבר לרשימת בנצ'מרקים ומאבחן את הפגמים שמערערים את אמוננו בתוצאותיהם.

זיהום דאטה (Data Contamination): קיים סיכון עצום שהמודלים אומנו על שאלות המבחן, מה שמוביל ל"תוצאות הערכה מנופחות" שאינן משקפות את מה שהמודל באמת יכול לעשות בעצמו. המאמר מדגיש את החשיבות של יצירת "בנצ'מרקים דינמיים ועמידים בפני זיהום" המשתמשים בנתונים חדשים או פרטיים כדי להבטיח מבחן הוגן.

הטיה תרבותית ולשונית (Cultural and Linguistic Bias): רוב מבחני הביצועים מתמקדים באנגלית, מה שאומר שהם אינם מעריכים באופן הוגן מודלים בשפות עם מבנים והקשרים תרבותיים שונים. "המיקוד האנגלוצנטרי" הזה עלול להסתיר ביצועים נמוכים ולהוביל לתמונה מעוותת של יכולותיו האמיתיות של המודל ברחבי העולם.

התעלמות מה"איך" ומהעולם האמיתי (Ignoring the "How" and the Real World): המאמר מצביע על נקודה עיוורת מרכזית: אכפת לנו בעיקר מהתשובה הסופית ואנחנו מתעלמים מאיך המודל הגיע אליה. התמקדות זו במדד דיוק יחיד נכשלת ב"תיאור מקיף של היכולות המורכבות של מודלי שפה גדולים" ועלולה להסתיר חשיבה פגומה. יתרה מכך, רוב המבחנים הם סטטיים ואינם משקפים את הטבע הדינמי והמשתנה של העולם האמיתי, שבו מודלים צריכים להסתגל.

על ידי ארגון מאות מבחני ביצועים למסגרת אחת, מובנת, והדגשת האתגרים הקריטיים העומדים בפנינו, "סקירה על מבחני ביצועים של מודלי שפה גדולים" הוא יותר מסקירה פשוטה – הוא מדריך חיוני. הוא מעצים מפתחים, חוקרים ומובילי תעשייה לחרוג מדירוגים פשטניים ולשאול שאלות עמוקות יותר. ערכו הסופי הוא בסיוע להסיט את השיח מסתם "מה מודלים יכולים לעשות" לשאלה החשובה הרבה יותר של "כיצד עליהם לפעול באחריות".

https://arxiv.org/abs/2508.15361