**המודל הכי טוב בבנצ'מארקים - האם זו הסיבה להתלהבות?**

**המאמר היומי של שמוליק ומייק: 12.10.25, סקירה 524**

**Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence**

בשנים האחרונות, עם העלייה הדרמטית בשימוש ב-LLMs, התפתח דור חדש של בנצ'מרקים: מבחנים שנועדו להעריך ולהשוות בין מודלים במגוון משימות. בניגוד לבנצ'מרקים הקלאסיים מעולם החומרה והתוכנה, שהיו לרוב סטנדרטיים וחד־ממדיים, המבחנים לעולם ה־GenAI מתאפיינים בגיוון רב מבחינת גישות ומתודולוגיות. חלקם מתמקדים בדיוק (accuracy) או יעילות (efficiency), אחרים ביצירתיות (creativity), בריזונינג או אפילו בהיבטים אתיים כמו הטיה ו־fairness.

גיוון זה מעיד מצד אחד על ההתפתחות המואצת של התחום, ומצד שני מעלה שאלות לגבי הצורך במסגרת אחידה שתאפשר השוואה הוגנת ורב־ממדית בין מודלים שונים. המאמר ניגש לנושא מזווית מאוד ביקורתית וטוען שיש ליקויים רבים בבנצ'מרקים הקיימים כיום וחייבים לעשות סדר ולפתח מתודולוגיה ברורה כדי שנוכל לסמוך עליהם.

המאמר בוחן 23 בנצ'מרקים מובילים - בהם MMLU (Massive Multitask Language Understanding - מבחן ידע כללי בסגנון בחינות סטנדרטיות), HumanEval (הערכת יכולות קוד ב־Python), LegalBench (משפט), MultiMedQA (רפואה), SciBench (בעיות מדעיות), BIG-bench (מבחנים כלליים שנוצרו ע"י מאות חוקרים), AGIEval (בחינות אמיתיות שמודלים נבחנים בהן כמו SAT), וגם בנצ'מרקים חדשים יחסית כמו AgentBench ו־ToolBench שבודקים אינטראקציה עם כלים חיצוניים.

ליבת הטיעון: Functionality ו־Integrity

הכותבים טוענים שכל בנצ'מרק טוב צריך להיבחן בשני מימדים:

1. Functionality - האם המבחן באמת מודד יכולות שימושיות ורלוונטיות לעולם האמיתי?
   * למשל, מבחן כמו MMLU יכול לגלות אם המודל יודע לענות על שאלות רב־ברירה, אבל זה לא בהכרח מלמד על היכולת שלו לשוחח, לעזור למתכנת או לכתוב תוכן יצירתי.
2. Integrity - האם המבחן חסין ממניפולציה ורמאויות?
   * כאן עולה בעיית האוברפיט: מודל גדול שלמד את כל האינטרנט עשוי פשוט לשנן את התשובות המופיעות במבחן (במודע או לא), מבלי לפתח יכולת הכללה אמיתית.

בעיות מרכזיות שנמצאו:

המחקר מצביע על שורה ארוכה של ליקויים נפוצים:

* Response variability״: התשובות של המודל רגישות בצורה קיצונית לשינויים קטנים. במאמר מביאים דוגמה: מעבר מפורמט “[A]” ל־“(A)” יכול להוביל לציון שונה לחלוטין, מה שיכול להעיד על השבריריות של המבחן.
* Single correct answer assumption: רוב ה־בנצ'מרקים מניחים שיש רק תשובה אחת נכונה. זה אולי עובד במתמטיקה, אבל בשאלות על דת, תרבות או אתיקה - התמונה מורכבת הרבה יותר. החוקרים מציינים שזה עיוות רציני שמקשה על הערכה אמינה.
* Ignoring conversational context: אינטראקציה עם LLM לרוב מתרחשת בשיחה מתמשכת, עם שאלות המשך, הבהרות ותיקונים. רוב ה־בנצ'מרקים מסתפקים בתשובה ראשונה בלבד.
* Helpfulness vs Harmlessness tension: המאמר מתעכב על הפער בין להיות מועיל (Helpful) לבין להיות בטוח ולא מסוכן (Harmless). למשל, במבחנים רפואיים כמו MultiMedQA, מודל יכול "להתחמק" מתשובה בשם בטיחות - ולקבל ציון גבוה, למרות שלא נתן תשובה מועילה למשתמש.
* הטיות בבנצ'מרקים שנבנו על ידי LLMs: חלק מהמבחנים עצמם נוצרו ע"י LLMs. למשל, ב־M3KE השתמשו במודל כדי ליצור שאלות חדשות. זה מוביל לבעיה של “LLM judging LLM”, שבה המודל גם מייצר את המבחן וגם נבחן עליו.
* Scalability and reproducibility issues: חלק מהמבחנים מורכבים להרצה ולא תמיד מתועדים היטב. ב־SciBench למשל, התקנת סביבת ההרצה מסובכת, מה שמקשה על הרצה מחדש ובחינה חוזרת שלו.
* Peer review: מתוך 23 בנצ'מרקים, רק 6 עברו ביקורת עמיתים. כלומר, רובם הופצו בלי תהליך אקדמי מלא, וזה מעלה שאלות על אמינותם.

מה מציעים החוקרים?

החוקרים לא מסתפקים רק בביקורת, אלא מציעים גם כיוון לפתרון מעולמות הCyber Security - PPT Framework (People, Process, Technology):

* People: מי בונה את המבחן ומי משתמש בו? האם יש מספיק גיוון בין מומחים שונים (למשל רופאים, משפטנים, מתכנתים)?
* Process: האם התהליך עקבי, מתועד וסטנדרטי, אך גם גמיש להתפתחות מהירה של המודלים?
* Technology: האם התשתיות (Datasets, APIs, מערכות הערכה) אמינות וסקיילביליות?

בנוסף, הם מדגישים את הצורך לעבור מבנצ'מרקים סטטיים לדינמיים כלומר, מבחנים דינמיים שמתעדכנים ומשתנים לאורך זמן, עם בדיקות תקופתיות שמזהות יכולות חדשות או התדרדרות בהתנהגות המודל.

ביקורת על המאמר

למרות התרומה המשמעותית שלו, המאמר סובל מכמה חולשות:

* הפתרונות המוצעים נשארים high-level. ה־PPT Framework נשמע טוב כתיאוריה, אבל אין בו פירוט מספק איך לבנות בפועל בנצ'מרק בעזרתו.
* החוקרים לא הריצו את הבנצ'מרקים בעצמם, אלא הסתפקו בסקירה. מהלך אמפירי היה עשוי לגלות בעיות נוספות.
* חלק מהבעיות שהחוקרים מציינים הם אינהרנטיות לבנצ'מרקים ואין להם ממש פתרון פשוט, לא תמיד ברור מה החוקרים ציפו שיוצרי הבנצ'מרק יעשו אחרת כדי לפתור את הבעיה.
* המאמר מתמקד בעיקר בבנצ'מרקים ותיקים יחסית כמו HumanEval או MMLU, אך לא סוקר חדשים יותר כמו Swe-Bench (שמודד תיקוני באגים בקוד) או Aider Benchmark, שכבר מנסים לטפל בחלק מהכשלים שהוזכרו.
* יש דגש גדול על פערים תרבותיים ואתיים שהוא אמנם חשוב, אבל לא מאוד רלוונטי לחלק גדול מהבנצ'מרקים, כמו אלה שמודדים יכולות כתיבת קוד.

סיכום

זהו מאמר ביקורתי ומרחיב אופקים שמספק מיפוי שיטתי של הכשלים בעולם ה־LLM בנצ'מרקים. הוא נותן שפה מסודרת לבעיות שמוכרות להרבה חוקרים, כמו Response variability, Overfitting ו־Lack of reproducibility, ומציע לחשוב מחדש על איך נבנים מבחנים: לא רק סטטיים, אלא דינמיים, רב־ממדיים ומבוססי מתודולוגיה סדורה.

החיסרון**:** הוא תורם רבות בניסוח הבעיות המרכזיות בתחום ובהצגת כיווני מחשבה חדשים, אך חסר כלים פרקטיים והצעות פרקטיות ליישום. לא ברור אילו מהבעיות שמציגים במאמר הן ברות פתרון ואלו אינהרנטיות לבנצ'מרקים של LLMs.

https://arxiv.org/abs/2402.09880