

Sundar Pichai על agents, סייבר, קוד פתוח ו-compute — ראיון על במת Google I/O

ראיון של כ-32 דקות עם מנכ"ל Google על agents ברשת, CodeMender, אסטרטגיית
Flash ומחסור ב-compute

Sundar Pichai, המראיין Matt (כנראה Matthew Berman) · שיחה על במת Google I/O, המקור המקורי לא זוהה
(הועלה מחדש ב-X על ידי spect) · הועלה ב-30.09.2026

משך: כ-32 דקות · ראיון

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זו שיחה על הבמה בין Sundar Pichai, מנכ"ל Google ו-Alphabet, לבין מראיין בשם Matt. לפי סימן המים שמופיע בפריימים זה כנראה Matthew Berman, יוצר תוכן בנושאי AI. לוגו I/O על הקיר מאחוריהם ואמירות כמו "announcements today" מרמזים שהשיחה צולמה במסגרת Google I/O. הסרטון הגיע מ-repost ב-X, והכיתוב שלו מייחס ל-Pichai ציטוט על בניית harness ל-agents. הציטוט הזה לא מופיע בתמליל.

הטענה המרכזית: agents יהפכו לחלק בסיסי מהשימוש ברשת, אבל האמון בהם צריך להיבנות בהדרגה. בסייבר, ההחלטה אם לשחרר מודל תלויה בגודל הקפיצה שהוא מביא. Google פועלת בגישה מאוזנת בקוד פתוח, ומשקיעה הרבה במודלי Flash כי ה-compute מוגבל והלקוחות חורגים מהתקציב. לפי Pichai, הביקוש ב-Google עולה על ה-compute הזמין, וצווארי הבקבוק נמצאים בכל שכבות ה-stack.

איך לקרוא את המסמך

- 1. **תקציר מנהלים:** הטענות המרכזיות וכל מה שלא נאמר, כולל הציטוט מהכיתוב שלא קיים בתמליל.
- 2. **agents כנקודת הכניסה לרשת:** שליטה, אמון, Gemini Spark והשכבה שנוצרת בין המשתמש לרשת.
- 3. **סייבר, Mythos ו-agentic workflows: CodeMender** בצוותי האבטחה ומדיניות שחרור של מודלים חזקים.
- 4. **קוד פתוח, Gemma ומודלים מסין:** הגישה המאוזנת של Google ואיך לבנות כך שאפשר להחליף מודלים.
- 5. **Flash כמודל עבודה:** למה Google משקיעה במודלים מהירים וזולים, ושילוב Pro עם Flash.
- 6. **מרוץ, שיפור עצמי ו-race condition: compute**, פער התפיסה בין המעבדות וצווארי הבקבוק.
- 7. **מה עושים עם זה, ו-8. מילון מונחים.**

הערות מקור המסמך הופק אוטומטית מהתמליל של ההקלטה ומהפריימים שלה. שמות ומונחים עלולים להשתבש בזיהוי הדיבור (ASR); למשל "Viz" הוא כנראה "JAMA", "Wiz" הוא Gemma ו-"iOS" הוא I/O. הציטוטים מועתקים מילה במילה ונבדקו מול חותמת הזמן שלהם. לא נוסף מידע ממקורות חיצוניים. בהקלטה אין שקפים, כל הפריימים הם צילומי במה.

1. תקציר מנהלים

זה ראיון ולא הרצאה מתוכננת. רוב התשובות של Pichai עוסקות בעקרונות ובכיוון ולא במספרים. הפרטים הקונקרטיים הם שמות מוצרים (Gemini Spark, CodeMender, Gemma 4, Flash), עיקרון ה-90 יום של Project Zero, נתון של 800 tokens לשנייה מדמו, ואמירה מפורשת שהביקוש עולה על ה-compute.

- **agents יהיו חלק מרכזי מהשימוש ברשת, בתנאי שהמשתמש שומר על שליטה:** Pichai רוצה שהמשתמשים ירגישו "a sense of control and agency and transparency" [1:33]. לדבריו, המפתחים כבר עברו בחודשים האחרונים מ-autocomplete ל-[0:00] agentic workflows.
- **Gemini Spark נפתח בהדרגה:** בשלב הראשון רק מול Gmail ו-Calendar, ורק אחר כך MCP, computer use ו-browser use לצד שלישי, אף שהיכולות כבר קיימות [3:09].
- **Google מפעילה agents באבטחה הפנימית ומתכננת להוציא את CodeMender החוצה:** הכלי מזהה פגיעויות, מייצר patches, בודק ומפיץ אותם, ורץ 24/7. בנוסף הוזכרה רכישת Wiz שהושלמה [9:19].
- **מדיניות שחרור במקרה של כל מודל לגופו:** השאלה היא אם המודל מוסיף 1-2% או קופץ ב-20%. מודל שמשנה את ה-frontier מצדיק עבודה עם הממשלה, אבל גם גישה רחבה מספיק כדי שאפשר יהיה לתקן מערכות [12:20].
- **קוד פתוח: גישה מאוזנת ולא מודל frontier פתוח:** Gemma 4 הוצג כגרסה חזקה, אבל לפי Pichai הגעה ל-frontier דורשת השקעה כבדה [13:51].
- **מודלים פתוחים מסין לא מטרידים אותו:** כשיש רישוי מתאים, המקור פחות משנה. מה שחשוב הוא לבנות כך שאפשר להחליף את המודל שמתחת [18:27], [19:58].
- **Flash הוא תשובה ישירה לחריגות בתקציב:** מנהלי IT (CIO) מתלוננים על עלויות, ובתוך Google משלבים Pro ו-Flash [21:30], [23:03].
- **הביקוש ב-Google גבוה מה-** "Absolutely" [29:14] **compute.** העלויות עולות, כולל מחירי זיכרון, וצווארי הבקבוק מתחלפים ביניהם בכל ה-[30:45], [27:43] stack.

מה לא נאמר

- **הציטוט מהכיתוב לא קיים:** המשפט "If you don't learn how to build a harness..." ורצף התפקידים Fixer → Reviewer → Tester → Writer → Researcher לא מופיעים בתמליל. אין לייחס אותם ל-Pichai.
- **אין מספרים על היקף התקיפות:** Pichai אומר שמזהים יותר פגיעויות ככל שהמודלים משתפרים, אבל לא נותן כמויות, שיעורי תיקון או נתונים על תקיפות שנעזרות ב-AI.
- **אין סף מוגדר לעצירת שחרור:** ה-1-2% מול 20% הם רק המחשה. לא הוצג מדד או benchmark.
- **אין נתוני עלות לטענות על** "Flash: remarkably cost efficient" **נאמר בלי מחירים ובלי השוואה.**
- **אין גודל לפער בין ביקוש ל-compute:** השאלה "What is the scale of that?" נשארה בלי תשובה.
- **אין תאריך שיחה מאומת:** ההקשר של I/O עולה מהלוגו ומהתוכן, ולא מכותרת רשמית.

2. agents כנקודת הכניסה לרשת



Sundar Pichai על הבמה, עם כיתוב שמזהה אותו כ-CEO of Google and Alphabet. בכל ההקלטה אין שקפים.

המראיין פותח בשאלה אם agents יהפכו לנקודת הכניסה לרשת. Pichai מתחיל מהמפתחים, שלדבריו נמצאים בחזית: לפני שנתיים הם התחילו לקבל השלמות קוד, "But over the last few months, developers are actually doing agentic workflows, right?" [0:00] הוא מביא כדוגמה את הדמו ב-Antigravity שבו נבנתה מערכת הפעלה.

בהמשך הוא מפריד בין מטלות שהן עול לבין חוויות שאנשים נהנים מהן. agent יכול למלא "farm fields 18" (כנראה form fields) כדי לחדש רישיון ב-DMV, אבל אדם שקונה מתנה לאדם אהוב לא ירצה להעביר את זה לאחר [1:33]. כשנשאל על אמון, הוא משווה ל-spam filter ב-Gmail ול-Waymo: אמון נבנה לאורך זמן, עם נתונים ועם ניסיון תפעולי [3:09].

Sundar Pichai [3:09] Gemini Spark is actually very powerful under the hood, but we are taking the — deliberate, careful step of making it work with your first-party surfaces, like Gmail, Calendar, etc. before we expose third-party with MCP and full computer use and browser use

על החשש שה-agents מוסיפים שכבה בין המשתמש לרשת הגולמית, Pichai מודה שקיימת "a layer of abstraction". מנגד הוא טוען שגם הכלים ליצירת תוכן מתרבים, ולכן "there'll be a new balance which we will find" [6:12]. YouTube משמש לו דוגמה לכך שאנשים מחפשים קשר ישיר עם יוצרים, ושגם ה-agents אמורים לחבר את המשתמש למערכת התוכן ולא להחליף אותה.

3. סייבר, Mythos ו-CodeMender

לדברי Pichai, צוותי האבטחה הפנימיים של Google משתמשים ב-agentic workflows כדי לזהות פגיעויות ולתקן אותן. בשנתיים האחרונות, ככל שהמודלים השתפרו, הם מזהים יותר פגיעויות [7:44]. את Mythos של Anthropic הוא מתאר כ-"a point of inflection", מודל שנבנה במיוחד למשימה הזו והגיע ל-frontier שלה [9:19].

הבשורה הקונקרטית היא CodeMender. לדבריו זו אחת ההכרזות שלא קיבלו מספיק תשומת לב ב-I/O (בתמליל: "at iOS CodeMender"). Google משתמשת בכלי בתוך החברה ובונה אותו כדי לשתף החוצה. בשילוב עם Wiz (בתמליל: "Viz"), שהרכישה שלה הושלמה ושמטמחה בניטור פגיעויות בזמן אמת, Google משתמשת בשניהם כדי להישאר בחזית. Pichai גם משבח את שיתוף הפעולה בין החברות בתעשייה, ומזכיר את SynthID ואת watermarking כדוגמאות.

Sundar Pichai — על CodeMender [9:19] it not only helps you identify the vulnerabilities, generate patches, test and verify that they work and deploy them

המראיין משווה בין ההחלטה של Anthropic לא לשחרר את Mythos לציבור לבין השחרור של GPT 5.5 cyber של OpenAI. Pichai-ב-OpenAI נשען על נוהלי תעשיית האבטחה: ב-Project Zero של Google מודיעים לספק על פגיעות ונותנים לו 90 יום לתקן לפני חשיפה פומבית [10:49]. כשמודל משנה את ה-frontier בצורה דרמטית, "I think it's important to work closely with the government on that" עם זאת, צריך לדאוג שמספיק גורמים יקבלו גישה כדי לתקן את המערכות שלהם [12:20].

Sundar Pichai [12:20] Is it a 1% to 2% improvement or the current state of the art? Or are you — taking these 20% jumps? That's where the judgment comes in

4. קוד פתוח, Gemma ומודלים מסין

המראיין טוען ש-Google ו-NVIDIA הן היחידות עם אסטרטגיית open source אמיתית, ושואל למה Google לא משחררת מודל frontier פתוח וגדול. Pichai מזכיר את Chromium, Android ו-Kubernetes, ואת סדרת Gemma, שהגרסה האחרונה שלה היא Gemma 4 (בתמליל: "JAMA"). לדבריו, הגעה ל-frontier דורשת השקעות CapEx ו-R&D גדולות, ולכן Google נוקטת "a balanced approach" [13:51], [15:23]. הוא מוסיף שהקוד הפתוח מדביק פערים כשקצב ההתקדמות של הטכנולוגיה מאט, ומפגר כשה-frontier רץ מהר [15:23].

על השאלה למה ארגון אמריקאי לא יבחר ב-DeepSeek, שמחירו נמוך בהרבה, התשובה שלו פרגמטית. ארגונים מחפשים צפיות, אמינות, עקביות ואבטחה. סביב מודלים פתוחים מתפתחת קהילה, ולכן בעיה בתוכנה לא תעבור בלי שישימו לב [18:27].

Sundar Pichai [18:27] I am more okay if it is open source with the right licenses, it should matter — less where it came from

מה שמדאיג אותו יותר הוא אם ארה"ב עושה מספיק כדי להישאר ב-frontier [19:58]. המראיין שואל אם מודלים סיניים יעברו אופטימיזציה לשבבים סיניים, וכך ייווצר תלות טכנולוגית. התשובה היא בעצם עצה ארכיטקטונית: המודלים משתנים מהר מדי, ולכן צריך לבנות "in a way in which you're able to evolve the models underneath". לדבריו, מוקדם לקבוע אם החשש מוצדק [19:58].

5. Flash כמודל עבודה

Pichai מסביר את ההשקעה בסדרת Flash דרך המשימה של Google להפוך את הטכנולוגיה לנגישה לכולם, ודרך הצורך לשרת מיליארדי משתמשים ב-Search וב-Gemini [21:30]. הנתון הבולט הוא אנקדוטלי: "What I've heard anecdotally from a lot of CIOs who are so concerned about how much their companies are blowing through budgets". הוא מעריך שהבעיה תחריף במהלך השנה [21:30], [23:03].

ההיגיון ההנדסי: ב-agentic workflow המודל נקרא שוב ושוב, ולכן מודל מהיר ויעיל חשוב במיוחד. "even" [23:03] "Flash is remarkably cost efficient for token use, accounting for token use, Flash is remarkably cost efficient". במקום אחר מוזכר דמו של 3.5 Flash ב-Antigravity שרץ ב-800 tokens לשנייה [27:43].

Sundar Pichai [23:03] We are using it as a blend of Pro and Flash, right, internally. And I think — most companies should learn to use it that way

שימו לב המונח המקורי "Flash-Lite" שובש בתמליל ל-"flashlight" [15:23] ו-"3.4" [29:14] "flash" הוא כנראה 3.5 Flash. מוזכר גם "Omni model" שאפשר היה לבנות טוב יותר אם לא היה מחסור ב-compute, בלי פירוט נוסף.

6. מרוץ, שיפור עצמי ו-compute

המראיין שואל אם הכול מסתכם במרוץ ל-AI שמשפר את עצמו. Pichai דוחה את המסגור הזה: "we all need to be careful to avoid this race condition at all costs" [24:33]. לדבריו, ההובלה בין המעבדות מתחלפת מהר כי לכל אחת קצב שונה של מחזורי pre-training, ו-"All this creates that perception gap which shifts widely in like four to six weeks". הוא מוסיף שכמה מעבדות נמצאות ב-frontier ואחריהן יש פער גדול [26:06]. אם recursive self-improvement יגיע, יהיה צורך להתמודד איתו כמו עם רגע הסייבר, בזהירות רבה יותר מהיום, כי "the more the more".

AI becomes advanced the more it's a societal conversation versus a single company conversation" [26:06].

לגבי compute: לפי Pichai, Google קיבלה החלטות נועזות ונכונות להשקעה ב-compute, ובכל זאת כל מנהל מצטער שלא השקיע יותר. גם העלויות עולות: "like for a given budget, you may be getting less compute than you had previously planned for". לפי התמליל, מחירי הזיכרון הם חלק מזה [27:43]. התכנון לטווח ארוך ל-Cloud נעשה בנפרד מהצרכים הפנימיים. המראיין מצטט אדם בשם Thomas Kurian, כנראה Thomas Kurian (השם לא אומת), שאמר "[26:06] we planned really well".

— Sundar Pichai [30:45] I think there are systemic bottlenecks across all layers of the stack now

כשנשאל מה צוואר הבקבוק המרכזי, Pichai מונה רישוי ובנייה של מרכזי נתונים, אספקת חשמל ורכיבי הליבה של המערכות. לדבריו, ברגע שפותרים צוואר בקבוק אחד, אחר תופס את מקומו [30:45]. יש לכך השלכה גם על המוצר: ההחלטה אם לבנות מודל בגודל Ultra תלויה גם בשאלה לכמה משתמשים אפשר יהיה לתת אותו [29:14].

7. מה עושים עם זה

- **לבנות כך שאפשר להחליף מודלים:** שכבת האפליקציה לא צריכה להיות קשורה לספק או למודל מסוים. זו ההמלצה הכי ישירה בשיחה [19:58].
- **לשלב מודל גדול עם מודל מהיר:** לשלוח את הקריאות החוזרות ב-agentic workflow ל-Flash, ולהשאיר ל-Pro את מה שבאמת דורש אותו [23:03].
- **לפתוח הרשאות של agents בהדרגה:** להתחיל במערכות הפנימיות ורק אחר כך להוסיף MCP, browser use, computer use-ו כמו שמתואר ב-[3:09] Gemini Spark.
- **לבחון אוטומציה של תיקון פגיעויות:** לעקוב אחרי מועד הזמינות החיצונית של CodeMender ולבדוק תהליך מלא של זיהוי, patch, בדיקה ו-[9:19] deploy.
- **לתכנן תקציב compute מתוך הנחה שהמחירים יעלו:** Pichai עצמו מעריך שבעיית התקציב תחריף השנה [27:43], [23:03].

שאלה פתוחה	מה נאמר	מה חסר
מתי CodeMender יהיה זמין מחוץ ל-Google?	"which we're building to share externally" [9:19]	תאריך, תמחור, דרך שילוב עם Wiz
איפה עובר הסף לעצירת שחרור של מודל?	1-2% מול קפיצות של 20% [12:20]	מדד או benchmark מוגדר
כמה חוסכים בפועל עם Flash?	"remarkably cost efficient" [23:03]	מחירים והשוואה מול Pro
מה גודל הפער בין ביקוש ל-compute?	"Absolutely" [29:14]	כל נתון כמותי
האם תלות במודלים סיניים מסוכנת?	"too early to predict" [19:58]	קריטריונים לבדיקת מודל פתוח
מתי Gemini Spark ייפתח ל-MCP ולצד שלישי?	אחרי משוב ושיפור [3:09]	לוח זמנים

8. מילון מונחים

מונח	פירוש
agent / agentic workflow	מערכת AI שמבצעת רצף פעולות באופן עצמאי בשביל המשתמש, ותהליך עבודה שמבוסס עליה
MCP	Model Context Protocol, פרוטוקול שמחבר מודלים לכלים ולמקורות מידע של צד שלישי
computer use / browser use	יכולת של agent להפעיל מחשב או דפדפן כמו משתמש אנושי
frontier	רמת היכולת הגבוהה ביותר הקיימת כרגע במודלים
CodeMender	כלי של Google שמזהה פגיעויות, מייצר patches, בודק ומפיץ אותם, ורץ ברציפות
Project Zero	צוות המחקר של Google לאיתור פגיעויות, שנותן לספקים 90 יום לתקן לפני חשיפה
Zero Trust	מודל אבטחה שבו אין אמון ברירת מחדל באף רכיב ברשת
SynthID / watermarking	סימון של תוכן שנוצר ב-AI כדי שאפשר יהיה לזהות אותו

פירוש	מונח
משפחת המודלים הפתוחים של Google, שהגרסה האחרונה שלה היא Gemma 4	Gemma
דרגות של מודלי Gemini: מודל מהיר וזול, מודל חזק, ומודל גדול במיוחד	Flash / Pro / Ultra
הוצאות הון, כמו השקעה במרכזי נתונים ובחומרה	CapEx
הרצת מודל מאומן כדי לייצר תשובות	inference
שלב האימון הראשוני והכבד של מודל	pre-training
מצב שבו מערכת AI משפרת את עצמה בלולאה חוזרת	recursive self-improvement
שבב ה-AI שמפתחת Google	TPU
כאן במובן של דינמיקת מרוץ מסוכנת בין המעבדות	race condition