

Jensen Huang ב-All-In: תקציב טוקנים, 100 agents לכל מהנדס, ומודלים פתוחים מול סגורים

מנכ"ל NVIDIA על עבודה עם agents, העתיד של תוכנת enterprise, מודלים פתוחים מול
מודלים קנייניים ושרשרת האספקה הגאופוליטית

Jensen Huang עם מגישי All-In · All-In (שידור חי מאולם תצוגה; האירוע עצמו לא נקרא בשמו). הוידאו פורסם מחדש
ב-X על ידי 2026-10-08 Komal Prajapati (תאריך הפרסום מחדש)

משך: כ-16 דקות · X

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זה קטע של כ-16 דקות מתוך שיחה חיה של פודקאסט All-In עם Jensen Huang, מנכ"ל NVIDIA. השיחה צולמה ליד דסק עם הלוגו של All-In, באולם תצוגה מלא בקהל. הקטע הגיע מפרסום מחדש ב-X, והוא נראה ערוך: יש קפיצות חדות בין נושאים, שאלה על auto research נענית בתשובה על OpenClaw, והסרטון נקטע באמצע שאלה על נהיגה אוטונומית. מדברים בו Huang, David Friedberg (מנכ"ל Ohalo; התמלול כותב Freberg), מראיין ש-Huang פונה אליו בשם Brad, ומגישים נוספים שאי אפשר לזהות בוודאות מהתמלול.

הטענה המרכזית של Huang: מהנדס טוב צריך לצרוך טוקנים בהיקף של חצי מהשכר שלו, ותכנות הופך לכתיבת מפרטים, ארגון צוותים של agents והגדרת קריטריונים להערכה. תוכנת enterprise לא תיעלם, כי agents ישתמשו בכלים הקיימים. ובעולם יהיו גם מודלים קנייניים וגם מודלים פתוחים. החלק השני של השיחה עוסק בגאופוליטיקה: השוק הסיני, ה-diffusion של טכנולוגיה אמריקאית, המזרח התיכון, ישראל וטאיוואן.



צילום רחב של הפאנל ליד הדסק של All-In, באולם תצוגה מלא בקהל. Jensen Huang במעיל עור שחור, לצד המגישים

איך לקרוא את המסמך

- 1. **תקציר מנהלים:** הטענות המרכזיות בשורה אחת כל אחת, ומה חסר בשיחה.
- 2. **תקציב טוקנים:** למה מהנדס שמשתכר \$500,000 צריך לצרוך טוקנים בשווי \$250,000.
- 3. **מקוד למפרטים:** איך נראית עבודה עם 100 agents.
- 4. **Ohalo:** החלפת stack ב-90 דקות ומחקר של שבע שנים ב-30 דקות. אלה עדויות של Friedberg, לא של Huang.
- 5. **OpenClaw ותוכנת enterprise:** למה agents יגדילו את השימוש בכלים הקיימים.
- 6. **מודלים פתוחים וסגורים:** A and B, לא A or B.
- 7. **גאופוליטיקה:** סין, ה-tech stack האמריקאי, ישראל והמזרח התיכון, טאיוואן והליום.
- 8-9. **מסקנות ומילון:** מה לעשות עם זה, שאלות פתוחות, ומונחים.

הערת מקור המסמך הופק אוטומטית מהתמלול של ההקלטה ומהפריימים שחולצו ממנה. אין בהקלטה שקפים, ובכל הפריימים רואים רק את הדוברים על הבמה. שמות ומונחים עלולים להשתבש בתמלול האוטומטי: Claude מופיע כ-Clawed וכ-Friedberg, Cloud כ-Freberg, ו-Lutnick כ-Lutnik. הציטוטים מועתקים מילה במילה מהתמלול ונבדקו מול חותמת הזמן שלהם. הכיתוב של הפוסט ב-X לא שימש כמקור, ולא נוסף שום מידע ממקורות חיצוניים.

1. תקציר מנהלים

Huang מתאר עולם שבו העלות העיקרית של מהנדס היא כבר לא רק השכר שלו אלא גם הצריכה שלו ב-compute, והעבודה עצמה עוברת מכתובת קוד לניהול צוות של agents. את רוב העדויות המספריות על האצה בשיחה סיפק דווקא Friedberg, מתוך Ohalo. בחלק הגאופוליטי Huang מדבר מנקודת המבט של NVIDIA: איבוד השוק הסיני, ורצון שה-tech stack האמריקאי יהיה הבסיס של העולם.

- **מהנדס בלי תקציב טוקנים הוא כמו מעצב שבבים בלי CAD:** מהנדס שמרוויח \$500,000 וצרך רק \$5,000 בטוקנים יגרום ל-Huang "to be deeply alarmed". הרף שלו הוא \$250,000 בשנה [0:00].
- **תכנות הופך לכתיבת מפרטים וניהול צוותים:** "every engineer is going to have 100 agents". העבודה תהיה כתיבת ideas, architectures, specifications והגדרת מה טוב ומה רע [1:34].
- **מחסומי "קשה מדי" ו"ייתקו יותר מדי זמן" נעלמים:** Huang משווה למהפכה התעשייתית, שבה אף אחד כבר לא אומר שבניין נראה כבד מדי [1:34–0:00].
- **החלפת stack בלילה אחד:** Friedberg מספר שבנה ופרס ב-Claude מערכת agentic שהחליפה stack שלם, ב-90 דקות ביום ראשון בערב, ושכל הנהלת Ohalo עשתה תרגיל דומה [3:05–1:34].
- **תוכנת enterprise תתחזק, לא תיהרס:** היום היא מוגבלת במספר המשתמשים האנושיים, ותקבל פי 100 agents שמשתמשים בכלים. הכלים האלה הם ה-ground truth שדרכו האדם שולט בתוצאה [6:09–4:37].
- **מודלים פתוחים וקנייניים יתקיימו יחד:** "models is a technology, not a product". מודלים אופקיים ישמשו את הצרכנים, ומודלים פתוחים ישמשו תעשיות שרוצות לשלוט במומחיות שלהן [7:44–6:09].
- **השוק הסיני של NVIDIA ירד מ-95% ל-0%:** יש רישיונות שאושרו, הזמנות רכש מחברות סיניות, ושרשרת האספקה מתחילה לפעול מחדש [9:21].
- **היעד: ה-tech stack האמריקאי ב-90% מהעולם:** מצב כמו בתחומי הסולארי, המגנטים והתקשורת, שבהם ארה"ב לא שולטת, הוא בעיני Huang תוצאה רעה לביטחון הלאומי [10:58].
- **ישראל וטאיוואן:** "We are 100% in Israel". מול טאיוואן הוא מציע שלושה צעדים: תיעוש מחדש של ארה"ב, פיזור של שרשרת האספקה, וריסון [14:11–12:31].

מה לא נאמר

- **אין "hundreds of agents":** הכיתוב של הפוסט ב-X מצטט "Every engineer is going to have and manage hundreds of agents". המשפט הזה לא מופיע בתמלול. Huang אמר "100 agents", ובלי המילה manage [1:34].
- **אין מתודולוגיה לרף הטוקנים:** היחס של 50% מהשכר מוצג כאינטואיציה. לא נאמר איך הוא נקבע ולא הוצגו נתונים על התפוקה שהוא מניב.
- **הסיפורים של Ohalo לא מאומתים:** לא נמסר מה בדיוק הוחלף ב-stack, ולא איזה ממצא פורסם פנימית. הטענה שמדובר בעבודה ששווה דוקטורט של שבע שנים ומאמר ב-Science היא הערכה של Friedberg.
- **Huang לא ענה על השאלה על auto research:** השאלה על 600 שורות קוד ועל ריצה מקומית נענתה (או נערכה) בתשובה על OpenClaw.
- **לא ניתן ציון ל-Huang diffusion:** נשאל אם המצב הוא A, B או C ולא נתן ציון.
- **לא הייתה התייחסות ל-BitTensor:** השאלה על אימון מבוזר של מודל LLaMA בגודל 4B נענתה בתשובה כללית על מודלים פתוחים.
- **הנושא של הנהיגה האוטונומית נקטע:** השאלה על BYD, Mercedes, Uber ופלטפורמה "Android-type" נשאלה, אבל הקטע נגמר לפני התשובה [15:42].

2. תקציב טוקנים: "all-star employees" עם יכולות

על-אנושיות

Huang פותח בתרחיש מספרי: מהנדס או חוקר AI שמשתכר \$500,000 בשנה ומדווח בסוף השנה שהוציא \$5,000 על טוקנים. התגובה שלו חריפה, והוא משווה את זה למעצב שבבים ב-NVIDIA שמודיע שיעבוד עם נייר ועיפרון בלי

כלי CAD. מראיין מוסיף אנלוגיה מה-NBA: LeBron James, שמוציא מיליון דולר בשנה על תחזוקת הגוף ועדיין משחק בגיל 41. Huang מסכים: אם אלה עובדי ידע יוצאי דופן, צריך לתת להם יכולות על-אנושיות.

כשהוא נשאל איך ייראה עובד כזה בעוד שנתיים-שלוש, הוא מונה שלוש מחשבות שייעלמו: "זה קשה מדי", "זה ייקח הרבה זמן" ו"נצטרך הרבה אנשים". מבחינת מי שמנהל צוות הנדסה, המשמעות היא שצריכת טוקנים הופכת לממד של עבודה תקינה ולא לסעיף הוצאה שצריך לקצץ.

Jensen Huang [0:00] If that \$500,000 engineer did not consume at least \$250,000 worth of — tokens, I am going to be deeply alarmed

3. מקוד למפרטים: 100 agents לכל מהנדס

Huang ממשיך את הטיעון בהשוואה למהפכה התעשייתית: היום אף אחד לא אומר שבניין נראה כבד מדי או שהר נראה גדול מדי. "[1:34] Everything that's too big, too heavy, takes too long. Those ideas are all gone." מה שנשאר הוא היצירתיות.

כשהוא נשאל איך עובדים עם agents, הוא עונה שזו פשוט צורה חדשה של תכנות. המהנדס יכתוב רעיונות, ארכיטקטורות ו-specifications, יארגן צוותים, ויגדיר איך מעריכים מה טוב ומה רע, איך נראית תוצאה מצוינת, איך עושים איטרציות ואיך עושים סיעור מוחות. בפועל זה תיאור של העבודה של tech lead: הגדרת evals ו-acceptance criteria במקום כתיבת קוד.

Jensen Huang [1:34] In the past, we code. In the future, we're going to write ideas, architectures, — specifications. We're going to organize teams

4. Ohalo: stack שלם ב-90 דקות ו"דוקטורט" ב-30 דקות

החלק הזה הוא עדות של David Friedberg, מנכ"ל Ohalo, ולא של Huang. Friedberg מספר שבילילה של יום ראשון, בין 22:00 ל-23:30, הריץ ב-Claude מערכת agentic שבנתה ופרסה תחליף ל-stack תוכנה שלם ולחלק גדול מה-workload. כל חברי ההנהלה התבקשו לעשות תרגיל דומה באותו סוף שבוע, וביום שני, לדבריו, "I was like, it's over".

בצד המדעי הוא מספר שהצוות הוריד מ-GitHub את auto research ביום שישי, הריץ אותו על מחשב שולחני עם הדאטה הגנומי שלהם, ותוך 30 דקות קיבל ממצא שלדבריו היה "PhD thesis that would take seven years" וראוי להתפרסם ב-Science. הוא גם מזכיר zero shot genomic modeling. אחד המראיינים מוסיף שהכלי פורסם "with 600 lines of code in a weekend" ורץ מקומית. כל המספרים האלה הם תיאור מפי הדובר, ולא הוצגו עליהם נתונים.

David Friedberg [3:05] 90 minutes on Claude ran this agentic system, built the whole thing, — deployed it

5. OpenClaw, tool use ולמה תוכנת enterprise לא תיעלם

לפי Huang, OpenClaw הצליח בעיקר בגלל התזמון שלו. הוא הגיע בדיוק כשהמודלים (Claude, GPT ו-ChatGPT) הגיעו לרמה טובה מאוד ופיתחו יכולת tool use. ההערכה של Huang היא ש-Peter, היוצר, כנראה לא היה מגיע לרעיון בלי זה. הכלים שנבנו לאורך השנים, כמו דפדפנים, Excel, ובתכנון שבבים, Synopsys, Cadence, Autodesk ו-Omniverse, ימשיכו לשמש.

כאן הוא יוצא נגד התזה שתוכנת enterprise תיהרס. לטענתו היא מוגבלת היום במספר המשתמשים האנושיים, ועוד מעט יעבדו עליה פי 100 agents: על SQL, על vector databases, על Blender ועל Photoshop. יש לזה שתי סיבות. הכלים עושים עבודה טובה, והם הממשק בין ה-agent לאדם. התוצאה צריכה לחזור לכלי שהאדם יודע לשלוט בו, ולכן ב-NVIDIA הוא רוצה שהכל יחזור ל-Synopsys ול-Cadence. שם נמצא ה-ground truth.

Jensen Huang [4:37] The enterprise software industry is limited by butts and seats. It's about to — get 100 times more agents banging on those tools

6. מודלים פתוחים מול קנייניים: A and B

השאלה נפתחת בדוגמה של המראיין: אימון מבוצע במסגרת BitTensor, Subnet3 (השם כפי שהופיע בתמלול), של מודל LLaMA בגודל 4 מיליארד פרמטרים, עם compute עודף שתרמו משתתפים רבים. המראיין שואל לאן זה מוביל את הקוד הפתוח. Huang לא מתייחס לפרויקט עצמו ועונה ברמת העיקרון: "models is a technology, not a product" [6:09]. לצרכן הרגיל, השכבה האופקית של general intelligence תישאר מוצר. הוא עצמו לא רוצה לעשות fine-tune ומעדיף ChatGPT, Claude, Gemini ו-X, שלכל אחד מהם אישיות משלו.

תעשיות עם מומחיות ייעודית, לעומת זאת, צריכות ללכוד את הידע שלהן בצורה שהן שולטות בה, ולדבריו זה אפשרי רק עם מודלים פתוחים. NVIDIA תורמת לתחום הזה, והוא "near the frontier". מראיין מציין שכל סטארט-אפ שהם משקיעים בו מתחיל בקוד פתוח ועובר אחר כך למודלים קנייניים. Huang מוסיף שעם router טוב יש גישה למודל הטוב בעולם כבר מהיום הראשון, ואז יש זמן להוזיל, לעשות fine-tune ולהתמחות.

— These two things are not A or B, it's A and B. There's no question about it [6:09] Jensen Huang.

7. גאופוליטיקה: סין, ה-tech stack האמריקאי, ישראל וטאיוואן

Huang נשאל לתת ציון ל-diffusion של AI אמריקאי שנה אחרי ביטול כלל ה-diffusion מתקופת Biden. הוא לא נותן ציון, ובמקום זה מציג נתון: NVIDIA ויתרה על נתח שוק של 95% בשוק השני בגודלו בעולם, ועכשיו היא ב-0% בסין. יש רישיונות שאושרו על ידי השר Lutnick (בתמלול: Lutnik), חברות סיניות שלחו הזמנות רכש, ושרשרת האספקה מתחילה לפעול מחדש. החזון שלו הוא שהעולם יבנה את ה-AI שלו, ציבורי או פרטי, על ה-tech stack האמריקאי, משבבים ועד פלטפורמות: "[10:58] 'I would love that the American tech stack is 90 percent of the world.' הדוגמאות השליליות שהוא מביא הן מנועים זעירים, מינרלים נדירים, מגנטים, סולארי ורשתות תקשורת.

על המזרח התיכון הוא אומר של-NVIDIA יש באזור 6,000 משפחות, כולל עובדים איראנים שמשפחותיהם באיראן. הוא חוזר על המחויבות לישראל, ומעריך שבסוף המלחמה המזרח התיכון יהיה יציב יותר ושזה יהיה מקום טוב להתרחב אליו עם AI. בנושא הליום הוא מעריך שזו עלולה להיות בעיה, אבל ששרשרת האספקה כנראה מחזיקה הרבה buffer.

— We are 100% in Israel. We are 100% behind the families there [12:31] Jensen Huang.

- **תיעוש מחדש של ארה"ב:** מפעלי שבבים, מפעלי מחשבים ו-AI factories. לדבריו בזכות שרשרת האספקה הטאיוואנית נבנו מפעלים באריזונה, בטקסס ובקליפורניה "[12:31–14:11] 'at incredible rates'.
- **פיזור שרשרת הייצור:** דרום קוריאה, יפן ואירופה, כדי להגדיל את העמידות [14:11].
- **ריסון:** "let's demonstrate restraint". לא ללחוץ בלי צורך בזמן שבונים גיוון ועמידות [14:11].

8. מסקנות ומה עושים מחר

- **למדוד צריכת טוקנים כמדד ולא רק כעלות:** בדקו כמה כל מהנדס בצוות צורך בפועל. אם זה בטווח של \$5,000 לשנה, לפי Huang זה סימן אזהרה ולא חיסכון. הרף של 50% מהשכר הוא דעה ולא נתון מבוסס.
- **להשקיע בכתיבת specs ו-evals:** הכישורים ש-Huang מונה הם ארכיטקטורה, מפרטים, הגדרה של "good versus bad" ואיטרציה. כדאי לתרגל אותם כבר עכשיו על agents כמו Claude Code.
- **לעשות תרגיל החלפה בסגנון Ohalo:** הקצו ערב אחד לבדוק האם agent יכול לבנות תחליף לכלי פנימי. Friedberg עשה את זה עם כל ההנהלה באותו סוף שבוע.
- **לחשוב על הכלים הקיימים כמשקל-agents:** אם Huang צודק, SQL, vector DB וכלי התכנון יקבלו הרבה יותר תעבורה מ-agents. כדאי לבדוק אם ה-API, ההרשאות וה-rate limits שלהם מוכנים לזה.
- **אסטרטגיית מודלים כפולה:** להתחיל עם מודל frontier דרך router, ובמקביל להכין מסלול ל-fine-tune של מודל פתוח על הדאטה הייעודי שלכם.

שאלות פתוחות

שאלה	למה היא פתוחה
האם 50% מהשכר בטוקנים מחזיר את ההשקעה?	לא הוצגו נתוני תפוקה, רק אנלוגיה ל-CAD ול-LeBron James.
מה בדיוק Ohalo החליפה בתוך 90 דקות?	לא נאמר איזה stack, באיזה היקף, ומה היה המצב אחרי חודש.
האם הממצא של auto research עומד בביקורת עמיתים?	פורסם פנימית בלבד. הטענה שמדובר בעבודה ברמה של Science היא הערכה של הדובר.
מה עמדתו של Huang על אימון מבוזר (BitTensor)?	השאלה נשאלה ולא נענתה ישירות.
מה האסטרטגיה של NVIDIA בנהיגה אוטונומית?	הקטע נקטע לפני התשובה על המודל של Android מול iOS.
כמה מהשוק הסיני NVIDIA תצליח להחזיר?	יש רישיונות והזמנות רכש, אבל לא נמסרו היקפים או לוחות זמנים.

9. מילון מונחים

מונח	פירוש
Tokens	יחידות הטקסט שמודל שפה מעבד. התמחור והצריכה של מודלים נמדדים לפיהן
Agent / agentic system	מערכת שבה מודל שפה מתכנן ומבצע משימות בכמה שלבים ומשתמש בכלים
CAD tools	כלי תכנון ממוחשב. בתכנון שבבים: Synopsys ו-Cadence
Specifications	מפרטים: תיאור כתוב של מה שצריך לבנות ושל קריטריוני ההצלחה
Tool use	היכולת של מודל להפעיל כלים חיצוניים כמו דפדפן, גיליון אלקטרוני או מסד נתונים
OpenClaw	פרויקט agent בקוד פתוח ש-Huang מייחס ל-Peter ומתאר כפריצת דרך בתזמון מושלם
Auto research	כלי מחקר אוטומטי שהורד מ-GitHub. לפי המראיין, כ-600 שורות קוד שרצות מקומית
Butts and seats	מודל הרישוי של תוכנת enterprise לפי מספר משתמשים אנושיים
Ground truth	מקור האמת. כאן: הכלי שבו האדם מאמת את התוצאה ושולט בה
Open weights / open source models	מודלים שהמשקלות שלהם זמינים להורדה, ל-fine-tune ולהרצה עצמאית
Fine-tune	אימון נוסף של מודל קיים על דאטה ייעודי
Router	שכבה שמנתבת כל בקשה למודל המתאים או הטוב ביותר
Zero shot genomic modeling	מידול גנומי בלי אימון ייעודי על דוגמאות מתויגות, כפי שתיאר Friedberg
Diffusion rule	כלל ממשל Biden להגבלת הפצה של שבבי AI ומודלים מחוץ לארה"ב
Tech stack	שכבות הטכנולוגיה, משבבים ומערכות מחשוב ועד פלטפורמות
AI factories	מרכזי נתונים שמוקדשים לאימון ולהרצה של מודלים, במינוח של NVIDIA