ירושלים, כ״ח טבת תשפ״ה
28 ינואר 2025

השקת מודל DeepSeek R1 הסיני מערערת את פרדיגמת פיתוח ה-AI במערב ומאיימת לשנות את המאזן בתחרות הגאו-טכנולוגית בין ארה״ב לסין

השקת מודל היסוד הסיני המתקדם DeepSeek R1 מהווה התפתחות משמעותית בעולם ה-AI - אימון המודל בעלויות נמוכות ועל תשתית חומרה פחות מתקדמת מהווה הפתעה אסטרטגית (״DeepSeek Moment״), **ומטלטלת את הפרדיגמות שהנחו את המערב במרוץ הטכנולוגי לביסוס עליונות בתחום הבינה המלאכותית:**

ברמה הטכנולוגית, המודל הסיני מערער לכאורה את המוסכמה כי התקדמות חייבת לעבור דרך הגדלה מאסיבית של כוח המחשוב בהשקעות עתק במרכזי נתונים עתירי שבבים מתקדמים.

ברמה האסטרטגית, הפריצה הסינית למודל יעיל וזול במשאבים שומטת את הקרקע מתחת לאסטרטגיה שסברה כי הגבלת יצוא שבבים מתקדמים לסין תאט את התקדמותה הטכנולוגית.

ברמת התחרות הטכנולוגית, סדרת מודלים סיניים שנחשפו במקביל בימים האחרונים מציבה סימן שאלה על ההנחה הרווחת כי ארה״ב מובילה בבטחה במודלי היסוד האיכותיים. החשיפה של חדשנות סינית מבוססת קהילת מפתחים מקומית בתחום ה-AI מעידה על פוטנציאל לקחת ההובלה העולמית בתחום.

ברמה הכלכלית, שחרור המודל הסיני כקוד פתוח מאפשר שימוש נרחב והתאמות קוד לכל מי שחפץ, מקזז יתרון מסחרי של החברות האמריקניות ומחייב אותן להיערך אחרת באשר למודל הרווחים בטווח הקצר לפחות.

עבור ישראל, המודל הסיני מסמן אופק חיובי באשר לסיכוייה להיות מהמובילות בבינה מלאכותית בעולם גם ללא השקעות עתק, שהיוו עד עתה חסם כניסה משמעותי למועדון זה באופן עצמאי.

Ministry of Innovation, Science & Technology
JTG, 2 Agudat Sport Hapoel Rd., Jerusalem, 9695102
Tel: +972 (2) 5411109, Fax: +972 (2) 5825581

משרד החדשנות, המדע והטכנולוגיה
גן המדע, דרך אגודת ספורט הפועל 2, ירושלים, 9695102
טלפון: 02-5411109, פקס: 02-5825581

1. חברת סטארט-אפ סינית בבעלות איש עסקים מקומי Liang Wenfeng, הפתיעה את העולם בהשקת מודל השפה הגדול החדש שלה DeepSeek R1. מדובר במודל גנרטיבי העומד בשורה אחת עם המודלים האמריקנים המתקדמים (671 מיליארד פרמטרים)[1], בעל יכולת הסקה, שהשיג ציונים מפתיעים במדדי איכות מובילים הקרובים לביצועי מודל GPT-4o1, הנחשב למתקדם ביותר בין המודלים שבשימוש (השוואה בנספח). יצוין כי במקביל שחררה אותה החברה מודל נוסף מתוצרתה - Janus-Pro-7B המולטימודלי[2], שלטענתה עולה על ביצועי DALL-E3 (גם הוא מודל אמריקני מבית OpenAi). כמו כן, הסינים שחררו מודל גדול נוסף של Kimi.ai[3], שגם לגביו נטען כי משתווה לביצועי המודלים האמריקנים. **המודלים הסינים שנחשפו מהווים עדות אפשרית לסגירת הפער במודלי היסוד ל-AI תוך גילוי גישה חדשנית וידע טכנולוגי נרחב .**

2. החברה הסינית טוענת כי עלויות האימון של מודל DeepSeek R1 היו זולות במיוחד - פחות מ-6 מיליון דולר בלבד, וכי האימון נעשה על בסיס תשתית צנועה של כ-2000 מאיצי AI בעלי יכולת מופחתת[4] וארך 55 ימים בלבד. לשם ההשוואה: עלות האימון של המודל של OpenAI הייתה מעל מאה מיליון דולר, והוא נעשה על תשתית של 25 אלף מאיצים מתקדמים יותר (H100) במשך כמאה ימים. זאת ועוד, החברה טוענת כי השימוש השוטף במודל שלה צפוי להיות זול בכ-90% בהשוואה למתחרה האמריקנית. כל זאת כאמור כשאיכותו לא נופלת מהמודלים האמריקנים המתקדמים. התייעלות זו מתאפשרת הודות לשורה של חידושים אלגוריתמיים בשלבים שונים של אימון והפעלת המודל.

3. **המודל הסיני החדש מערער על הפרדיגמה המרכזית לפיה התנהלה תעשיית ה-AI העולמית בשנתיים האחרונות, הגורסת כי שיפור ביצועים במודלי יסוד עובר בהכרח דרך הגדלת כוח המחשוב.** הגדלת כוח מחשוב מחייבת השקעה חסרת תקדים בהקמת מרכזי נתונים (Data Centers) ייעודיים. תפיסה זו השתקפה אך ימים ספורים קודם להשקת המודל בהכרזת הנשיא טראמפ על מיזם Stargate - **השקעה בהיקף של כחצי טריליון דולר בתשתיות חישוב לבינה מלאכותית בארבע השנים הקרובות**[5]. ואכן, חשיפת המודל, וההבנות הראשונות באשר לחדשנות שביצירתו, התקבלו בהפתעה והובילו

[1] להשוואה, המודל המקביל של OpenAI כולל 1.76 טריליון פרמטרים, בעוד של META כולל 405 מיליארד פרמטרים.

[2] מודל מולטימודלי מסוגל לצרוך ולייצר סוגי חומר שונים מלבד טקסט - תמונה, קול, מוזיקה וכו'.

[3] בבעלות חלקית של Alibaba.

[4] החברה השתמשה בשבבי GPU מסוג H800 של אנבידיה שמתחת לרף מגבלות הייצוא של שבבי AI מתקדמים לסין.

[5] מיזם Stargate בנוי מקונסורציום חברות הכולל את Oracle (תשתיות) ו-OpenAI (תוכנה) האמריקניות, יחד עם קרן ההון-סיכון היפנית Softbank (מימון) - בחסות ובעידוד הממשל. עוד לפני חשיפת המודל הסיני, עלו ספקות בנוגע להיתכנות להעמיד חצי טריליון לפרויקט. יצוין כי גם Microsoft הכריזה לאחרונה על השקעה מתוכננת של כ-80 מיליארד דולר ב-2025, בעוד Meta הכריזה על השקעה בהיקף של 63 מיליארד דולר בשנת 2025, והשקעות עתק בתשתיות תוכננו גם על ידי Amazon ו-xAI של מאסק.

Ministry of Innovation, Science & Technology
JTG, 2 Agudat Sport Hapoel Rd., Jerusalem, 9695102
Tel: +972 (2) 5411109, Fax: +972 (2) 5825581

משרד החדשנות, המדע והטכנולוגיה
גן המדע, דרך אגודת ספורט הפועל 2, ירושלים, 9695102
טלפון: 02-5411109, פקס: 02-5825581

לצניחת מדדי ענקיות הטכנולוגיה, באופן המשקף את מה שכבר זכה לכינוי DeepSeek Moment של תעשיית הבינה המלאכותית האמריקנית[6]

4. יצוין כי מודל DeepSeek R1 שוחרר בקוד פתוח ונועד לשימוש חופשי של קהילת המפתחים, ובשונה מרוב מהמודלים האמריקנים המקבילים[7], המודל הסיני כולל גם גרסה מוקטנת שניתנת להרצה על מחשבי PC. מאפיינים אלו עשויים להפוך את המודל לשימושי ונפוץ, ולאפשר לחברות רבות ברחבי העולם לשחזר ולשפר את המודל באופן שייתר את הישגי תעשיית ה-AI האמריקנית עד כה, ויפגע באופן אנוש ברווחיה בטווח הקצר לפחות.

5. שחרור המודל שפותח באופן בלעדי על ידי חברה סינית ועל ידי מדענים סינים (הסינים מדגישים כי איש מאנשי החברה לא למד בחו"ל), מערער את הנחת היסוד שהייתה מקובלת במערכת העולמית בדבר עליונות מוחלטת של מודלי היסוד האמריקנים (אחד התחומים הטכנולוגיים-מדעיים האחרונים בהם הסינים נתפסו כמי שמפגרים מאחור). מפתיע במיוחד שהמודל הושק על ידי חברת סטארט-אפ אלמונית ולכאורה דלת משאבים, כאשר ברקע תוכניות להשקתם של מודלים מתקדמים של ענקיות הטכנולוגיה הסיניות[8].

6. ככל שיתבהר כי המודל הסיני אכן מציג פריצת דרך ביכולת לפתח מודלים חזקים בעלויות מזעריות וללא תלות בשבבי העיבוד החזקים ביותר, הרי שחשיפתו מציבה סימן שאלה על האסטרטגיה האמריקנית להובלה טכנולוגית, שהתמקדה בהבטחת השליטה על שרשראות האספקה של שבבים מתקדמים ובהגבלת הנגישות של סין ומדינות נוספות לשבבים אלו. יתרה מכך, אסטרטגיה זו דחפה את הסינים לחדשנות ולמציאת פתרונות עוקפים ויעילים יותר להשגת המובילות ב-AI. יצוין כי כוח מחשוב עודף עודנו מהווה יתרון למי שמחזיק בו, גם בעידן של מודלים יעילים יותר, ולכן החיבור בין חדשנות המודל הסיני - ששוחרר כאמור בקוד פתוח - לעוצמה חישובית האמריקנית, עשויה להוביל לקפיצת מדרגה נוספת בהתפתחות ה-AI העולמי.

7. המודל הסיני פותח את הזירה לתחרות עבור שחקנים נוספים, ובכללם גם ישראל, שמתקשים להתמודד על חסם הכניסה הגבוה בתשתיות להובלה עולמית בזירת ה-AI.

[6] רפרנס ל-Sputnik Moment - ההפתעה של הציבור האמריקני כאשר הסובייטים הקדימו אותם בשיגור לווין ספוטניק לחלל ב-1957.

[7] להוציא מודל LLaMA של META

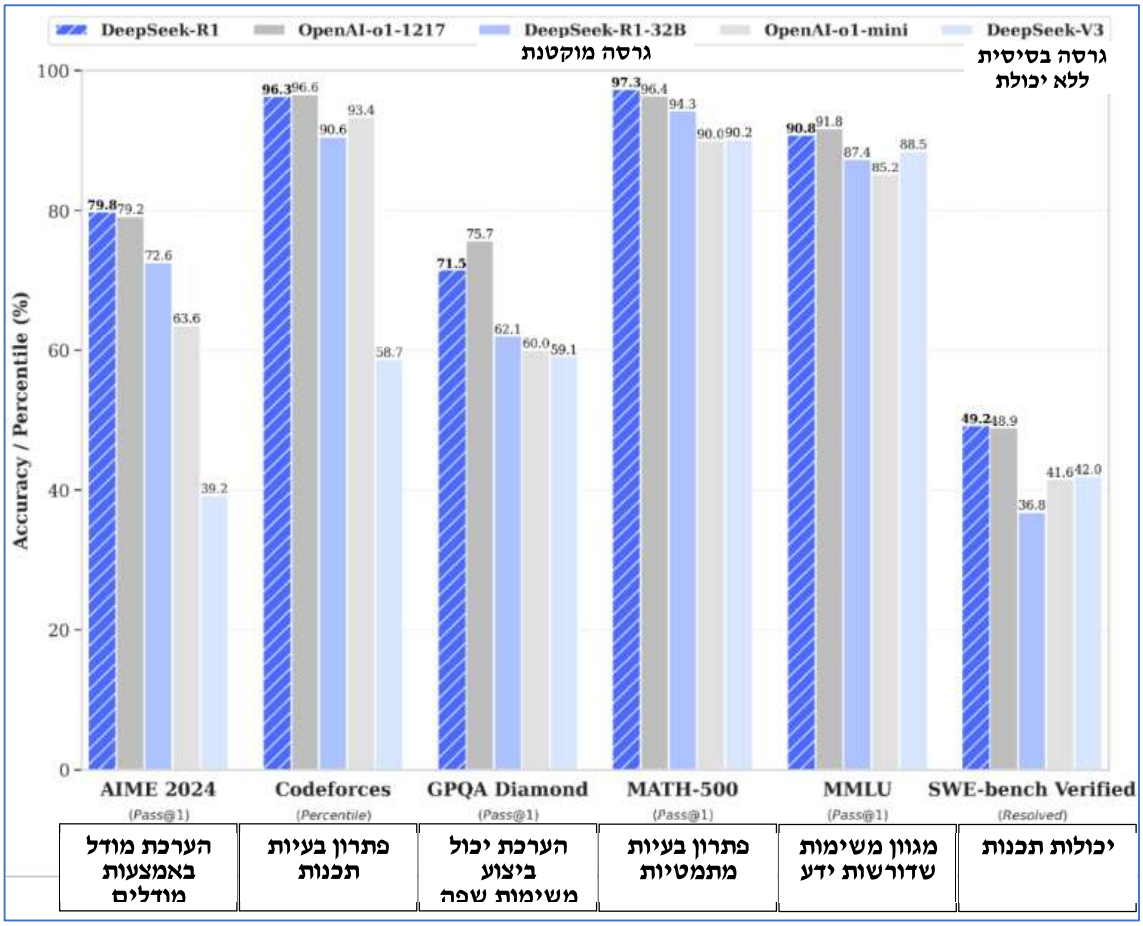
[8] Ernie (Baidu), Doubao7 (ByteDance), PanGu (Huawei), M6 (Alibaba), Wu Dao (Beijing Academy of AI).

Ministry of Innovation, Science & Technology
JTG, 2 Agudat Sport Hapoel Rd., Jerusalem, 9695102
Tel: +972 (2) 5411109, Fax: +972 (2) 5825581

משרד החדשנות, המדע והטכנולוגיה
גן המדע, דרך אגודת ספורט הפועל 2, ירושלים, 9695102
טלפון: 02-5411109, פקס: 02-5825581

נספח

ציונים השוואתיים במבחני איכות של המודל



מקור הגרף: https://github.com/deepseek-ai/DeepSeek-R1

Ministry of Innovation, Science & Technology
JTG, 2 Agudat Sport Hapoel Rd., Jerusalem, 9695102
Tel: +972 (2) 5411109, Fax: +972 (2) 5825581

משרד החדשנות, המדע והטכנולוגיה
גן המדע, דרך אגודת ספורט הפועל 2, ירושלים, 9695102
טלפון: 02-5411109, פקס: 02-5825581