

איך אני משתמש במודלי שפה

סיכום ההרצאה "How I use LLMs" של אנדריי קרפתי – המודל המנטלי, הכלים, ומה
מהם עדיין רלוונטי

אנדריי קרפתי · הרצאת וידאו ביוטיוב, פברואר 2025

נצפה כהעלאה מחדש ב-X (@0xnicc0, 5.10.2026) · 2:07:32

1. על המסמך

המסמך מסכם את "How I use LLMs" של אנדריי קרפתי (לשעבר ב-OpenAI ובטסלה): סיור מעשי של כשעתיים בצ'אטבוטים, בכלים שלהם ובאופן שבו הוא עצמו משתמש בהם ביום-יום. זו ההרצאה השנייה בסדרה. הראשונה, "Deep Dive into LLMs", הסבירה איך המודלים נבנים, וזו מסבירה איך עובדים איתם.

הזמנים בסוגריים [ש:דד] מתייחסים לגרסה שהועלתה ב-X, ולא לסרטון המקורי ביוטיוב.

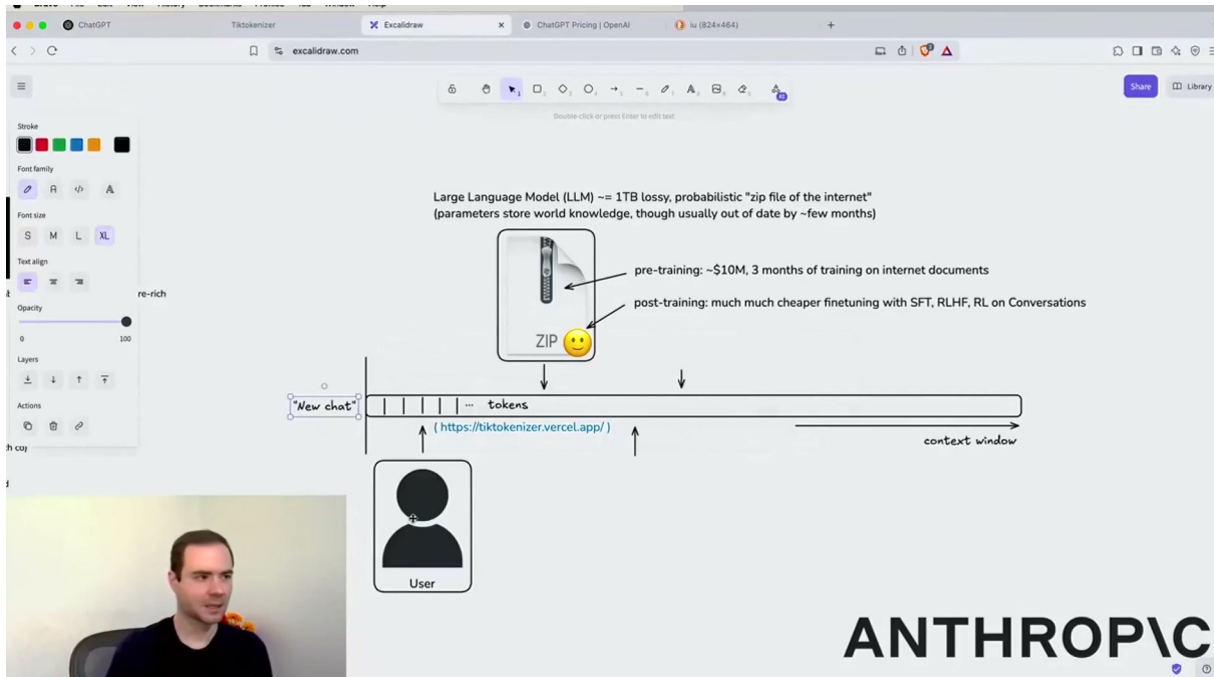
על המקור: זה לא "השבוע שעבר" הפוסט ב-X מציג את ההרצאה כאילו קרפתי פרסם אותה בשבוע שעבר. זה לא נכון: מדובר בסרטון מפברואר 2025. התוכן מסגיר את התאריך. GPT-4o הוא "the flagship biggest model that's" [12:28] Claude 3.7 Sonnet, currently available as of today" יצא [45:02] "while I was filming this video", ו-Grok 3 "just released" [15:34]. בגרסת X נחתכו כמה הדקות הראשונות (הסקירה של שוק האפליקציות), ובפינת המסך הודבק לוגו ANTHROPIC גדול שאינו קשור לתוכן. מעבר לשימוש ב-Claude כאחת האפליקציות שהוא משווה, ההרצאה לא מזכירה שום קשר לחברה.

מה התיישן הרעיונות בהרצאה עדיין תקפים: המודל כ"קובץ zip", חלון ההקשר, מתי להשתמש במודל חושב ומתי בכלי חיפוש. שמות המודלים, המחירים והמגבלות לכל מסלול מתארים את פברואר 2025 בלבד. בפרק 9 ריכזתי מה מהם כבר לא נכון.

2. תקציר מנהלים

- **אתם מדברים עם קובץ zip:** המודל הוא דחיסה "אבודה" והסתברותית של האינטרנט (כטרה-בייט, כטריליון פרמטרים). הידע שלו מעורפל ומעודכן רק עד מועד האימון. האישיות שלו נוספה באימון שאחרי (post-training).
- **חלון ההקשר הוא הזיכרון העובד:** מה שבחלון זמין למודל באופן ישיר. לכן פותחים צ'אט חדש בכל החלפת נושא: טוקנים לא רלוונטיים מסיחים אותו, מאטים אותו ומייקרים את התשובה.
- **דעו באיזה מודל אתם משתמשים:** מודלים גדולים יודעים יותר וכותבים טוב יותר, ומודלים קטנים מזיפים יותר. כדאי לשלם על המודל הטוב אם משתמשים בו מקצועית, ולשאול כמה מודלים במקביל: "מועצת LLM".
- **מודלים חושבים לבעיות קשות:** מודלים שאומנו עם RL חושבים דקות ומשיגים דיוק גבוה יותר במתמטיקה ובקוד. לשאלות פשוטות הם מיותרים.
- **כלים הם מה שמשנה את התמונה:** חיפוש באינטרנט למידע עדכני, Deep Research לדוחות ארוכים, העלאת קבצים לקריאה משותפת, ומפרש Python לחישובים ולגרפים. אפליקציה שאין בה כלי מתאים פשוט תמציא תשובה.
- **קוד ברצינות עושים מחוץ לצ'אט:** Cursor ודומיו עובדים על הקבצים במחשב. "vibe coding" זה לתת לסוכן לבצע ולבדוק את התוצאה.
- **קול, תמונה ווידאו:** קרפתי מדבר עם המודלים בערך בחצי מהמקרים, ובטלפון ב-80% מהם. הוא מבחין בין קול "מזויף" (המרה לטקסט וחזרה) לבין קול "אמיתי" (טוקנים של אודיו בתוך המודל).
- **כל תוצר הוא טיוטה ראשונה:** המסר חוזר לאורך כל ההרצאה: בודקים מקורות, קוראים את הקוד, ומאמתים מספרים.

3. המודל המנטלי: עם מה אתם מדברים



התרשים שקרפתי בונה לאורך ההרצאה: המשתמש כותב טוקנים לחלון ההקשר, והמודל ("קובץ zip" עם פראוק מחיך) עונה בטוקנים משלו

כשנלוחצים "צ'אט חדש", מתחיל רצף טוקנים חד-ממדי. אנחנו כותבים, המודל עונה, ופולט טוקן מיוחד שאומר "סיימתי". יחד אנחנו בונים את חלון ההקשר (context window), שהוא הזיכרון העובד של השיחה [0:00].

המודל עצמו נבנה בשני שלבים. האימון המקדים (pre-training) דוחס את כל האינטרנט לפרמטרים של רשת נוירונים. התהליך יקר ("עשרות מיליוני דולרים", כשלושה חודשים) ולכן נדיר, ומכאן שהמודל תמיד קצת לא מעודכן (knowledge cutoff). אחריו בא האימון שאחרי (post-training), שמחליף את הדאטה בשיחות שבני אדם כתבו, וכך המודל לובש את הפרסונה של עוזר [3:07–1:33].

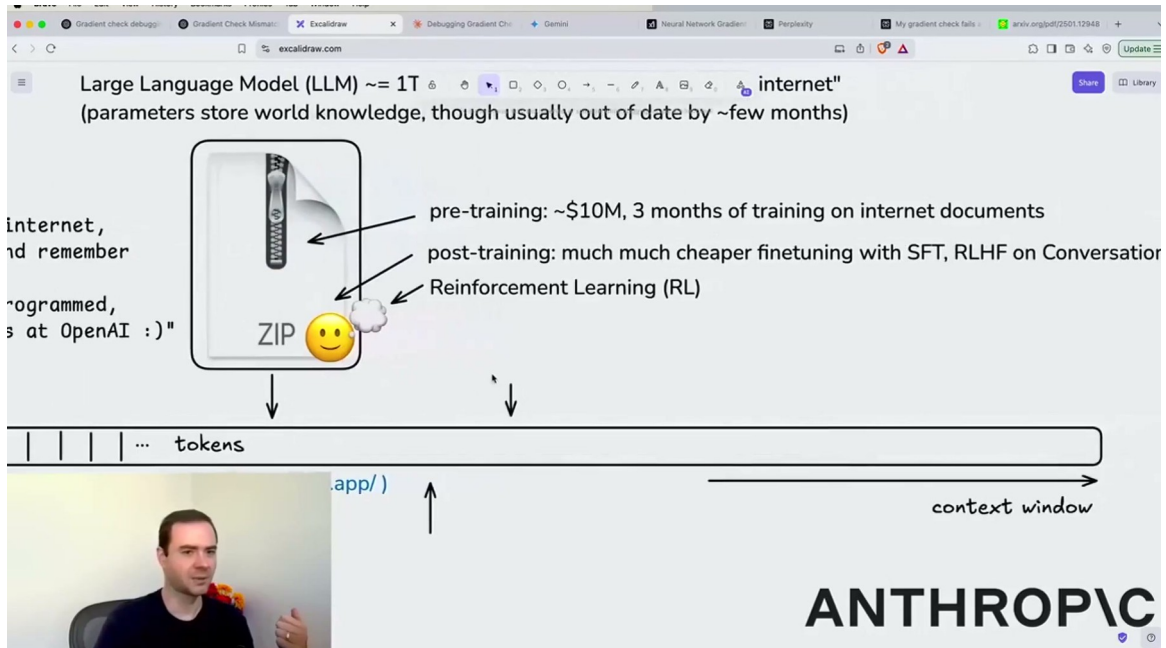
— קרפתי, [4:42] Hi, I'm ChatGPT I'm a one terabyte zip file My knowledge comes from the internet which I read in its entirety about six months ago and I only remember vaguely

המשמעות המעשית: על דברים שמופיעים באינטרנט שוב ושוב המודל יזכור טוב, ועל דברים נדירים או חדשים פחות. קרפתי שאל כמה קפאין יש באמריקנו (כ-63 מ"ג), ואיזו תרופה עוזרת לנזלת, ואז אימת מול הקופסה עצמה [6:15–9:23]. אלה שאלות נפוצות, לא עדכניות, ובלי סיכון גבוה, ולכן מתאימות לזיכרון של המודל.

שני הרגלים בסיסיים

- **צ'אט חדש בכל החלפת נושא:** "think of the tokens in the context window as a precious resource" [10:56]. טוקנים לא רלוונטיים מסיחים את המודל, מורידים את הדיוק ומייקרים כל טוקן נוסף.
- **דעו מה המודל ומה המסלול:** בלי התחברות, לא ברור עם איזה מודל מדברים, וייתכן שזה מודל קטן ו"טיפש" יותר. המסלולים (נכון לפברואר 2025): חינם, Plus ב-20 דולר ו-Pro ב-200 דולר לחודש. קרפתי משלם על הרבה מסלולים, כי "in my professional work... this is, still very cheap for me" [14:01]. מועצת המודלים: קרפתי שואל כמה מודלים את אותה שאלה. כשביקש המלצה לעיר לחופשה, גם Claude וגם Gemini המליצו על צרמט בשווייץ, והוא נסע לשם [14:01]. I kind of refer to all these models as my LLM" [15:34].

4. מודלים חושבים



השלב השלישי באימון: Reinforcement Learning. המודל מגלה בעצמו אסטרטגיות חשיבה בזמן שהוא מתרגל על בעיות

בשלב ה-RL, המודל מתרגל על הרבה בעיות מתמטיקה וקוד ומגלה דרכי חשיבה שמובילות לתוצאה טובה: לנסות רעיונות, לחזור אחורה ולבדוק הנחות. את האסטרטגיות האלה קשה לכתוב ידנית, ולכן רק RL מגלה אותן. קרפתי מתאר את השלב כפריצת דרך יחסית חדשה, והמאמר של DeepSeek הוא הראשון שדיבר עליו בפומבי [15:34–17:08].

My gradient check doesn't work can you see why?

I'll analyze your gradient check code to identify why it's failing. Let me trace through your implementation step by step.

I've found an issue in your gradient computation. The problem is with how you're handling the gradients for each sample in your loss function. Let me explain:

Analysis >

Let me examine your gradient calculation more carefully:

Analysis >

I've pinpointed the issue in your implementation! The problem is in how you're packing the gradients. Here's what's happening:

1. Your parameters are structured as: [weight1, bias1, output_weight1, weight2, bias2, output_weight2, ...]
2. But when calculating gradients for each data point, you're packing them as:

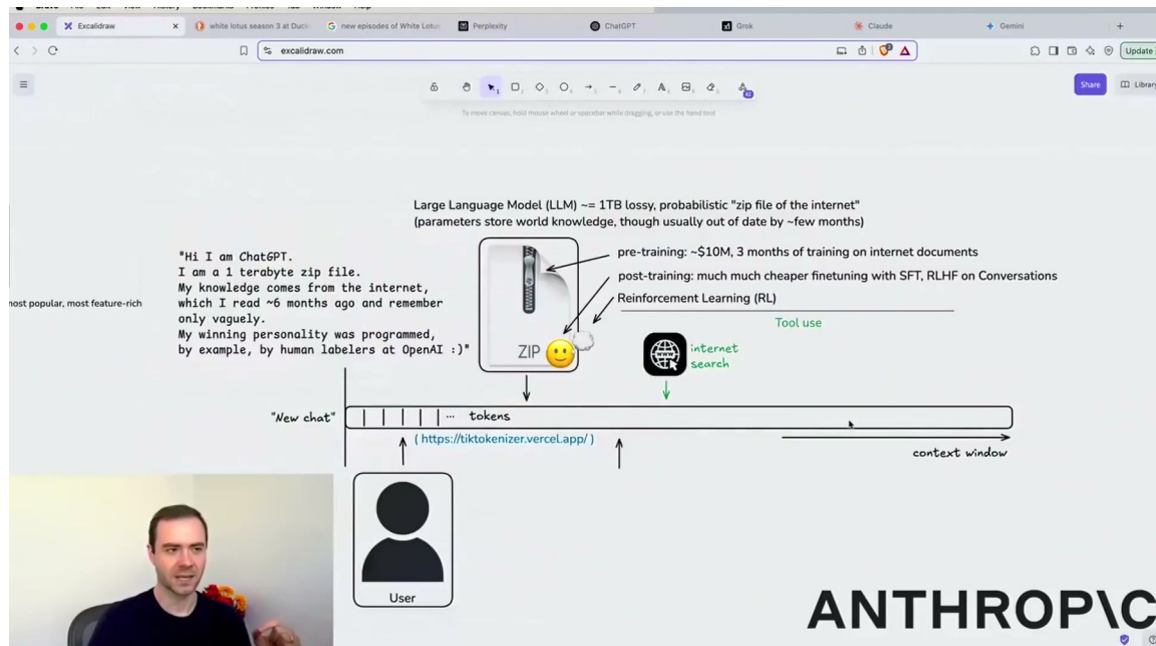
```
def loss_and_grad(params):  
    # unpack the parameters  
    num_params = len(params)  
    assert num_params % 3 == 0  
    n_hidden = num_params // 3  
    weights = params[1::3]  
    biases = params[2::3]  
    output_weights = params[3::3]  
    sigmoid = lambda x: 1 / (1 + exp(-x))  
    loss = 0.0  
    grads = [0.0 for _ in params]  
    for i in range(N):  
        x1, y1 = x[i], y[i]  
        # forward pass the network for this datapoint  
        hidden = (sigmoid(w * x1 + b) for w, b in zip(weights, biases))  
        output = sum(w * h for w, h in zip(output_weights, hidden))  
        # calculate the loss  
        loss += 1/N * (output - y1)**2  
        # calculate the gradients  
        doutput = 2/N * (output - y1)  
        doutput_weights = [doutput * h for h in hidden]  
        dhidden = [doutput * w for w in output_weights]  
        dhidden = [dhidden[i] * (1 - hidden[i]) for i in range(len(hidden))]  
        dweights = [dhidden[i] * x1 for i in range(len(hidden))]  
        dbiases = [dhidden[i] for i in range(len(hidden))]  
        # DEBUG: Correctly pack gradients to match parameter structure  
        for j in range(n_hidden):  
            grads[j*3] += dweights[j] # weight gradients  
            grads[j*3 + 1] += dbiases[j] # bias gradients  
            grads[j*3 + 2] += doutput_weights[j] # output weight gradients  
    return loss, grads
```

הדוגמה: באג בבדיקת גרדיאנט. GPT-4o לא מצא אותו, o1 Pro מצא אחרי דקה של חשיבה, ו-Claude 3.5 Sonnet מצא גם בלי חשיבה

- **מתי להשתמש:** בבעיות קשות במתמטיקה, בקוד ובהגיון. לשאלה על יעדי חופשה "there's no need to wait for one minute" [23:20].
- **השיטה שלו:** מתחיל עם מודל רגיל ומהיר. אם התשובה נראית חלשה, עובר למודל חושב [23:20].
- **איך מזהים (פברואר 2025):** ב-OpenAI, כל מודל שמתחיל ב-o1 (o1, o3-mini, o1 Pro), ב-Grok, כפתור Think; ב-Perplexity, DeepSeek R1 עם המחשבות הגולמיות [18:42–23:20].

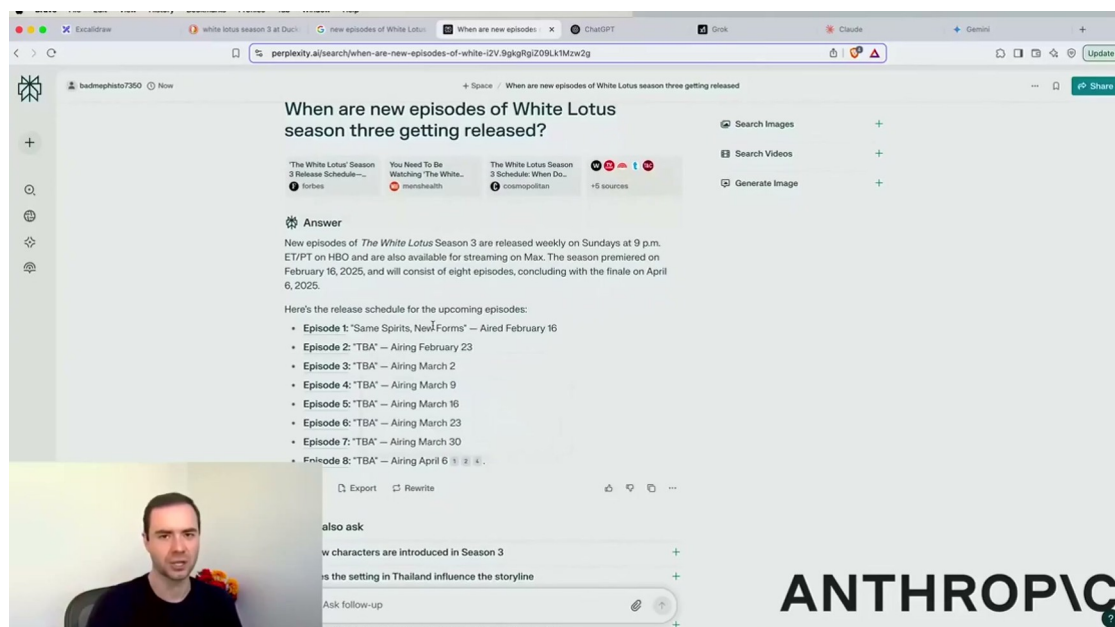
5. כלים: חיפוש, מחקר וקבצים

חיפוש באינטרנט



כלי החיפוש: המודל פולט טוקן "חפש", האפליקציה מורידה את הדפים ומכניסה את הטקסט שלהם לחלון ההקשר

כשהמודל פולט טוקן חיפוש, האפליקציה עוצרת, מחפשת, מורידה דפים ומדביקה את הטקסט שלהם לחלון ההקשר. מכאן המודל עונה מתוך הזיכרון העובד ולא מתוך הזיכרון המעורפל שלו [27:58–24:53]. הדוגמה של קרפתי: מתי יוצא הפרק הבא של White Lotus. את התשובה המודל לא יכול לדעת, כי היא נקבעה אחרי האימון.



Perplexity עונה על שאלת White Lotus עם לוח שידורים וציטוטים

מתי הוא משתמש בחיפוש: כשהתשובה נמצאת "בכמה הלינקים הראשונים בגוגל", כשמדובר במשהו עדכני או משתנה, או בנושא נישתי. הדוגמאות שלו: האם הבורסה פתוחה היום, האם Vercel מציעה PostgreSQL, למה מניית Palantir עולה, ומה העניין עם משהו שראה בטוויטר [35:45–32:39]. לחיפוש כאלה הוא נוהג להשתמש ב-Perplexity.

Deep Research

חיפוש וחשיבה שרצים במשך עשרות דקות, ובסופם דוח עם מקורות. הדוגמה: Ca-AGK, רכיב בתוסף התזונה של בריאן ג'ונסון. ChatGPT חשב חמש דקות, קרא 27 מקורות וכתב דוח על מחקרים בתולעים, בזבובים, בעכברים ובבני אדם. גם Perplexity ו-Grok הריצו את אותה שאלה, וקרפתי מצא את הדוח של ChatGPT היסודי ביותר [41:58–35:45].

anthropic

anticoagulants should exercise caution, as AGK may enhance nitric oxide production and bleeding risk.

Limitations in Current Evidence

While animal data are compelling, human trials lack rigor:

- **Sample Size:** Existing studies (e.g., Rejuvant®) are underpowered (n<50)
- **Duration:** Long-term effects beyond 12 months unknown
- **Sex-Specific Responses:** Female-biased benefits in mice may not translate to humans

Conclusion

CaAGK emerges as a multifaceted longevity therapeutic, targeting inflammaging, epigenetic dysregulation, and mitochondrial decay. Its inclusion in the Blueprint Protocol reflects a strategic emphasis on metabolic health and inflammation control—key pillars of geroprotection. While human data remain preliminary, the compound's favorable safety profile and pleiotropic mechanisms justify its use as a prophylactic agent. Future research must prioritize large-scale, placebo-controlled trials to validate morbidity compression claims and optimize dosing regimens across diverse populations.

[Citations embedded per sentence as per provided sources]

Share Export Rewrite

People also ask

- How does CaAGK compare to other supplements in the Longevity Mix
- What specific health benefits has CaAGK shown in animal models
- Are there any long-term studies on the effects of CaAGK in humans
- How does CaAGK interact with other supplements in the Longevity Mix
- What are the potential side effects of taking CaAGK regularly

Ask follow-up

ANTHROPIC

דוח Deep Research על Ca-AGK: מסקנות, מגבלות המחקר הקיים והפניות

— קרפתי, [41:58] Treat this as your first draft Treat this as papers to look at but don't take this as definitely true

ChatGPT 4o

Major LLM Research Labs in the USA

Below is a table of prominent U.S.-based labs focused on large language model (LLM) research. It includes independent organizations and corporate AI divisions, with their founding year, approximate team size, and total funding or investment (where available):

Lab	Founded	Employees (approx.)	Total Funding / Investment
OpenAI	2015	~770 (2023)	~\$13 billion (Microsoft ~49% stake)
Anthropic	2021	~500 (2024)	>\$10 billion (incl. \$4B Amazon, \$2B Google)
Google DeepMind	2010	~2,600 (2024)	~\$500 million (acquired by Google in 2014)
Meta AI (FAIR)	2013	Hundreds (internal team)	N/A (internal; Meta planning \$60–65B on AI in 2025)
Microsoft Research (AI)	1991	~1,000+ researchers	N/A (internal; ~\$10–14B annual R&D spend)
Hugging Face	2016	~170 (2023)	~\$400 million (total venture funding)
Allen Institute for AI (AIZ)	2014	~200+	~\$100 million per year (non-profit)
EleutherAI	2020	Community-driven (dozens)	N/A (non-profit; relies on grants & donated compute)

Each of the above labs has played a significant role in advancing AI language models. OpenAI and Anthropic are independent startups with billions in funding focused on safe AGI/LLM development.

Ask anything

Search Deep research

ANTHROPIC

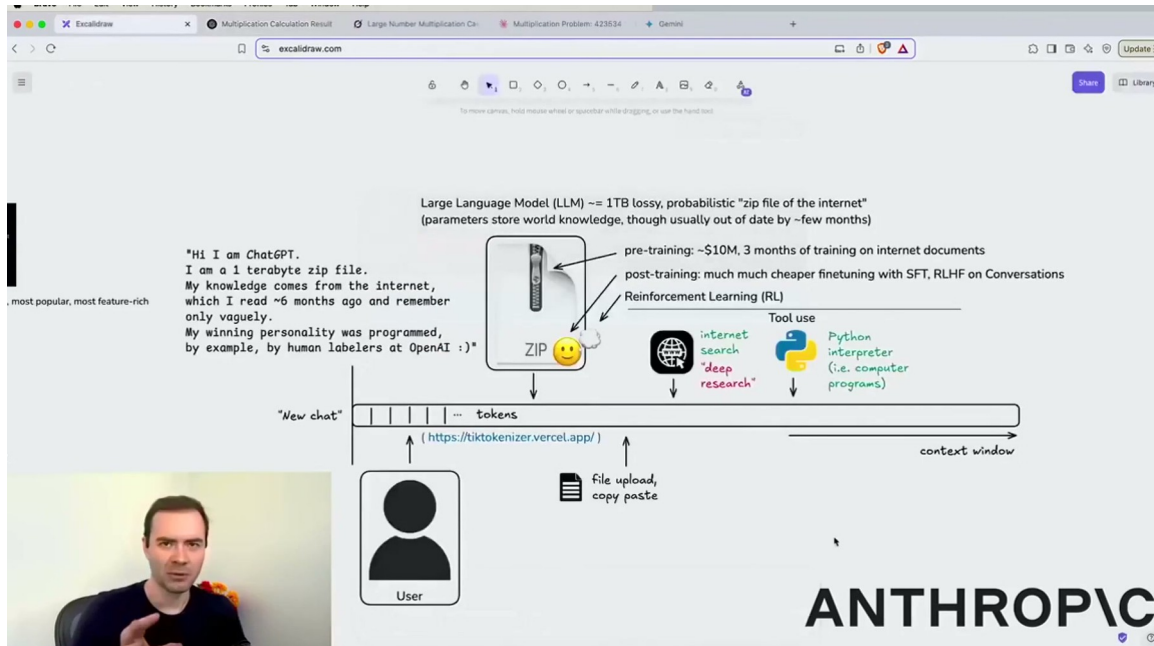
דוגמת כישלון: טבלת מעבדות ה-LLM בארה"ב. AI אסרה, ו-Hugging Face ו-EleutherAI לא היו צריכות להופיע

העלאת קבצים וקריאה משותפת

מעלים PDF, והטקסט שלו נכנס לחלון ההקשר. קרפתי מעריך שהתמונות כנראה נזרקות [48:14–45:02]. כך הוא קרא מאמר על Evo 2, מודל שפה שאומן על DNA. גם ספרים הוא כבר לא קורא לבד: הוא מעתיק פרק "עושר העמים" (מפרויקט גוטנברג) ל-Claude, מבקש סיכום, ואז קורא ושאל תוך כדי [52:51–48:14].

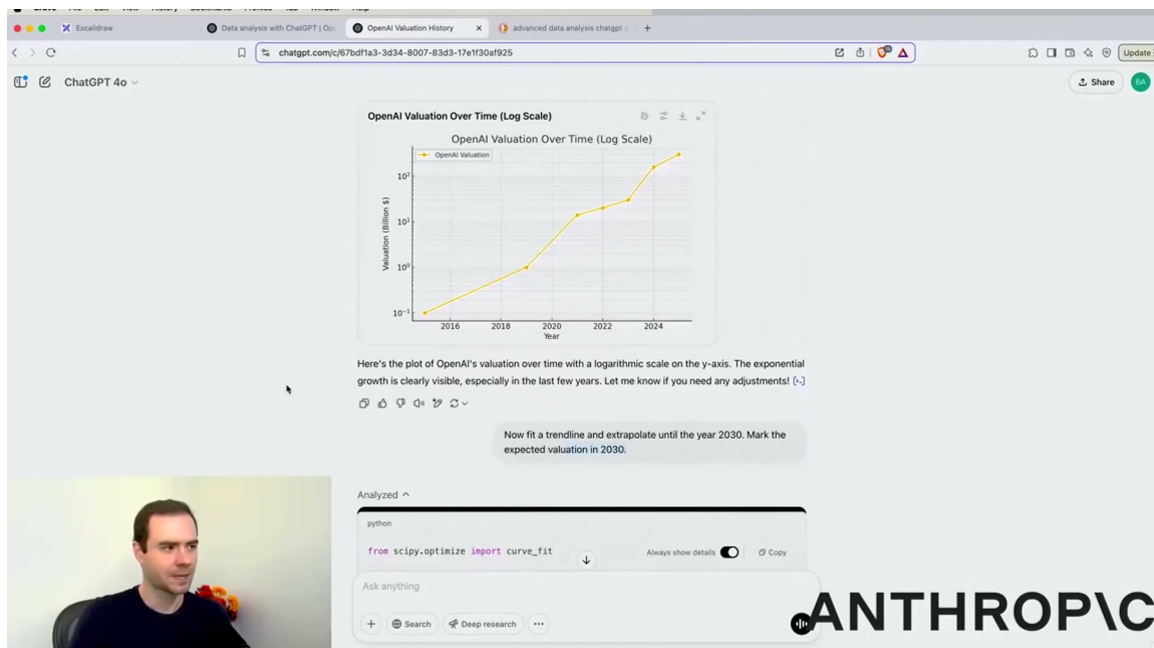
6. כלים: קוד

מפרש Python



התרשים מתרחב: לצד חיפוש ו-Deep Research נוספו מפרש Python והעלאת קבצים

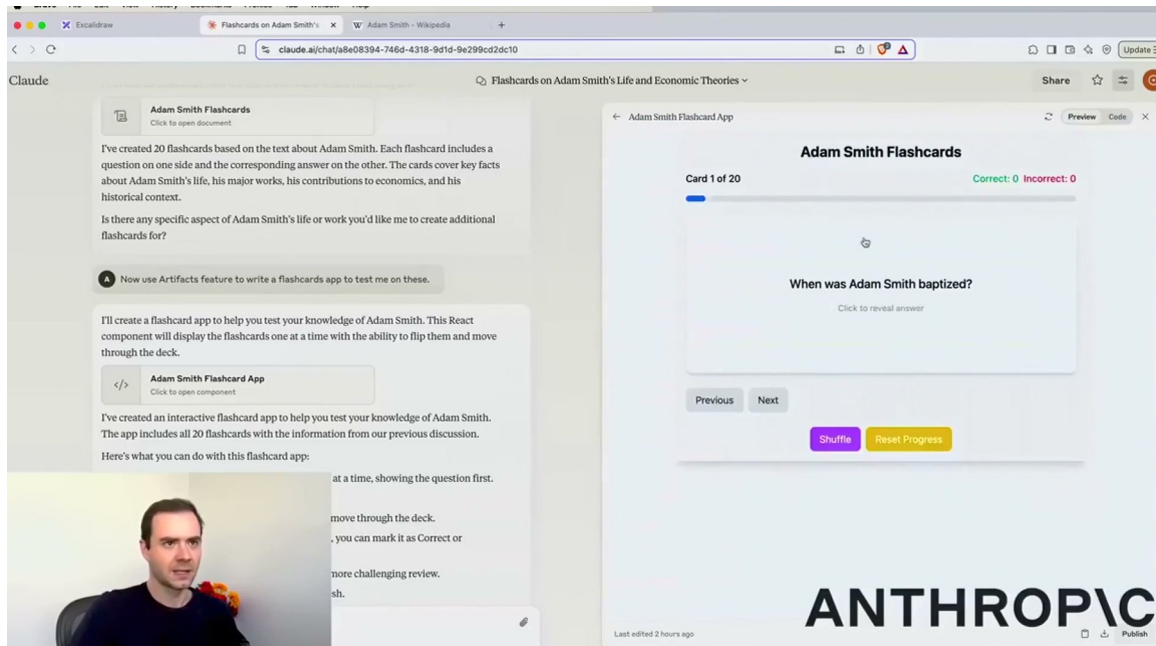
את 9x30 המודל מחשב בראש. כפל של מספרים גדולים ChatGPT מעביר לקוד Python, מריץ ומחזיר את התוצאה [55:59–53:00]. לא לכל האפליקציות היה הכלי הזה: Grok 3 חישב בראש וטעה, Claude כתב JavaScript וצדק, ו-Gemini 2.0 Pro צדק בכפל הראשון אבל טעה בכפל קשה יותר [59:07–55:59]. כשאין כלי, המודל ימציא תשובה.



Advanced Data Analysis: הערכות השווי של OpenAI על סקאלה לוגריתמית, והטעות שהסתתרה בקוד

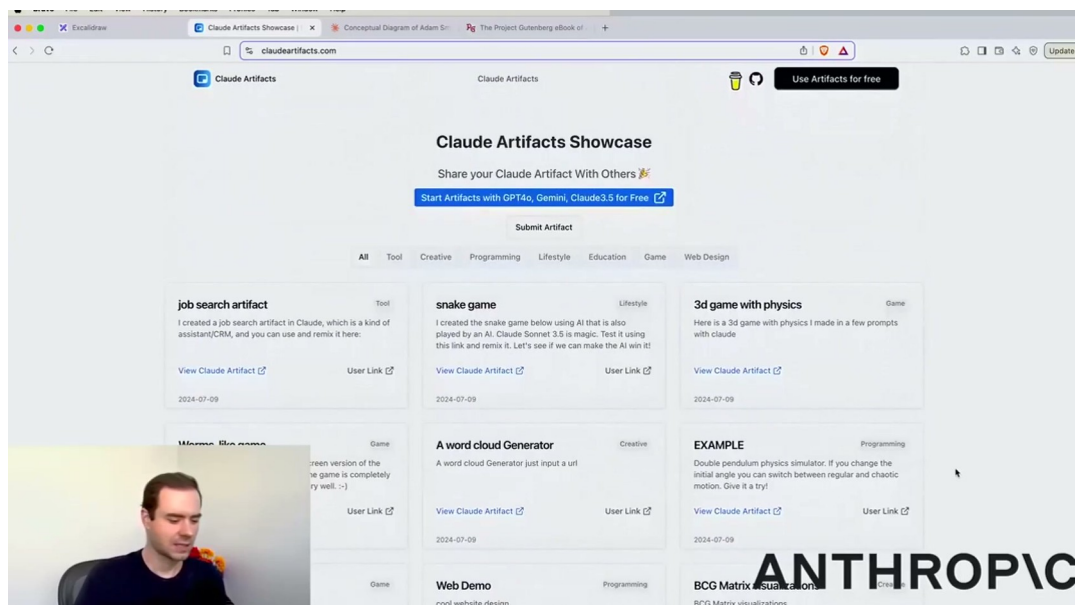
ב-Advanced Data Analysis, ChatGPT חיפש את הערכות השווי, שרטט ואז חישב אקסטרפולציה ל-2030. הוא הניח בשקט שהשווי ב-2015 היה 100 מיליון, וגם דיווח ש-2030 יגיע ל-1.7 טריליון, בזמן שהמשתנה בקוד הראה כ-20 טריליון. כשקרפתי עימת אותו, ChatGPT התנצל [1:02:17–59:07]. "it is kind of like a very very junior". [1:02:17] "data analyst: חזק, אבל צריך לקרוא את הקוד.

Claude Artifacts



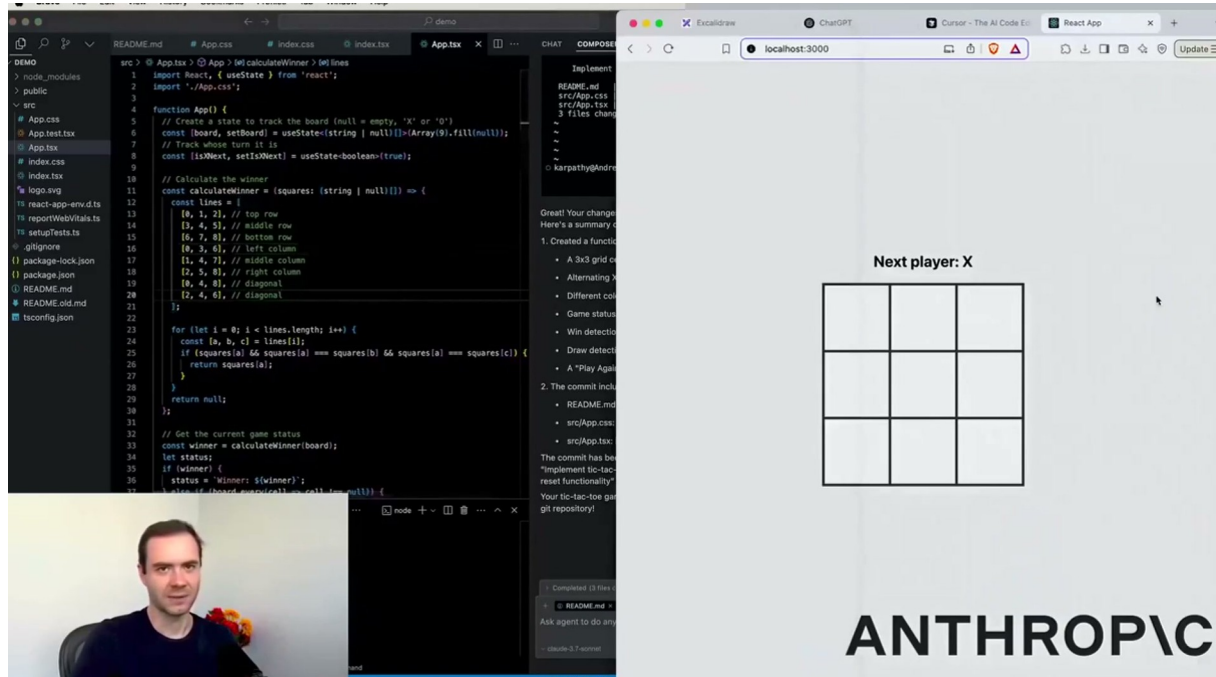
Claude כתב אפליקציית כרטיסיות על אדם סמית', שרצה בדפדפן

Claude כותב אפליקציית React קטנה שרצה בדפדפן. הדוגמה: 20 כרטיסיות על אדם סמית', ואפליקציה לתרגול אותן. קרפתי מודה שהוא לא משתמש בזה כל יום, אבל משתמש הרבה בתרשימי Mermaid לסיכום פרקים: "what is the layout of the argument?" [1:03:51–1:08:31]



Claude Artifacts Showcase: אפליקציות קטנות שאנשים שיתפו

vibe coding-ו Cursor



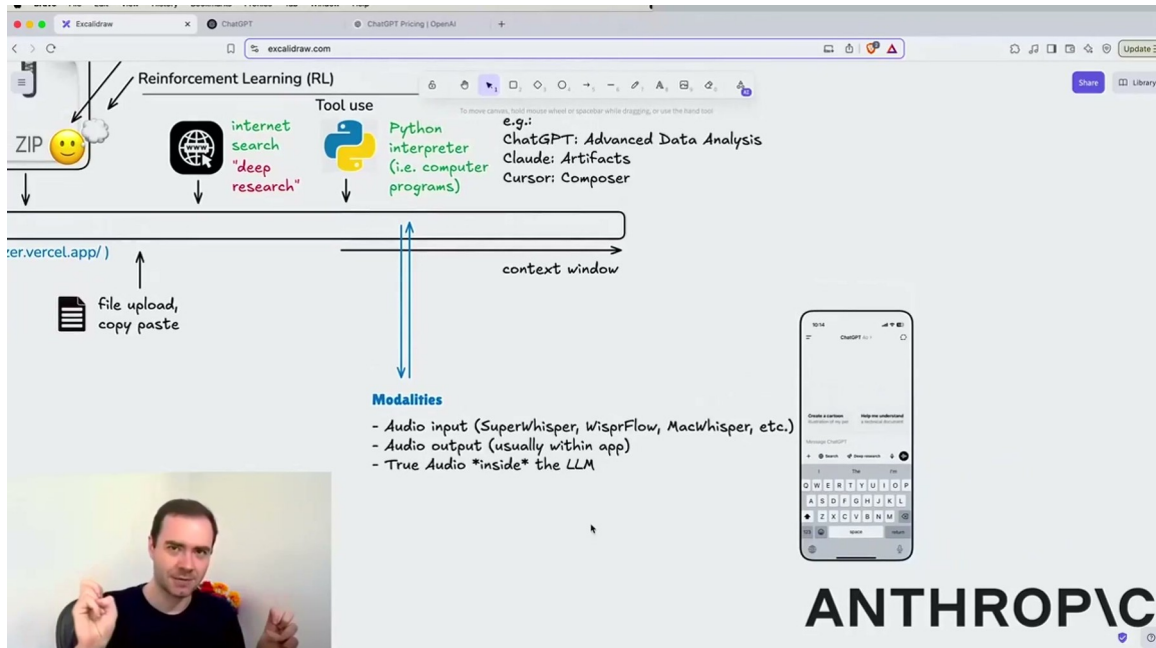
Cursor עם Claude 3.7 Sonnet: איקס-עיגול ב-React מחמש משפטים, ואחר כך קונפטי וציליל ניצחון

רוב זמן העבודה של קרפתי הוא כתיבת קוד, אבל לא בצ'אט, כי שם אין הקשר. הוא עובד ב-Cursor, שרואה את כל הקבצים בתיקייה ושולח את הבקשות ל-Claude דרך ה-API. ההתקדמות לפי הקיצורים: Cmd+K (שינוי שורה), אחר כך Cmd+L (הסבר), ועכשיו Cmd+I, Composer, שהוא סוכן שעורך כמה קבצים ומריץ פקודות [1:08:31–1:13:07].

— קרפתי, [1:13:07] the name for this is called vibe coding a name with that I think I probably minted

ההדגמה: הוא ביקש קונפטי כשמישהו מנצח. Composer התקין ספרייה, הוסיף עיצוב, ואז החליט בעצמו להוסיף ציליל ניצחון והוריד קובץ שמע מכתובת שקרפתי לא מכיר [1:13:07–1:16:15]. ואם משהו נשבר, "we can always fall back to standard programming".

7. מודאליות: קול, תמונה ווידאו

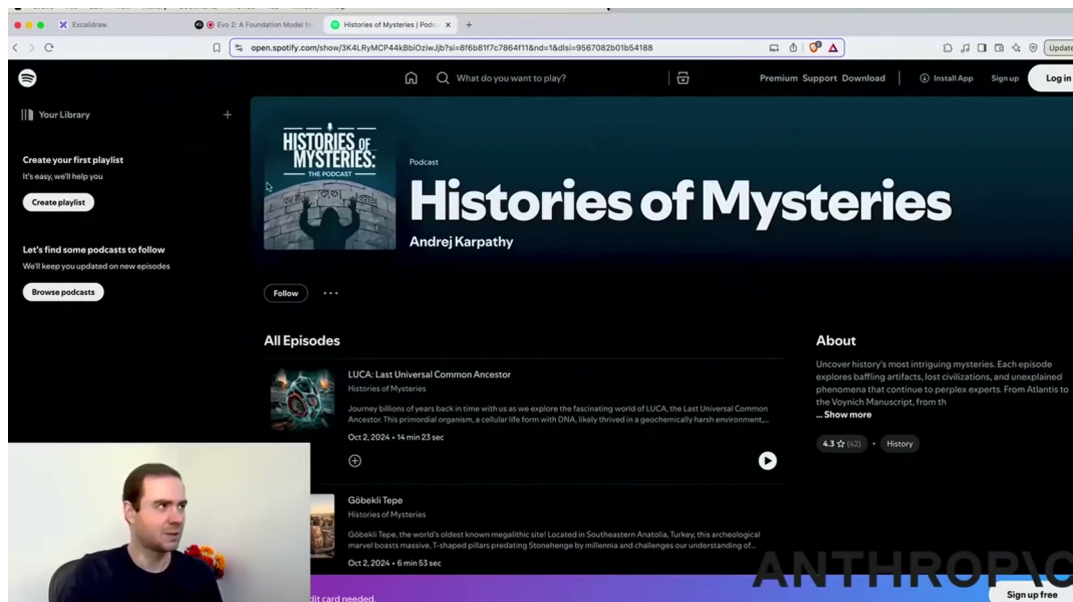


שכבת המודאליות בתרשים: קלט קול (SuperWhisper, WisprFlow, MacWhisper), פלט קול, ו"קול אמיתי" בתוך המודל

קול

- **"קול מזויף"**: המרת דיבור לטקסט ובחזרה. במחשב הוא משתמש ב-SuperWhisper: לוחץ F5, מדבר, ולוחץ F5 שוב [1:19:21]. שמות של מוצרים וספריות הוא מעדיף להקליר.
- **"קול אמיתי"**: האודיו מפורק לטוקנים, והמודל שומע ומדבר ישירות, בלי טקסט בכלל. ב-ChatGPT זה Advanced Voice Mode. הדגמה: יודה, פיראט, וספירה מהירה [1:22:29–1:27:11]. החיסרון לדבריו: "very, very cagey", מסרב להרבה.
- **Grok**: גם שם יש קול מתקדם, עם מצבים כמו Unhinged ו-"if you prefer a lot of Conspiracy, entertainment" [1:28:47].
- **NotebookLM**: מעלים מקורות ומקבלים פודקאסט מותאם ("Deep Dive"), כולל מצב שבו אפשר לקטוע ולשאול. קרפתי העלה לספוטיפיי סדרה שיצר כך, "Histories of Mysteries" [1:31:58–1:35:04].

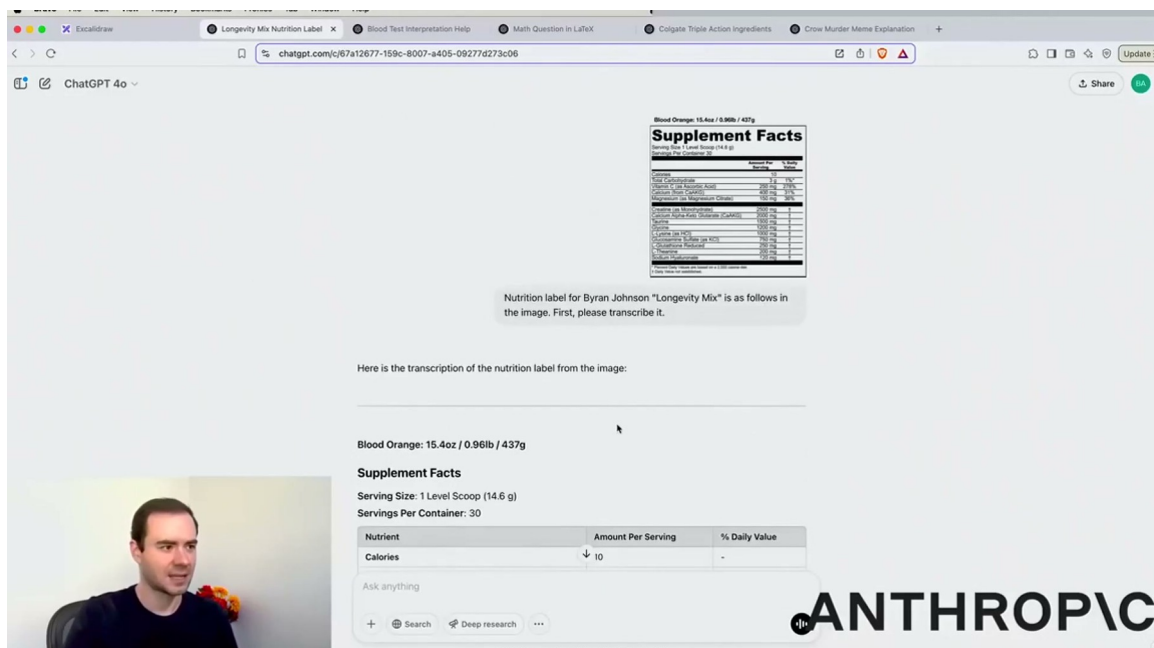
— קרפתי, [1:22:29] the big takeaway here is don't type stuff out use voice



Histories of Mysteries: פודקאסט שקרפתי יצר ב-NotebookLM והעלה לספוטיפיי

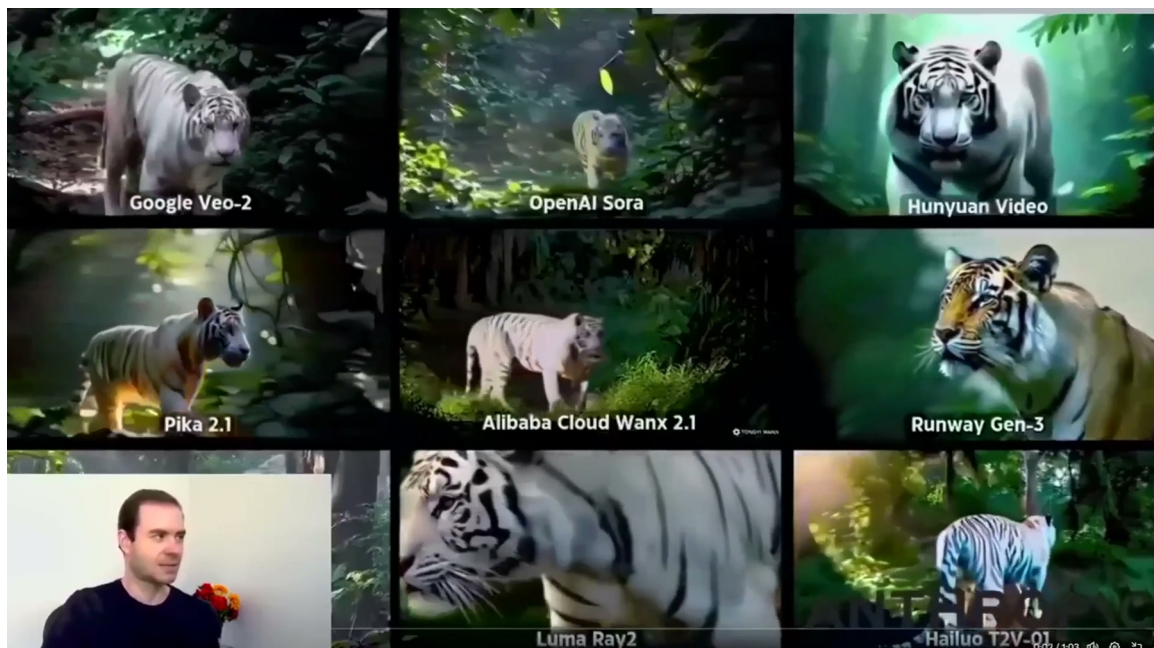
תמונות

גם תמונות מתפרקות לטוקנים, משבצת אחר משבצת, והטרנספורמר בכלל "לא יודע" אם הטוקן הוא טקסט, אודיו או תמונה [1:36:38–1:35:04]. השיטה שלו לקלט תמונה היא בשני שלבים: קודם מבקשים מהמודל לתמלל את התמונה לטקסט ובדקים שקרא נכון, ורק אז שואלים. כך עשה עם התווית של תוסף התזונה, עם בדיקות הדם ("a really good first draft" לשיחה עם הרופא), עם רכיבי משחת השיניים ועם מסע עורבים [1:42:48–1:36:38].



השיטה בשני שלבים: קודם תמלול תווית ה-Supplement Facts לטבלה, ורק אחר כך שאלות

ליצירת תמונות (DALL-E 3, Ideogram) הוא משתמש בעיקר לתמונות ממוזערות ולאייקונים ליוטיוב. ב-ChatGPT של אז, מודל נפרד יצר את התמונה מתוך כיתוב שהמודל הראשי כתב [1:44:21–1:42:48].

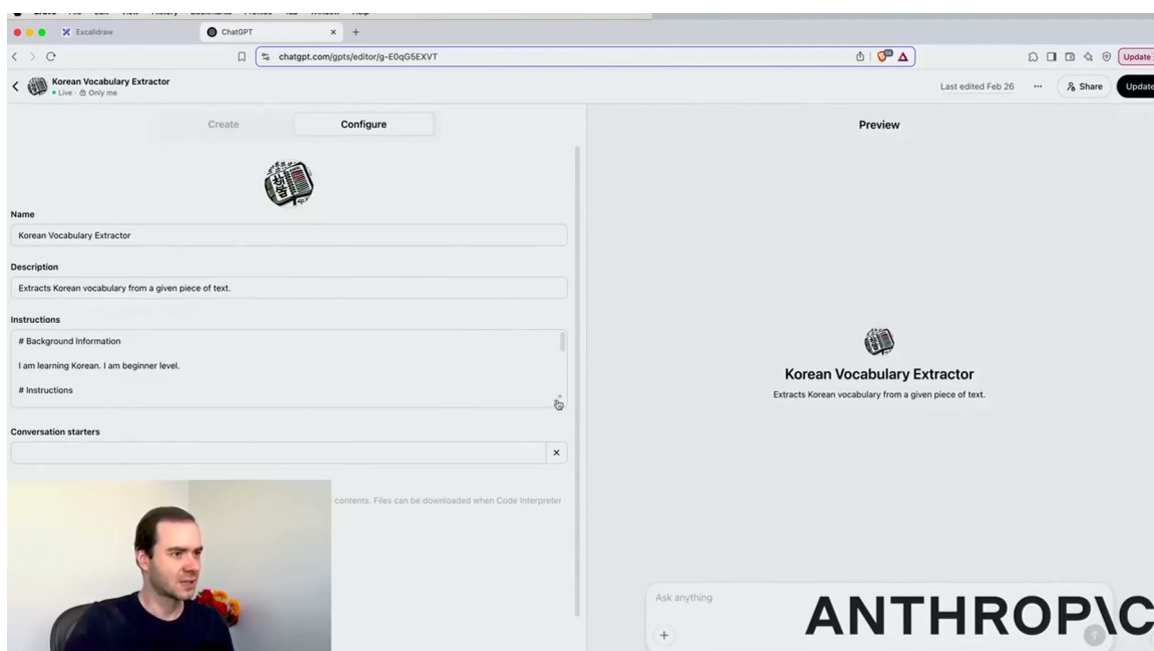


השוואת מודלי וידאו על "נמר בג'ונגל": Veo 2, Sora, Hunyuan, Pika ואחרים

וידאו: באפליקציה בנייד אפשר להפעיל מצלמה במצב הקול המתקדם ולהראות למודל דברים. בהדגמה הוא זיהה ספר של פיינמן, מד CO_2 ומפה של הארץ התיכונה [1:44:21–1:47:28]. קרפתי לא משתמש בזה בעצמו, אבל היה מראה את זה להורים שלו.

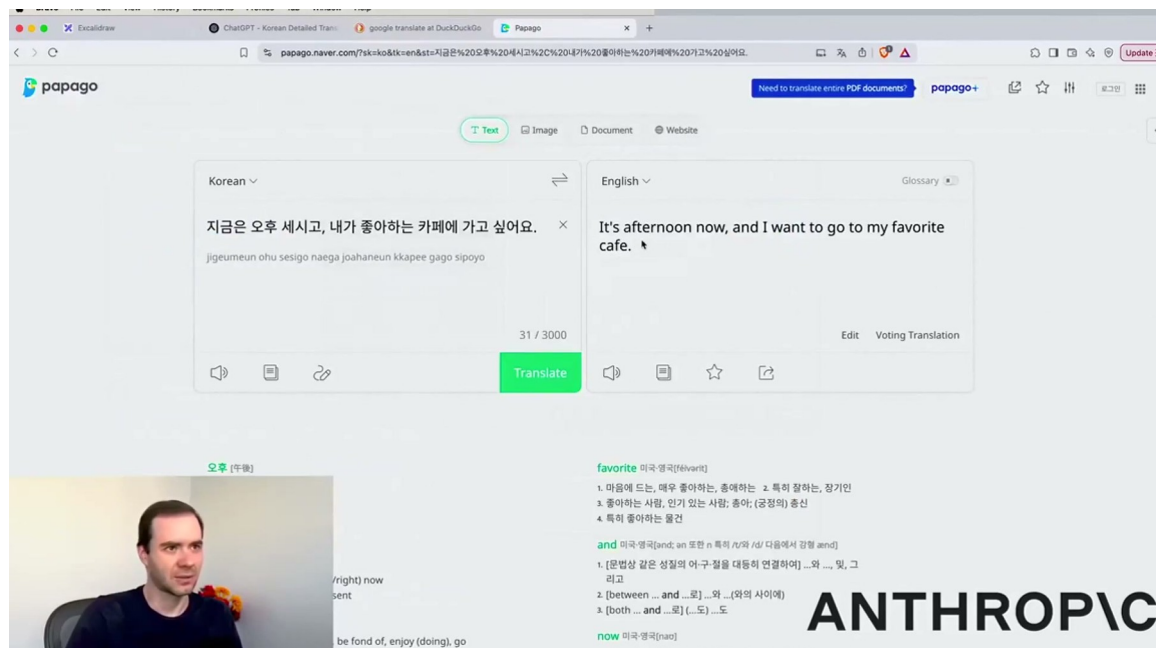
8. נוחות: זיכרון, הנחיות ו-GPTs

- **זיכרון (Memory):** ChatGPT שומר עובדות עליכם במאגר טקסט שמתווסף לתחילת כל שיחה. לפעמים מספיק לבקש "remember this" קרפתי מודה: "I'm definitely coming around" [1:49:06–1:52:12].
- **הנחיות מותאמות (Custom Instructions):** ההנחיה שלו: "don't be like an HR business partner just" [1:52:12–1:53:44], ותעדיף הסברים ותובנות. היא כוללת גם רמת רשמיות קבועה לקוריאנית.
- **Custom GPTs:** פרומפט שמור. העיקרון שהוא מדגיש הוא few-shot: לתת למודל דוגמאות קונקרטיות, ולא רק תיאור של המשימה, כי "this always increases the accuracy" [1:55:20].



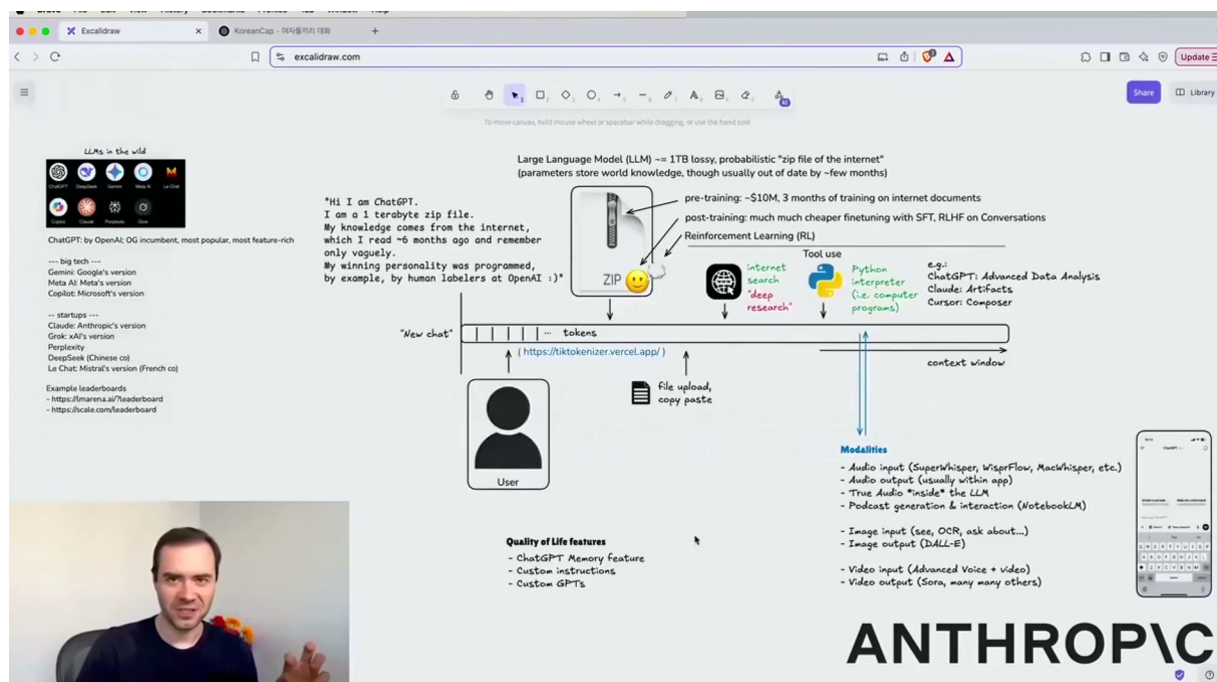
Korean Vocabulary Extractor: ה-GPT שמחלץ אוצר מילים בפורמט שמתאים ל-Anki

שלושת ה-GPTs שהוא בנה ללימוד קוריאנית: מחלץ אוצר מילים (מוכן לייבוא ל-Anki); מתרגם מפורט שמפרק את המשפט חלק אחר חלק ושעליו אפשר לשאול שאלות, ולדבריו הוא טוב בהרבה מ-Google Translate ומ-Papago; ו-"Korean Cap", שמקבל צילום מסך של כתוביות מתוכנית ריאליטי ומבצע OCR, תרגום ופירוק [1:53:44]–[2:01:32]. בדוגמאות ל-GPT הוא משתמש בתגיות בסגנון XML, כדי שהמודל ידע איפה כל דוגמה מתחילה ונגמרת.



Papago מתרגם, אבל לא מסביר. בהשוואה, המתרגם המפורט ב-ChatGPT מפרק את המשפט ומאפשר לשאול

9. התמונה המלאה, ומה התיישן מאז



התרשים בסוף ההרצאה: האפליקציות, המודל כ-*zip*, הכלים, המודאליזציות ותכונות הנוחות. דף עזר אחד

אם אתם רוצים...	מה להשתמש (לפי ההרצאה)	שימו לב
תשובה מהירה מידע כללי	מודל רגיל, בלי חשיבה	זיכרון מעורפל. מאמרים כשזה חשוב
בעיה קשה במתמטיקה או בקוד	מודל חושב (RL)	לוקח דקות. מנסים קודם מודל רגיל
מידע עדכני או משתנה	חיפוש (Perplexity, ChatGPT, Search)	בודקים את הציטוטים
סקירה עמוקה של נושא	Deep Research	טיוטה ראשונה בלבד

אם אתם רוצים...	מה להשתמש (לפי ההרצאה)	שימו לב
להבין מסמך או ספר	העלאת קובץ או הדבקה	תמונות כנראה נזרקות
חישוב, נתונים, גרף	מפרש Python / ADA	לקרוא את הקוד, לחפש הנחות מוסתרות
אפליקציה קטנה או תרשים	Claude Artifacts	אין צד שרת
פיתוח אמיתי	Cursor / Composer	לבדוק מה הסוכן התקין והוריד
לחסוך הקלדה	הכתבה (SuperWhisper וכו')	שמות מוצרים מקלידים
משימה שחוזרת על עצמה	Custom GPT עם דוגמאות	few-shot עדיף על תיאור

מה כבר לא נכון (אוקטובר 2026)

- **שמות מודלים:** GPT-4o, o1 Pro, Claude 3.5/3.7 Sonnet, Grok 3 ו-Gemini 2.0 הוחלפו מאז בדורות חדשים.
- **הפערים בין האפליקציות:** בהרצאה, ל-Claude לא היה חיפוש, ל-Gemini Pro לא היה חיפוש, ול-Grok לא היה מפרש קוד. רוב הפערים האלה נסגרו, וכדאי לבדוק מחדש לפני שמסיקים מהם משהו.
- **"הזיכרון ייחודי ל-ChatGPT":** זה היה נכון בפברואר 2025. מאז, גם אפליקציות אחרות הוסיפו זיכרון.
- **מחירים ומגבלות:** \$200 / \$20 ו-"80 הודעות כל שלוש שעות" הם נתוני פברואר 2025.
- **מה שלא התיישן:** המודל המנטלי (zip, חלון הקשר, כלים), ההרגלים (צ'אט חדש, מועצת מודלים, תמלול לפני שאלה), והעיקרון שכל תוצר הוא טיוטה.

10. מונחים

מונח	הסבר
טוקן	יחידת הטקסט (או האודיו או התמונה) שהמודל קורא וכותב
חלון הקשר (Context window)	רצף הטוקנים של השיחה הנוכחית. הזיכרון העובד של המודל
Pre-training	דחיסת האינטרנט לפרמטרים. יקר, נדיר, וקובע את ה-knowledge cutoff
Post-training	אימון על שיחות שנותן למודל את פרסונת העוזר
RL / מודל חושב	אימון בתרגול שבו המודל מגלה אסטרטגיות חשיבה
Deep Research	חיפוש וחשיבה ממושכים שבסופם דוח עם מקורות
Vibe coding	לתת לסוכן קוד לבצע, ולבדוק את התוצאה במקום את הקוד
Few-shot prompt	פרומפט שכולל כמה דוגמאות של קלט ופלט רצוי
Omni / multimodal	מודל שמעבד קול, תמונה או וידאו בתוכו, ולא דרך מודל נפרד