

כלכלת הטוקנים: כיצד יחידת המדידה החדשה של הבינה המלאכותית מעצבת את הכלכלה, התשתיות והריבונות הדיגיטלית

הדס לורבר ועידו בן-דוד¹ | גיליון 2186 | 18 באוגוסט 2026

הבינה המלאכותית אינה משנה רק את יכולות העיבוד של מחשבים, אלא גם יוצרת כלכלה חדשה המבוססת על הטוקן כיחידת המדידה המרכזית של השימוש במודלים, הביקוש לכוח מחשוב והעלות הכלכלית של הפעלתם. הטוקן משקף את היקף עבודת ה-Inference – העיבוד שמבצע המודל בזמן אמת – ולכן מהווה מדד ישיר לצריכת מחשבי מחשוב, ובהם מעבדים גרפיים, מרכזי נתונים ואנרגיה. במובן זה, הטוקן אינו הגורם להשקעות בתשתיות AI אלא המדד שבאמצעותו ניתן להבין את הביקוש לכוח מחשוב – המשאב האסטרטגי המרכזי בעידן הבינה המלאכותית. המאמר מציג את מושג "כלכלת הטוקנים" כמסגרת אנליטית חדשה להבנת התפתחות שוק ה-AI ומראה כיצד הוא מסביר תהליכים כגון השקעות גלובליות במרכזי נתונים, עליית חברות ה-NeoCloud, המרוץ להקמת תשתיות AI לאומיות והתעצבות ריבונות דיגיטלית המבוססת על שליטה בכוח מחשוב. לצד זאת, המאמר מדגיש כי כלכלת הטוקנים משנה גם את אופן הניהול הארגוני של מערכות בינה מלאכותית, המחייב מעבר ממדידת ביצועי מודלים בלבד לניהול יעיל של צריכת טוקנים, עלויות Inference והערך העסקי המופק מכל אינטראקציה. לאור זאת, המאמר טוען כי הבנת כלכלת הטוקנים היא מפתח להבנת המודל העסקי של תעשיית ה-AI ושל התחרות הגלובלית על תשתיות מחשוב וריבונות טכנולוגית.

כלכלת הטוקנים: הרבה יותר מיחידת תמחור

מאז השקת ChatGPT בסוף שנת 2022, התמקד הדיון הציבורי בבינה מלאכותית בעיקר ביכולות המודלים: איכות הכתיבה, יכולת ההסקה, יצירת קוד, תמונות וסרטונים. אולם, מאחורי ההתקדמות הטכנולוגית מתפתחת מהפכה כלכלית עמוקה יותר, הנשענת על מושג שרוב המשתמשים כמעט ואינם מודעים לקיומו – הטוקן.

¹ עידו בן דוד - מנהל שותפויות טכנולוגיות בבניס ואחראי על הפעילות מול רשות החדשנות סביב הקמת מחשב-העל בישראל.

מבחינה טכנית, טוקן הוא יחידת המידע הבסיסית שבאמצעותה מודלי שפה מעבדים מידע. הוא אינו זהה למילה ואינו זהה לתו בודד; טוקן יכול להיות מילה שלמה, חלק ממילה, סימן פיסוק או רצף תווים, בהתאם לשיטת ה-Tokenization שבה משתמש המודל. גם מידע שאינו טקסטואלי – תמונות, קטעי קול, מסמכים או וידאו – מומר בסופו של דבר לייצוגים הניתנים לעיבוד באמצעות יחידות חישוב דומות.

הבדלים אלה אינם רק טכניים. בשפות שונות נדרש מספר שונה של טוקנים כדי לייצג את אותו רעיון. בעברית, למשל, מספר הטוקנים הנדרש עדיין גבוה יחסית בהשוואה לאנגלית. עובדה זו נובעת הן מהמורפולוגיה של השפה והן מכך שמרבית המודלים אומנו בעיקר על טקסטים באנגלית, בעוד שאותיות עבריות מיוצגות במספר גדול יותר של Bytes. הפערים בין השפות הולכים ומצטמצמים ככל שטכנולוגיות ה-Tokenization משתפרות.

עם זאת, הטוקן אינו רק יחידת עיבוד. עבור ספקיות מודלים כמו OpenAI, Google, Anthropic, הוא הפך גם ליחידת התמחור המרכזית של שירותי הבינה המלאכותית. במקום לחייב לפי מספר השאלות שנשאלו, החיוב מתבצע בדרך כלל לפי מספר הטוקנים שהמודל קולט (Input Tokens) ומייצר (Output Tokens). מאחורי כל שאלתה מסתתר לעיתים נפח הרבה יותר גדול של מידע: הוראות מערכת, היסטוריית השיחה, מסמכים שנשלפו ממערכות RAG (מערכות המשלבות אחזור מידע ממאגרי ידע חיצוניים בזמן אמת), תוצאות של כלים חיצוניים והקשר נוסף שמוזרם למודל. כולם מתורגמים בסופו של דבר לטוקנים.

אולם, החשיבות האמיתית של הטוקן אינה נעוצה דווקא בכך שהוא משמש יחידת תמחור. גם אילו היו ספקי המודלים בוחרים לחייב לפי זמן עיבוד, לפי מספר שאלות או לפי צריכת אנרגיה, הצורך בתשתיות מחשוב היה נותר דומה. חשיבות הטוקן נובעת מכך שהוא הפך ליחידת המדידה הטובה ביותר של הביקוש לעיבוד בזמן אמת (Inference). ככל שמעובדים יותר טוקנים, כך נדרש יותר כוח מחשוב; וככל שנדרש יותר כוח מחשוב, כך גדל הצורך במרכזי נתונים, במעבדים גרפיים ובאנרגיה. מנקודת מבט זו, כלכלת הטוקנים אינה רק מודל תמחור. היא דרך חדשה להבין כיצד נוצר ביקוש למשאב האסטרטגי החשוב ביותר בעידן הבינה המלאכותית: כוח מחשוב.

מטוקנים ל-Inference: מדוע הבינה המלאכותית היא כלכלה שונה

כדי להבין את כלכלת הטוקנים, חשוב להבחין בין כלכלת הבינה המלאכותית לבין מודל התוכנה המסורתי (Software as a Service – SaaS). במשך שני העשורים האחרונים התבססה תעשיית התוכנה על הנחה פשוטה: העלות העיקרית היא פיתוח המוצר. לאחר שהמוצר פותח, העלות של כל משתמש נוסף נמוכה יחסית. לכן, רוב שירותי הענן מתומחרים באמצעות מנוי חודשי קבוע, גם כאשר היקף השימוש בפועל משתנה בין משתמש למשתמש.

בינה מלאכותית גנרטיבית פועלת באופן שונה. גם לאחר שהמודל אומן, כל אינטראקציה עם המשתמש מחייבת ביצוע חישובים בזמן אמת. כל שאלה, מסמך, תמונה או קטע קוד שמועברים למודל מפעילים שרשרת פעולות חישוביות המבוצעות על גבי מאיצים גרפיים (GPU), צורכות זיכרון, רוחב פס וחשמל. במילים אחרות, בניגוד לתוכנה מסורתית, העלות אינה מסתיימת בפיתוח המודל אלא נמשכת לאורך כל מחזור חייו, בכל פעם שהוא מופעל.

זו הסיבה שמושג ה-Inference הפך לאחד המשתנים הכלכליים החשובים ביותר בשוק הבינה המלאכותית. בעוד ששלב האימון (Training) הוא השקעה חד-פעמית, שלב ההרצה (Inference) הוא פעילות מתמשכת, שהיקפה גדל ככל שמספר המשתמשים והיישומים גדל. למעשה, ככל שהבינה המלאכותית הופכת לחלק אינטגרלי מתהליכים עסקיים, כך עיקר העלות הכלכלית עובר משלב פיתוח המודל לשלב הפעלתו.

הטוקן הוא המדד שבאמצעותו ניתן למדוד את אותה פעילות. מספר הטוקנים המעובדים משקף, בקירוב, את היקף עבודת ה-Inference שהמודל נדרש לבצע. לכן, עבור ספקיות המודלים, מפעילי מרכזי הנתונים ואף עבור משקיעים, הטוקן הפך למדד מרכזי להבנת הביקוש העתידי לכוח מחשוב. לכאורה, ניתן היה לצפות שההתקדמות הטכנולוגית תקטין את חשיבותן של תשתיות המחשוב. ואכן, העלות לעיבוד טוקן נמצאת במגמת ירידה מתמשכת. דורות חדשים של מאיצים גרפיים, שיפורים בארכיטקטורת המודלים, טכניקות Quantization, מנגנוני Routing חכמים ושיטות אופטימיזציה שונות מאפשרים לבצע את אותה משימה בעלות נמוכה יותר מאשר לפני שנה בלבד.

אולם דווקא ירידת המחיר צפויה להגדיל את הביקוש הכולל לעיבוד. תופעה זו מוכרת בכלכלה כפרדוקס ג'בונס (Jevons Paradox): כאשר השימוש במשאב נעשה יעיל וזול יותר, הצריכה הכוללת שלו נוטה לגדול ולא להצטמצם. כך אירע בעבר בפחם, בחשמל וברוחב הפס, וכך עשוי לקרות גם בבינה מלאכותית.

מנכ"ל Microsoft, סאטיה נאדלה, התייחס לאחרונה לעיקרון זה בהקשר של AI. לדבריו, ככל שהעלות להפעלת מודלים תרד, כך ארגונים ישלבו אותם במספר גדול יותר של תהליכים, יגדילו את היקף השימוש וייצרו ביקוש גבוה אף יותר לכוח מחשוב. במילים אחרות, הירידה בעלות לטוקן אינה צפויה להפחית את הצורך בתשתיות, אלא דווקא להגדיל אותו.

מודלים "חושבים" משנים את מבנה העלויות

לצד הירידה בעלות לטוקן, מתפתחת מגמה נוספת בעלת משמעות כלכלית עמוקה: המעבר למודלי Reasoning. בניגוד למודלים המסורתיים, שמטרתם להפיק תשובה במהירות האפשרית, מודלים אלה מבצעים תהליך הסקה מורכב יותר, הכולל שלבי ביניים, בדיקות עצמיות ולעיתים אף יצירת שרשרת מחשבה (Chain of Thought).

התוצאה היא שיפור משמעותי באיכות התשובות, במיוחד במשימות מורכבות כמו כתיבת קוד, מחקר, ניתוח מסמכים או קבלת החלטות. אולם שיפור זה כרוך גם בעלייה בכמות הטוקנים המעובדים ובזמן החישוב הנדרש. ככל שיותר ארגונים יאמצו מודלים מסוג זה, צפוי לגדול גם הביקוש הכולל ל-Inference, אף אם מספר המשתמשים יישאר דומה.

במקביל, מערכות מבוססות AI Agents צפויות להעצים עוד יותר את התופעה. בניגוד למשתמש אנושי, סוכן AI עשוי לבצע עשרות ואף מאות קריאות למודל במסגרת ביצוע משימה אחת – חיפוש מידע, כתיבת קוד, הפעלת כלים, אימות תוצאות וניסוח תשובה. לכן, בעתיד הקרוב, מספר הטוקנים שצורך "משתמש" אחד עשוי לגדול באופן משמעותי, גם אם מספר המשתמשים הכולל לא ישתנה.

מגמות אלו יצרו תחום חדש ההולך ומתפתח – Token FinOps. בדומה ל-Cloud FinOps, שמטרתו לנהל באופן מושכל את עלויות הענן, Token FinOps עוסק בניהול הכלכלי של השימוש בבינה מלאכותית, תוך יצירת קשר בין צריכת משאבי המחשוב לבין הערך העסקי המופק מהם.

ניהול כזה אינו מסתכם במדידת מספר הטוקנים. הוא כולל בחירה במודל המתאים לכל משימה, קיצור חלונות הקשר (Context Windows), שימוש במנגנוני Caching, המאפשרים לעשות שימוש חוזר בתוצאות שכבר חושבו בעבר והפניית משימות פשוטות למודלים קטנים וזולים יותר. במודלים מסוימים, עיבוד מידע שנשלף מה-Cache מתומחר בעשירית ואף פחות מעלות עיבוד חדש, ולכן תכנון נכון של זרימת המידע עשוי להפחית משמעותית את העלויות. כלומר, בדיוק כפי שארגונים למדו בעשור האחרון לנהל את צריכת הענן שלהם, הם יידרשו בשנים הקרובות ללמוד לנהל גם את כלכלת הטוקנים.

מטוקנים למרכזי נתונים: כיצד כלכלת ה-Inference מעצבת את שוק התשתיות

הבנת כלכלת הטוקנים מסייעת להסביר מדוע עיקר תשומת הלב של השוק אינה מופנית כיום דווקא למודלים עצמם, אלא לתשתיות המאפשרות להפעילם. אם הטוקן הוא יחידת המדידה של היקף עבודת ה-Inference, הרי שהגידול במספר הטוקנים המעובדים מתורגם, כמעט באופן ישיר, לביקוש גדל והולך לכוח מחשוב.

מנקודת מבט זו, המרוץ העולמי להקמת מרכזי נתונים אינו תוצאה של אופנה חולפת או של מודל תמחור מסוים, אלא של שינוי מבני בכלכלת הבינה המלאכותית. ככל שיותר ארגונים משלבים AI בתהליכי העבודה שלהם, וככל שמודלים הופכים מורכבים יותר ומערכות סוכנים מבצעות מספר הולך וגדל של קריאות למודלים, כך גדל הביקוש ל-Inference. ביקוש זה מחייב הרחבה מתמדת של תשתיות מחשוב, הכוללות מאיצים גרפיים, רשתות תקשורת, מערכות קירור ואספקת חשמל בהיקפים שלא נדרשו בעבר.

מגמה זו באה לידי ביטוי בגל ההשקעות חסר התקדים שמאפיין כיום את הענף. OpenAI השלימה השנה גיוס הון לפי שווי של מאות מיליארדי דולרים, בעוד שמזם Stargate צפוי להשקיע עד 500 מיליארד דולר בהקמת תשתיות AI בארצות הברית בעשור הקרוב. במקביל, ענקיות הטכנולוגיה ממשיכות להרחיב בקצב מואץ את מרכזי הנתונים שלהן ולהשקיע עשרות מיליארדי דולרים נוספים ברכישת מאיצים גרפיים, בהקמת מתקני מחשוב חדשים ובהבטחת אספקת אנרגיה.

גם העלייה המטאורית בביקוש למעבדי NVIDIA אינה נובעת רק מהצלחת החברה לפתח חומרה מתקדמת, אלא מהעובדה שה-GPU הפך למשאב הייצור המרכזי של כלכלת הבינה המלאכותית. כשם שנפט היה חומר הגלם של הכלכלה התעשייתית, כך כוח מחשוב הופך בהדרגה למשאב היסוד של הכלכלה הדיגיטלית החדשה.

אחת ההתפתחויות המעניינות ביותר בשנים האחרונות היא הופעתן של חברות NeoCloud – ספקיות ענן המתמקדות כמעט באופן בלעדי בתשתיות AI. בניגוד לענן המסורתי, שנועד לשרת מגוון רחב של עומסי עבודה, חברות אלה בונות את פעילותן סביב אספקת כוח מחשוב ייעודי לאימון ולהרצת מודלים.

חברות כמו CoreWeave, Crusoe, Lambda ו-Nebius אינן מתחרות בעיקר על נפח אחסון או שירותי IT מסורתיים, אלא על היכולת לספק גישה מהירה ויעילה ל-GPU, לרשתות מהירות ולמערכות תוכנה המותאמות לעומסי עבודה של בינה מלאכותית. במובן זה, הן משקפות את המעבר מכלכלת ענן "כללית" ל"כלכלת Compute" – כלכלה שבה היתרון התחרותי נמדד בראש ובראשונה בזמינות כוח מחשוב.

מתשתיות לריבונות מחשוב

השלכותיו של מרוץ התשתיות אינן מסתיימות בעולם המחשוב. ככל שמרכזי הנתונים מתרחבים, כך גדל גם הביקוש לחשמל, למים לקירור ולקרקע מתאימה להקמת מתקנים חדשים.

לא במקרה חברות טכנולוגיה גדולות משקיעות כיום בתחנות כוח, בחוות סולאריות, בכורים גרעיניים קטנים (SMR) ובהסכמי אספקת אנרגיה ארוכי טווח. בשונה מהעשור הקודם, שבו צוואר הבקבוק היה בעיקר ייצור השבבים, האתגר המרכזי של השנים הקרובות צפוי להיות היכולת לספק למרכזי הנתונים אנרגיה יציבה וזמינה.

מגמה זו משנה גם את מפת ההשקעות הגלובלית. מדינות ואזורים המסוגלים להציע קרקע, חשמל, קישוריות רגולטורית ותשתיות מתאימות הפוכים ליעדים מבוקשים להקמת מרכזי נתונים, בעוד שמדינות שאינן מסוגלות לספק תנאים אלה עלולות למצוא את עצמן מחוץ לשרשרת הערך של הבינה המלאכותית.

התפתחות זו משנה גם את תפיסת הביטחון הלאומי. במשך שנים נתפסה ריבונות דיגיטלית בעיקר כשאלה של שליטה בנתונים, בהגנת סייבר או ברגולציה על פלטפורמות דיגיטליות. כיום הולכת ומתבססת תפיסה רחבה יותר, שלפיה גם הגישה לכוח מחשוב היא נכס אסטרטגי.

מדינות מבינות כי ללא גישה יציבה ל-GPU, למרכזי נתונים ולמודלי AI מתקדמים, הן עלולות להיות תלויות בהחלטותיהן של מספר מצומצם של חברות וטכנולוגיות זרות. ההתפתחויות האחרונות סביב הגבלת הגישה למודלים מתקדמים מסיבות של ביטחון לאומי בארצות הברית המחישו כי הגישה ליכולות AI עצמן מתחילה להפוך לכלי מדיניות, בדומה לבקורות הייצוא על שבבים מתקדמים.

מסיבה זו, ארה"ב, האיחוד האירופי, מדינות המפרץ ומדינות נוספות משקיעות כיום סכומי עתק בהרחבת יכולות המחשוב הלאומיות שלהן. המטרה אינה רק לעודד חדשנות, אלא להבטיח שלמדינה תהיה גישה עצמאית לתשתיות שעליהן תישען הכלכלה בעשורים הקרובים.

במילים אחרות, אם בעשור הקודם דובר על ריבונות נתונים, הרי שבעשור הקרוב צפויה לעמוד במרכז הדיון ריבונות מחשוב – היכולת של מדינה להבטיח גישה מתמשכת, אמינה ובטוחה למשאבי AI. כלכלת הטוקנים מספקת את העדשה שבאמצעותה ניתן להבין את המעבר הזה: לא משום שהטוקן הוא מקור ההשקעות, אלא משום שהוא משקף את היקף הביקוש לכוח מחשוב, שהופך בהדרגה למשאב האסטרטגי החשוב ביותר של עידן הבינה המלאכותית.

משמעויות לישראל

השינויים המתחוללים בכלכלת הבינה המלאכותית אינם פוסחים על ישראל. במשך שנים נשען היתרון היחסי של המדינה על ההון האנושי, על מערכת אקדמית חזקה ועל אקוסיסטם מפותח של סטארט-אפים. אולם ככל שהבינה המלאכותית הופכת מתוכנה לתשתית, מתברר כי יתרונות אלה, חשובים ככל שיהיו, אינם מספיקים עוד. גם מדינות חדשניות נדרשות להבטיח לעצמן גישה מתמשכת לכוח מחשוב מתקדם.

הבנה זו באה לידי ביטוי בהחלטת הממשלה 4255 מיוני 2026, המגדירה את הבינה המלאכותית כנכס אסטרטגי בעל השלכות על הצמיחה הכלכלית, הביטחון הלאומי והחוסן החברתי. אחד מחידושיה המרכזיים של ההחלטה הוא ההכרה בכך שהאתגר אינו מסתכם עוד בעידוד מחקר או בהגדלת מספר הסטארט-אפים, אלא בהבטחת גישה ארוכת טווח למשאבי מחשוב מתקדמים.

בהתאם לכך, קבעה הממשלה יעד שלפיו תובטח למשק הישראלי, בטווח של חמש שנים, נגישות לכוח עיבוד השקול לביצועיהם של כ-100,000 מאיצי בינה מלאכותית (GPU), שישרתו את האקדמיה, המגזר הציבורי והתעשייה. אין מדובר רק ביעד טכנולוגי; מדובר בתפיסה אסטרטגית חדשה שלפיה מחשוב הופך לחלק מתשתיות היסוד של המדינה – בדומה לחשמל, מים או תקשורת.

הקמת מחשב העל הלאומי לבינה מלאכותית (2025) היא ביטוי ראשון לתפיסה זו. חשיבות המיזם אינה נובעת רק מהגדלת היצע כוח המחשוב, אלא מהנחת התשתית לאקוסיסטם לאומי שיאפשר לאקדמיה, לסטארט-אפים, לתעשייה ולממשלה לפתח ולהפעיל יישומי AI מבלי להיות תלויים באופן מלא בגישה למשאבים מחוץ לישראל.

עם זאת, יש להכיר גם במגבלות. ישראל אינה צפויה להתחרות בארצות הברית, בסין או במדינות המפרץ בהיקף ההשקעות במרכזי נתונים או במספר המאיצים הגרפיים. יתרונה ימשיך להישען בעיקר על יכולתה לפתח יישומים חדשניים, אלגוריתמים, פתרונות סייבר, בריאות דיגיטלית וטכנולוגיות ביטחוניות. לכן, האתגר האסטרטגי של ישראל אינו להפוך למעצמת מחשוב עולמית, אלא להבטיח שלחוקרים, לחברות ולמוסדות המדינה תהיה גישה אמינה ורציפה למשאבי AI מתקדמים.

במובן זה, ריבונות מחשוב אינה מחייבת בעלות מלאה על כל שרשרת הערך. היא מחייבת בראש ובראשונה ביטחון בגישה (Access Security) – היכולת להבטיח שגם בתקופות של משבר גיאופוליטי, מחסור עולמי או מגבלות ייצוא, יעמדו לרשות המשק הישראלי משאבי מחשוב בהיקף מספק.

עורכי הסדרה: ענת קורץ, רינת חרש ואלדד שביט.