המאמר היומי של מייק: 14.08.25  
FormulaOne: Measuring the Depth of Algorithmic Reasoning Beyond Competitive Programming

מאמר 🇮🇱 לבן של כמה מחברים שאני מכיר באופן אישי….

במרוץ הבלתי פוסק לעבר בינה מלאכותית כללית (AGI), היכולת של LLMs לחשוב בצורה אלגוריתמית נותרה חזית קריטית ושנויה במחלוקת. במשך שנים, מדד הביצועים (הבנצ'מארק) העיקרי שלנו היה תכנות תחרותי – תחום פופולרי מאוד ששימש כאינדיקציה לא רעה לסוג מסוים של חשיבה חישובית. אבל ככל שהמודלים שלנו הופכים מתוחכמים יותר, עולה שאלה מטרידה: האם אנחנו עדיין מודדים את הדבר הנכון? אני מדברים הרבה על בנצ'מארקים לאחרונה בצורה ביקורתית ולכן החלטתי לסקור את המאמר.

מאמר חדש וכחול לבן למעשה קורא לשינוי פרדיגמה. החוקרים מציגים בנצ'מארק חדש שנועד לבחון את עומק החשיבה האלגוריתמית, מימד שלטענתם נעלם כמעט לחלוטין ממסגרות ההערכה הקיימות. בעוד שמודלי שפה מראים תוצאות מרשימות על בנצ'מארקים כמו תכנות תחרותי, הם בעיקר פותרים בעיות שניתן לפתור על ידי שילוב של מספר אלגוריתמים מוכרים. אנחנו מתעניינים ביכולת לפתור בעיות הדורשות תהליך חשיבה עמוק ויצירתי יותר.

וזו בדיוק הנקודה. הבנצ'מארקים הנוכחיים בודקים את יכולתו של מודל לגשת וליישם את ספריית הפתרונות המוכרים העצומה שלו. FormulaOne שואל שאלה עמוקה יותר: האם מודל יכול לחשוב כמו מדען מחשב?

### מעבר לאזור הנוחות של התכנות התחרותי

פלטפורמות כמו Codeforces ו-LeetCode היו בעלות ערך אדיר. הן דחפו את גבולות היכולת של המודלים. עם זאת, כפי שמציין שלו-שוורץ, הן מטפחות סוג מאוד מסוים של פתרון בעיות; כזה המבוסס על זיהוי תבניות (pattern recognition) ורקומיבנציה. המאמר על FormulaOne מותח ביקורת מרומזת על הפרדיגמה הזו ומצביע על מגבלותיה:

מיקוד במהירות: תכנות תחרותי מתגמל לרוב את הפתרון הנכון המהיר ביותר, לא בהכרח את האלגנטי ביותר או זה שניתן להכללה.

חשיבה שטחית: בעיות רבות הן וריאציות על נושא מוכר, הניתנות לפתרון על ידי זיהוי תבנית ויישום אלגוריתם סטנדרטי. זה בוחן את "אוצר המילים האלגוריתמי" של המודל, לא את יכולת החשיבה שלו.

"קביים" של דאטה האימון: קיימת סבירות גבוהה שפתרונות לבעיות פופולריות רבות מסתתרים אי שם בתוך נתוני האימון (training data) של המודל, מה שמקשה על הערכת יכולת פתרון בעיות אמיתית ומקורית.

אתגר ה-FormulaOne: סוג חדש של בנצ'מארק

כאן FormulaOne נכנס לתמונה. זה לא רק דאטהסט חדש; זו פילוסופיית הערכה חדשה המטרה היא למדוד את עומק החשיבה הנדרש כדי להמציא אלגוריתם חדש מאפס.

החוקרים משיגים זאת באמצעות גישה "מתמטית" ומתוחכמת, הממנפת מושגים מתחום הסיבוכיות הפרמטרית (parameterized complexity) ותורת הגרפים (graph theory) כדי לייצר בעיות עם שיפוע קושי מבוקר ומדויק. אחד הכלים המתמטיים המרכזיים שהם משתמשים בו הוא רוחב-עץ (treewidth) של הפתרון, מדד לכמה גרף הוא "דמוי-עץ". בעיות עם treewidth נמוך ניתנות לרוב לפתרון באמצעות תכנות דינאמי, אך ככל שה-treewidth עולה, היצירתיות האלגוריתמית הנדרשת נוסקת.

זה מאפשר להם ליצור בעיות שנראות פשוטות באופן מטעה על פני השטח, אך דורשות תובנה עמוקה ולא מובנת מאליה. הם מכוונים לבעיות שבהן "הפתרון הוא תוכנית פשוטה לכתיבה, אך הדרך לגילוי התוכנית הזו היא מורכבת ומסועפת".

כדי להגדיר זאת באופן פורמלי, הצוות משתמש ב-לוגיקה מסדר שני מונודית (MSO). זוהי מסגרת לוגית חזקה המאפשרת להם להגדיר תכונות של גרפים ולייצר באופן אוטומטי סט עצום ומגוון של בעיות. באופן קריטי, תהליך יצירת הדאטה הסינתטי הזה מבטיח שהבעיות הן חדשות ואינן מופיעות בשום דאטה אימון, מה שמאלץ את המודלים לחשוב מהעקרונות הראשונים.

### התוצאות המפכחות והדרך קדימה

ממצאי המאמר הם קריאת השכמה. בעוד שמודלים החזקים ביותר כרגע כמו GPT-4 ו-Claude 3 Opus (סליחה gpt5) מראים יכולות מסוימות, ביצועיהם על בעיות FormulaOne נמוכים משמעותית מאשר על בנצ'מארקים מסורתיים. זה מדגים באופן חד את הפער בין התאמת תבניות לחשיבה עמוקה ואמיתית. המודלים מתקשים בדיוק בנקודה שבה נדרש גילוי אלגוריתמי יצירתי ורב-שלבי.

זוהי התובנה החדה וההוליסטית מ-FormulaOne. זה לא עוד Leaderboard שצריך לטפס בו; זהו כלי אבחוני שחושף את המגבלות הנוכחיות של ה-LLMs שלנו. המאמר מציע שפשוט להגדיל ארכיטקטורות קיימות ודאטה אימון אולי לא יספיק כדי לגשר על התהום ל-AGI. עלינו להתמקד בארכיטקטורות ובשיטות אימון המטפחות פתרון בעיות יצירתי ואמיתי.

FormulaOne מספק נתיב קונקרטי ומבוסס מתמטית למדוד את ההתקדמות שלנו. הוא מאתגר את קהילת ה-AI לצאת מאזור הנוחות של בעיות מוכרות ולהתחיל להתמודד עם האתגר הקשה הרבה יותר, והחשוב הרבה יותר, של ללמד את המודלים שלנו איך לחשוב. המירוץ החל.

וכמו שאמרתי כבר לכמה אנשים לדעתי העתיד הוא לא מודלים חכמים בצורה מטורפת אלא המודלים שיודעים להפעיל כלים בצורה מטורפת

https://arxiv.org/abs/2507.13337

מעבר למילים: למה Large Action Models הם הצעד האמיתי אל AI שפועל בעולם