

scale, ChatGPT, Codex intelligence-ו-utility-כ

שיחה בקורס systems בסטנפורד על הסיבה שהימור על scale ממשיך להשתלם, איך
נולדו ChatGPT ו-Codex, ולמה העתיד הוא inference זול

Sam Altman עם מנחה הקורס (שמו לא נאמר בתמלול) · Stanford, לפי הלוגו בווידיאו · המקור המקורי לא זוהה; פורסם
מחדש ב-X ע"י Hayyan | Creative AI · 21.08.2026

משך: כ-27 דקות · X

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זו שיחה של כ-27 דקות בין Sam Altman, מנכ"ל OpenAI, לבין מנחה של קורס systems בסטנפורד. על המסמך מופיע לוגו Stanford, ושם המנחה לא נאמר בתמלול. הווידאו הגיע מ-repost ב-X, והכיתוב שם מייחס ל-Altman את המשפט "You no longer need to write prompts" ומתאר הרצאה של 38 דקות. שני הפרטים לא מופיעים בתמלול, ולכן לא השתמשנו בהם. זו שיחת כורסה בלי שקפים, עם כמה שאלות קהל בסוף.

הטענה המרכזית של Altman היא שה-scale מייצר emergent properties ותשואות שה-consensus לא צופה, ושרוב האנשים לא דוחפים אותו מספיק. את זה הוא מדגים בסיפור של YC, ב-scaling של מודלים ובלידה של ChatGPT ו-Codex. משם הוא עובר לחזון שלפיו intelligence תהפוך ל-utility, וממליץ להשקיע ב-inference זול. בסוף הוא דוחה את הטענה שמיוחסת ל-Yann LeCun, שלפיה LLMs הם מבוי סתום.

איך לקרוא את המסמך

- 1. **תקציר מנהלים:** הטענות המרכזיות ומה לא נאמר.
- 2. **עידן ה-automated coding:** מה אפשר לבנות היום עם תקציב tokens.
- 3. **Scale כתכונה של מערכת:** YC, ריצות האימון והצד האנושי.
- 4. **ChatGPT ו-Codex:** איך שני ה-killer apps התגלו בפועל.
- 5. **ה-pipeline והחוקר האוטומטי:** היעדים ל-September ול-March 2028.
- 6. **Intelligence כ-utility:** האנלוגיה לחשמל, וההבדל בין tokens ל-compute.
- 7. **One-person frontier lab ושאלות קהל:** inference, ה-Erdős problem ו-LeCun.
- 8. **מה עושים עם זה:** פעולות ושאלות פתוחות. אחריו מילון מונחים.

הערות מקור המסמך הופק אוטומטית מהתמלול של ההקלטה ומהפריימים שלה. בווידאו אין שקפים, ורואים בו רק את הדוברים. ה-ASR עלול לשבש שמות ומונחים: למשל "GPP4" במקום "GPT-4", "Jan Lecun"/"Lacuna" במקום Yann LeCun, ומספר קורס משובש ("Cardlean 5x183"). הציטוטים מועתקים מילה במילה ובבדקו מול חותמת הזמן שלהם. לא הוספנו שום מידע ממקורות חיצוניים.



שני הדוברים על הבמה בכיתה בסטנפורד: מנחה הקורס (משמאל) ו-Sam Altman (מימין)

1. תקציר מנהלים

Altman מציג תזה אחת שחוזרת לאורך כל השיחה: כשמשוהו כבר עובד בקטן, כדאי לדחוף אותו ל-scale שאף אחד עוד לא ניסה. רוב האנשים, כולל חוקרים מבריקים, לא עושים את זה מספיק. מכאן הוא גוזר את סיפורי המוצר של OpenAI ואת התחזיות שלו להמשך.

- **Startup** קטן יכול היום לעשות עבודה של צוות של מאה: עם "affordable amount of spend on tokens" אפשר להגיע לתפוקה של צוות הנדסה של מאה איש, וזה משנה את רמת השאפתנות האפשרית [0:00].
- **Scale** מייצר דברים שלא קיימים בקטן: הדוגמה היא אפקט הרשת בתוך batch של YC, שהתגלה רק כשימנו הרבה חברות בבת אחת [3:06].
- בדרך ל-scale דברים נשברים, ואת השבירה מפרקים לבעיות: טכנולוגיה (ריצה על 10,000–100,000 GPUs), הון ותרבות מחקרית. כל בעיה מטופלת בנפרד [6:13].
- **ChatGPT** נולד כ-research demo: ל-GPT-3 API היה run rate של אולי 10–20 מיליון דולר. מפתחים השתמשו ב-API key רק כדי לצ'וטט, ו-OpenAI בנתה chatbot סביב 3.5 [10:46][12:17].
- **Codex** היה התוכנית המקורית: לפי האמונה הפנימית של OpenAI, קוד הוא ה-actuator של המודל במחשב. ה-inflection point הגיע עם 5.5 [13:47].
- **יעד פנימי: חוקר AI אוטונומי:** עד September השנה, "AI research intern" על 500,000 GPUs בשווה ערך A100. עד March 2028, חוקר מלא שממציא ארכיטקטורות [15:21].
- **Intelligence** תהיה utility, ואת האנלוגיה הנכונה עוד לא מצאו: כמו שחברות החשמל מכרו "light at night" ולא חשמל [18:27].
- **Inference** הוא הצוואר בקבוק שלא מושקע בו מספיק: לדעת Altman כל ה-frontier labs יצטרכו להפוך ל-inference companies [21:31].
- **להמר נגד LLM scaling זו טעות:** הראיה שהוא מביא: מודל של OpenAI הפריך השערה מה-Erdős problems [23:03][24:35].

מה לא נאמר

- **אין תאריך לשיחה ואין שם למנחה:** מהתמלול אפשר רק להסיק שהשיחה התקיימה ב-2026 ("by September of" "this year", "early this year").
- **אין פרטים על ה-Erdős problem:** לא נאמר איזו בעיה, איזה מודל, ומי אימת את ההפרכה.
- **"5.5" לא מוסבר:** הכוונה כנראה לגרסת מודל, אבל לא נאמר בדיוק מה ומתי.
- **אין מספרי הכנסות עדכניים:** המספר היחיד הוא run rate היסטורי של ה-API, "10 or a \$20 million".
- **אין הסבר ל-Altman scale:** אומר במפורש: "I offer no theory that I find satisfying to explain it".
- **אין עצות פרומפטינג:** המשפט "You no longer need to write prompts" מהכיתוב ב-X לא מופיע בתמלול.

2. עידן ה-automated coding: אילו בעיות לתקוף

השיחה נפתחת באמצע תשובה. לדברי Altman, מה שהשתנה הוא מרחב האפשרויות של startup: מה אפשר לקחת על עצמך, באיזו מהירות זזים וכמה דברים עושים במקביל. כשהמנחה שואל איזה תרגיל היה נותן לסטודנטים, הוא עונה שלחלק בעיות לא עובד, כי רעיון שברור לו כנראה ברור להרבה אחרים. מה שצריך לחפש זה המקבילה של OpenAI בהתחלה: אחד מאולי ארבעה מאמצי AGI בעולם (ה-ASR כתב "ADI"). כלומר שוק עתידי של טריליונים שרק ארבע חברות עובדות עליו כרגע.

Sam Altman [0:00] You, with like an affordable amount of spend on tokens, you can do what a — hundred person, incredibly great engineering team would do as a startup

3. Scale כתכונה של מערכת

המנחה מזכיר ש-Altman כבר דיבר על scale בקורס ב-2014, ושואל איך הוא מפרק את זה היום כ-attribute של systems design. Altman מסייג שאין לו תיאוריה. הטענה שלו אמפירית: כל הדברים המעניינים שראה בקריירה קשורים ל-scale.

הדוגמה המרכזית היא YC. אנשים חכמים לחצו לקטין את ה-batches חזרה לעשר חברות. ההיגיון היה שהחברות הטובות ממילא בולטות, ושאר החברות במימון לא תורמות. בפועל אפקט הרשת בתוך ה-batch היה חלק מרכזי מהקסם, ושום דבר כזה לא היה קיים בעשירית או במאית מה-scale.

— **Sam Altman [1:31]** all of the most interesting ones have had something to do with emergent properties at scale or scale continuing to provide returns far beyond what the consensus thinks will work.

למה לא עושים את זה יותר? כי "stuff breaks uh at an accelerating rate and in an unpredictable way as you scale it" [4:42]. ב-scaling של המודלים היו שלושה סוגי התנגדות. הראשון טכני: אף אחד לא חשב על ריצה על פני 10,000 או 100,000 GPUs. השני הוני: איך יהיה לזה business. השלישי תרבותי: חוקרים שאלו למה לשים את כל ה-compute על פרויקט אחד במקום לפזר אותו על הרבה פרויקטים. הפתרון של Altman הוא לפרק את המכלול לסיבות נפרדות ולטפל בכל אחת לחוד [6:13].

לגבי הצד האנושי, Altman מונה מה צריך: מטרה ברורה, תוכנית ברורה ודרך ברורה לקבל החלטות. כשהחליטו להמר על scaling של deep learning, ההימור עצמו ("If we're wrong, we'll fail") היה כלי ארגוני חזק. הקושי העמוק יותר הוא שבני אדם לא מבינים exponentials באופן אינטואיטיבי, ולכן צריך לשבת עם אנשים ולנמק מ-first principles [7:44].

— **Sam Altman [7:44]** we did not evolve to be good at thinking about exponentials

4. ChatGPT ו-Codex: איך נמצאו ה-killer apps

המנחה מציג את ChatGPT כ-killer app לצרכנים ואת הקוד כ-killer app ל-enterprise. Altman מפרט את ההיסטוריה. ל-OpenAI היה GPT-3 בלי מוצר, והיא הייתה צריכה הכנסות כדי לבנות מחשבים של מיליארדים. לכן הוציאו API בקיץ 2020, ואמרו לעצמם שמישהו אחר ימצא את המוצר. חודש אחרי זה ה-API התפוצץ ב-Twitter, אבל העסק היחיד שעבד באמת היה copywriting. מה שבלט בנתונים הוא שמפתחים השתמשו ב-API key רק כדי לצ'וט עם המודל [10:46].

— **Sam Altman [10:46]** If you go back and use GPT-3 or 3.5, you will be astonished at how bad the models were then relative to the amount of excitement they generated at the time

GPT-4 כבר היה מוכן אצלם (ה-ASR כתב "GPP4"), אבל בינתיים היה מודל 3.5 עם post-training חדש ל-instruction following. ה-API היה אז "like a 10 or a \$20 million run rate kind of business". לפי עקרון YC, "see what your users love", הם בנו chatbot. המטרה הייתה research demo שישכנע חברות אחרות לבנות מוצרי chat ולשלם על ה-API [12:17]. במשך חמישה ימים התנועה זינקה ונפלה וחזרה לשיא גבוה יותר. ביום החמישי Altman כינס את כולם:

— **Sam Altman [12:17]** this is an emergency. This is a good kind of emergency, but we have to build a company and a product all at once

משם הגיעו חודשיים של scaling מטורף. התמחור נקבע בהתחלה רק כדי לא לשרוף את חשבון ה-compute, ו"That also turned out just to work" [13:47]. הלקח מ-YC שהוא מצטט: "when something really starts growing and it's not very good, you have like a guaranteed hit on your hands" [12:17].

לגבי Codex, התוכנית לפני ChatGPT הייתה ללכת all-in על קוד. ההנחה הפנימית: קוד הוא הדרך של המודל לשלוט במחשבים, ורובוטים הם הדרך שלו לשלוט בעולם הפיזי. Codex השתפר מאוד בתחילת השנה, אבל "5.5 is when we saw this real inflection point" [13:47].

5. ה-pipeline והיעד: חוקר AI אוטומטי

המנחה שואל אם pipeline היכולות המקובל (pre-training, mid-training, post-training) וולולא RL supervised feedback-וי יישאר יציב. Altman עונה: "I think that is definitely the current pipeline. I expect we will go through a major rewrite that's a research problem for the AIs to figure out" [15:21].

Sam Altman [15:21] we've set this goal that by September of this year, we will use 500,000 A100 — equivalent GPUs, like a lot of computing power, as an AI research intern. And by March of 2028, that we will have a full end-to-end, very talented researcher, like figuring out complete new architectures

שימו לב אלה יעדים פנימיים ש-OpenAI הציבה לעצמה, לא הישגים. בשיחה אין שום מדד שמגדיר מתי "intern" או "researcher" נחשב כמי שהושג, ואין דיווח על התקדמות לקראת היעדים.

6. Intelligence כ-utility: האנלוגיה לחשמל

המנחה שואל על הגבולות של אנלוגיות כמו "AI intern" כשהן יוצאות מ-Silicon Valley לעולם. Altman אומר ש-OpenAI יוצרת utility חדש, ויש מעט מאוד דוגמאות כאלה להשוות אליהן: חשמל, אינטרנט, מים. חברות החשמל גילו שאי אפשר למכור "חשמל". זה נשמע מסוכן ולא ברור, ומה שעבד בסוף היה למכור אור בלילה. ההבטחה "it'll wash your clothes for you someday" הייתה קפיצה גדולה מדי בשביל הלקוחות [16:53].

Sam Altman [18:27] getting from us is not electricity. It's light at night —

החזון שהוא מתאר: "you will have like a OpenAI token subscription that you will plug into everything", כש-agent רץ בשבילך כל הזמן. מכירה של "intelligence" לא מתחברת לאנשים, ו-Altman מודה שעוד לא מצאו את המקבילה ל-"light at night" [18:27].

המנחה מזכיר ש-Jensen (שם פרטי בלבד בתמלול) תיאר ב-session קודם של הקורס דווקא את ה-compute כ-utility. Altman משיב שצרכנים ועסקים יחשבו במונחי tokens, או אפילו רמה אחת מעל. השאלות שלהם יהיו אם אפשר להשתמש בזה הרבה, אם זה זול ואם זה עושה עבודה טובה. החומרה תעבור abstraction, כמו שבחשבון הסולר משלמים על airtime וגיגה בלי לחשוב על ה-base station [20:00].

7. One-person frontier lab, inference ושאלות קהל

פרויקט הגמר של הקורס הוא "one-person frontier lab". כל סטודנט מדמה מעבדה של אדם אחד, עם credits של מאות אלפי דולרים מ-Cloudflare ואולי גם tokens של Altman. OpenAI נשאל על מה היה עובד אילו היה בכיתה, ועונה inference. מודלים מצוינים יגיעו בכל מקרה, "I promise you", אבל מה שחסר הוא היכולת לספק intelligence זולה בכמויות ענק [21:31].

Sam Altman [21:31] But I think we have not invested enough in being able to deliver at scale huge amounts of cheap intelligence —

בשאלות הקהל נשאל Altman על הטענה של Yann LeCun ש-LLMs הם מבוי סתום. לדבריו המודלים כבר עוקפים בני אדם בחלק מהמשימות וגרועים מהם בהרבה באחרות, במיוחד במשימות long-horizon שדורשות שיקול דעת. כדוגמה הוא מביא ש-"yesterday, we had one of our models discover, disprove a conjecture, one of the Erdős problems" [23:03]. לדבריו מתמטיקאים שואלים עכשיו אם המתמטיקה נגמרה. הוא מסכים ש-world models חשובים, למשל לרובוטיקה, אבל:

Sam Altman [24:35] I think the field was honestly held back by a generation of scientists who just — were way too certain on what scaling was not going to produce

על השאלה אם מעצבן להיות ה-"Altman", I told you so guy עונה שכבר לא מטרידים אותו הטרוולים שקוראים לחברה "fraud". לדעתו מי שבונה את הזהות שלו סביב אמונה שמהו יעבוד או לא יעבוד לא מסוגל לשחרר

כשהנתונים מפריכים אותו, "You can't see the truth". הוא מציין שהתזכורת הזו נכונה לשני הכיוונים [24:35]
[26:24].

8. מה עושים עם זה

- **בחנו מחדש את גבולות השאפתנות:** אם תקציב tokens נותן תפוקה של צוות של מאה (לפי Altman), כדאי לשקול מחדש פרויקטים שנפסלו בגלל כוח אדם.
- **חפשו מה עובד בקטן ונסו להגדיל פי עשרה:** זה העיקרון שחוזר בכל הדוגמאות. כשדברים נשברים, פרקו את השבירה לבעיות טכניות, הוניות ותרבותיות.
- **הסתכלו מה המשתמשים עושים בפועל:** ChatGPT נולד מכך שמפתחים השתמשו ב-API בצורה שלא תוכננה.
- **למי שבונה תשתית: inference:** לפי Altman זה המקום שבו חסרה השקעה. מכאן גם הציפייה שה-labs יהפכו ל-inference companies.
- **תמחרו לפי ערך ולא לפי טכנולוגיה:** השיעור של "light at night" רלוונטי לכל מי שמוכר AI ללקוחות שאינם טכנים.

שאלה פתוחה	למה זה חשוב
מה יחליף את ה-pipeline של pre/mid/post-training ו-RL?	Altman צופה "major rewrite" אבל משאיר את זה "for the Als to figure out".
איך יימדד ה-"AI research intern" של September?	בלי הגדרה ברורה אי אפשר לבדוק את היעד.
איזו ה-Erdős problem הופרכה, ומי אימת את זה?	זו הראיה המרכזית נגד LeCun, אבל אין עליה פרטים.
מה יהיה ה-"light at night" של AI?	Altman עצמו אומר שאין לו תשובה.
האם ה-utility יתומחר ב-tokens או רמה אחת מעל?	זה משפיע ישירות על מודלים עסקיים של מוצרים שבנויים על LLMs.

מילון מונחים

מונח	משמעות
Scaling laws	הקשר האמפירי בין גודל המודל, הדאטה וה-compute לבין הביצועים
Emergent properties	תכונות שמופיעות רק מעל scale מסוים ולא קיימות בקטן
Batch (YC)	מחזור של startups ש-Y Combinator מממנת ומלווה יחד
Post-training	שלב אחרי pre-training שמלמד את המודל למשל instruction following
Pre-/Mid-training	שלבי האימון הראשוניים של המודל על דאטה רחב
RL	Reinforcement learning, לולאת למידה מתגמול
Instruction following	היכולת של מודל לבצע הוראות בצורה שמקלה לשוחח איתו
Run rate	הכנסה שנתי מוערכת לפי קצב ההכנסות הנוכחי
Inference	הרצת מודל מאומן כדי להפיק תשובות, השלב שמספק את ה-tokens
Utility	שירות תשתיתי בסיסי כמו חשמל או אינטרנט
Tokens	יחידות הטקסט שמודל מעבד, והבסיס לתמחור
A100 equivalent	יחידת מדידה של compute לפי שבב A100 של NVIDIA
Frontier lab	מעבדה שמפתחת את המודלים המתקדמים ביותר
World models	מודלים שמייצגים את העולם הפיזי, למשל לרובוטיקה
Erdős problems	אוסף בעיות פתוחות במתמטיקה שקרויות על שם Paul Erdős