

Claude Opus 5.5 לעבודה: וובינר

הדגמה של Anthropic

מה השתנה מ-Opus 5 במחיר, במהירות ובתקשורת, ואיך זה נראה ב-Cowork עם Docs ו-Slides

Amelia Pedone, Nathan Wu, Nick Lin · וובינר של Anthropic (המקור המקורי לא זוהה; הועלה מחדש ב-X על ידי
2026-10-06 · (Codez

משך: כ-52 דקות · X

סיכום אוטומטי — AI Briefing (aibriefing.dev)

על המסמך הזה

זה וובינר של Anthropic בשם Opus 5.5 for work, באורך של כשעה, ומיועד במפורש לשימושים של לא-מפתחים. מציגים בו שלושה אנשים: Amelia (על השקף מופיע Amelia Pedone), שמנחה ומציגה את השינויים במודל; Nathan Wu מצוות Applied AI, שמריץ את הדמו של ניתוח הנתונים; ו-Nick Lin מצוות המוצר (בחלון הווידאו כתוב Nicholas Lin), שמציג את התרחישים הפיננסיים. רוב הזמן הוקדש לעבודה חיה ב-Claude Cowork.

הטענה המרכזית: Opus 5.5 זול בכ-40% מ-Opus 5, מהיר יותר בייצור פלט וכותב תשובות ברורות יותר, ולכן Anthropic רוצה שתשתמשו בו כ-daily driver בעבודה. כדי להמחיש את זה הם עוברים על פרויקט אחד: חברת ריטייל שהצמיחה שלה מואטת. מנתחים את הנתונים, בונים מסמך אינטראקטיבי ב-Claude Docs, עובדים עליו כמה אנשים במקביל, ובסוף ממירים אותו למצגת לדירקטוריון. בדרך יש גם השוואה ישירה של מצגות ש-Opus 5 ו-Opus 5.5 יצרו. חשוב לציין: הכיתוב של ההעלאה מחדש ב-X מדבר על motion design ועל "premier motion studio", אבל שני הנושאים האלה לא עולים בכלל בתמליל.

איך לקרוא את המסמך

- 1. **תקציר מנהלים:** הטענות המרכזיות עם חותמות זמן, ומה לא נאמר.
- 2. **מה השתנה מ-Opus 5:** מחיר, מנופי התמחור, מהירות ותקשורת.
- 3. **בטיחות ו-alignment:** prompt injection, safeguards, ומה קורה כשבקשה נחסמת.
- 4. **איפה Opus 5.5 מצטיין, וטיפים לפרומפטרים:** פיננסים, vision, ומה כדאי לשנות בפרומפטים.
- 5. **הדמו: ניתוח נתונים ו-Claude Docs:** בדיקת השערות, ממצא חדש ועבודה של כמה משתמשים במקביל.
- 6. **Opus 5 Slides מול Opus 5.5:** השוואה חזותית והמצגת לדירקטוריון.
- 7. **שאלות ותשובות:** גישה ל-Sonnet, brand guidelines, effort, artifacts.
- 8-9. **מסקנות ומילון מונחים:** מה לעשות עכשיו, שאלות פתוחות ומונחים.

הערת מקור המסמך הופק אוטומטית מהתמליל של ההקלטה ומהשקפים שצולמו ממנה. התמליל נוצר ב-ASR, ולכן שמות ומונחים עלולים להופיע משובשים. למשל, cash reads הוא כמעט בוודאות cache reads, ו-co-work הוא Cowork. במקומות שבהם יש ספק סימנו זאת במפורש. הציטוטים מועתקים מילה במילה מהתמליל ונבדקו מול חותמת הזמן שלהם. לא נוסף שום מידע ממקורות חיצוניים.

1. תקציר מנהלים

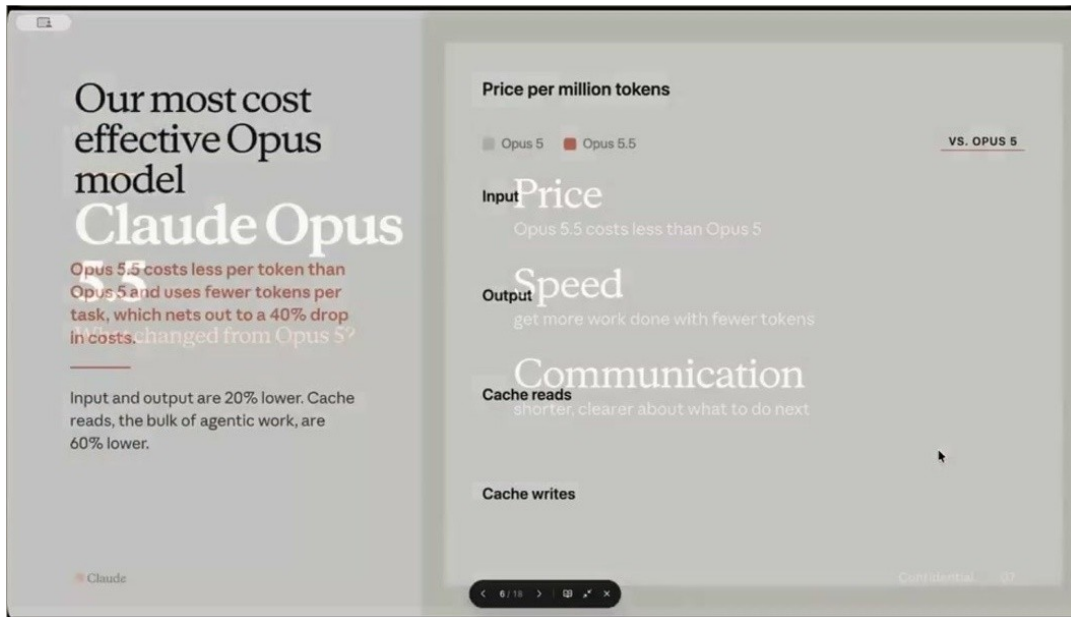
זה וובינר השקה ולא הרצאה טכנית. המספרים בו הם הערכות פנימיות של Anthropic, והדמו מתנהל לפי תסריט שהוכן מראש. ובכל זאת יש בו כמה דברים פרקטיים: כיצד לשנות את הפרומפטים מול המודל החדש, איזו רמת effort לבחור כברירת מחדל, ואילו יכולות מוצר חדשות נוספו (Docs, Slides ו-Design כ-artifacts, וגם עבודה של כמה משתמשים במקביל).

- **עלות נמוכה בכ-40%:** המחיר לכל token ירד, וגם מספר ה-tokens למשימה ירד. לפי Anthropic, input ו-output זולים ב-20% ו-cache reads ב-60% [1:33].
- **פלט מהיר ביותר מ-30%:** Opus 5.5 מייצר פלט מהר יותר מ-Opus 5 ומגיע לתשובה בפחות [3:04] tokens.
- **תקשורת טובה יותר:** התשובות קלות יותר לסריקה והמידע החשוב מופיע ראשון. הודגם על דוגמה של order exports שנכשלו [3:04].
- **הכי קשה לחטוף עד היום:** הוראות נסתרות הצליחו רק ב-1% מהמקרים, בתוך 15 ניסיונות. כשבקשה נחסמת, העבודה ממשיכה על מודל קודם במקום לקבל סירוב [4:38].
- **פיננסים ו-65% vision:** על benchmark פיננסי (השם לא ברור בתמליל). בדמו של merger model הוא היה מהיר בכ-30% וזול בכמעט 90% במשימה הזו [6:15], [7:47].
- **להפסיק לכתוב think hard:** המודל תמיד חושב לפני שהוא עונה, ורמת ה-effort קובעת כמה. ברירת המחדל המומלצת היא [12:27], [10:55] medium.
- **להגדיר מראש איך נראה done:** מטרה אחת בהודעה הראשונה במקום הוראות צעד אחר צעד. כך נבנה גם הפרומפט בדמו [12:27], [18:40].
- **Docs כמשטח עבודה משותף:** בדמו Claude הפריך את ארבע ההשערות של ההנהלה והציע הסבר חדש: חנויות חדשות שנפתחו קרוב מדי לחנויות קיימות [25:01].
- **Context הוא המנוף הכי חזק:** לחבר את Claude למערכות הארגון דרך [50:04] connectors.

מה לא נאמר

- **אין מחירים מוחלטים:** הוצגו רק אחוזי ירידה, בלי מחיר לכל מיליון tokens.
- **אין מתודולוגיה:** ה-40% מבוסס על "in our evaluations", וה-90% מתייחס למשימה בודדת. לא נמסר על אילו משימות נמדדו ובאיזה פיזור.
- **שם ה-benchmark הפיננסי ושם המודל להשוואה לא ברורים:** בתמליל מופיעים big finance bench ו-Fable 5.1. ייתכן שאלה שיבושי ASR.
- **ציון ה-alignment לא פורסם:** נאמר רק שהוא היה exceptional.
- **motion design לא הוזכר:** הכיתוב של ההעלאה מחדש ב-X לא תואם את תוכן הוובינר.
- **אין לוח זמנים:** לא נאמר מתי יגיעו Sonnet החדש ותוכן למפתחים.

2. מה השתנה מ-Opus 5: מחיר, מהירות, תקשורת



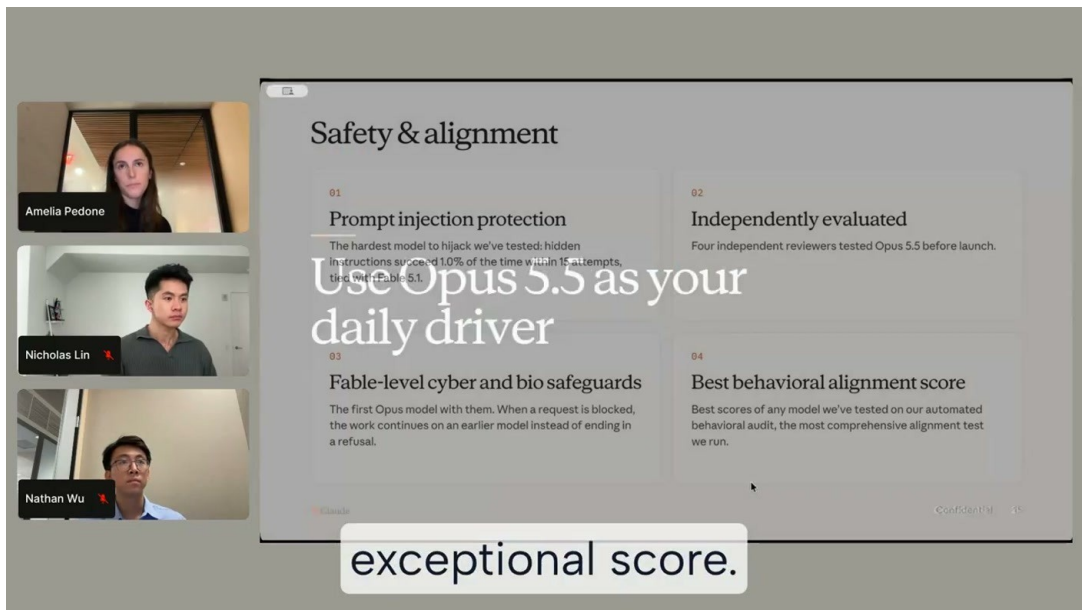
שקף התמחור: Our most cost effective Opus model, עם טבלת Price per million tokens שמשוה בין Opus 5 ל-Opus 5.5 בשורות cache writes- input, output, cache reads

Amelia מחלקת את השינויים לשלושה: מחיר, מהירות ותקשורת. במחיר יש שני אפקטים שמצטברים: כל token זול יותר, ובכל משימה נצרכים פחות tokens. Anthropic מעבירה את החיסכון למשתמשים דרך שלושה מנופים: input -20%, output -20%, ו-cache reads -60%. הירידה הכי גדולה היא ב-cache reads, ויש לזה הסבר: בעבודה agentic, "every time you type a new answer, quad has to read and reload the content that came before it" [3:04]. לכן Cowork היה יקר יותר מ-chat. הנקודה הזו חוזרת גם ב-Q&A [42:18].

במהירות, הפלט מהיר ביותר מ-30% של Opus 5. בתקשורת, הדוגמה הייתה דוח על orders שנכשלים בגלל timeout. בגרסה של Opus 5 קשה לעקוב אחרי ההסבר. Opus 5.5 פותח בעובדה (ה-exports נכשלו פעמיים החודש), ואחריה מפרט פעולות לטווח המיידי ולטווח הארוך. הניסוח שלה: "it's just writing a little bit more closely to how we think" [4:38].

— **Amelia Pedone [1:33]** Opus 5.5 costs less per token than Opus 5 and also uses fewer tokens per task. So in our evaluations, this nets out to about a 40% drop in costs

3. בטיחות ו-alignment



שקף Safety & alignment עם ארבעה רכיבים: Prompt injection protection, Independently evaluated, Fable-level cyber and bio safeguards ו-Best behavioral alignment score

לפי Anthropic, Opus 5.5 הוא "the hardest model to hijack that we've ever tested" [4:38]. הוראות נסתרות הצליחו רק ב-1% מהמקרים בתוך 15 ניסיונות, תוצאה שזהה לזו של מודל שנקרא בתמליל Table 5.1. השם מופיע גם בשקף (Fable-level safeguards), אבל לא ברור לנו לאיזה מודל הוא מתייחס, ולכן מסומן כלא ודאי. לפני ההשקה המודל עבר הערכה של ארבעה גופים בלתי תלויים, ויש בו safeguards ל-cyber ול-bio.

השינוי ההתנהגותי שהכי משפיע על המשתמש: "When a request is blocked because of the topic or the type of prompt that you asked, the work will continue on an earlier model instead of ending in a refusal" [4:38]. בנוסף נאמר שהמודל קיבל את ציון ה-behavioral alignment הטוב ביותר באודיט האוטומטי ש-Antropic מריצה על כל מודל. המספר עצמו לא נמסר.

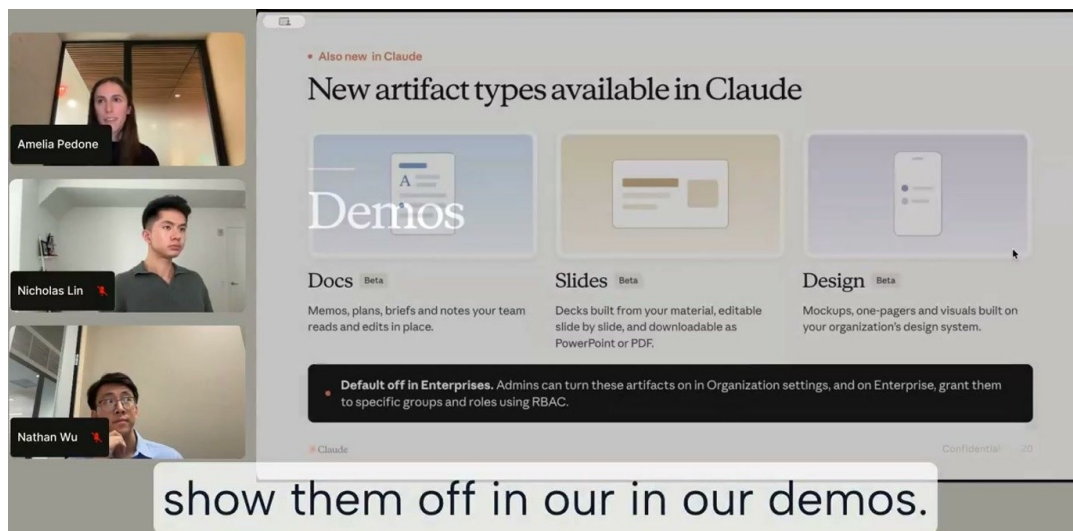
4. איפה Opus 5.5 מצטיין, וטיפים לפרומפטים

Nick מונה שלושה תחומים שבהם המודל חזק: איכות פלט ("follows instructions precisely and states its scope", כך שגיליונות, שקפים ומסמכים צריכים מעט מאוד עריכה), ניתוח נתונים ופיננסים, ו-vision (חילוץ מידע ממסמכים, גרפים, צילומי מסך ודיאגרמות) [6:15]. בתחום הפיננסי הוא מציין 65% על benchmark פיננסי, וארבע תכונות: קריאה של מקורות צפופים, שמירת ה-context לאורך עסקה שלמה, הצגה ואימות של העבודה, ופלט שמוכן לשיתוף. הדמו המוקלט הציג merger model ו-football field ב-Excel וב-Opus 5.5 מול PowerPoint, Opus 5.5 "we're seeing that opus 5 5 is about 30 faster and on this task specifically almost 90 cheaper than opus 5" [7:47].

אחר כך Amelia עוברת ל-AI fluency, כלומר לבדוק מחדש מה אפשרי עם כל מודל חדש. היא מציעה ארבעה דברים לנסות: לבקש את הקובץ הסופי ולא רק outline, לבקש בדיקה מדוקדקת של מסמכים ארוכים (באחד המבחנים המודל זיהה גרף בדוח שלא התאים למספרים בדק), מחקר פתוח ארוך, ויצירתיות חזותית [9:19]. הטבלה מסכמת את השינויים שהיא ממליצה לעשות בפרומפטים.

טיפ	מה נאמר	זמן
להוריד think hard / think carefully	המודל תמיד חושב לפני שהוא עונה, ורמת ה-effort קובעת כמה. אם מורידים את ההוראה, התשובה מתחילה מוקדם יותר.	10:55
ברירת מחדל: effort medium	"the best calibrated for cost and performance". מעבר ל-max או extra max לא תמיד משפר את התוצאה.	43:50, 12:27

זמן	מה נאמר	טיפ
12:27	מטרה ופלט רצוי בהודעה אחת. המודל טוב יותר מ-Opus 5 בעבודה ארוכה עם כמה שלבים.	להגדיר done בהודעה הראשונה
12:27	בלי הכוונה המודל חוזר לכמה סגנונות ברירת מחדל. "avoid generic" רק מחליף ברירת מחדל אחת באחרת. רשימה של patterns ו-anti-patterns עובדת טוב יותר.	לפרט אילו סגנונות לא רוצים
14:04	לא צריך להקליד מחדש נתונים מגרף או מצילום מסך, המודל קורא אותם בדיוק.	לשתף את הגרף עצמו



שקף New artifact types available in Claude: Docs, Slides ו-Design, כולם בתיוג Beta. בתחתית: Default off in Enterprises, ואדמינים מפעילים אותם ב-Organization settings לקבוצות ספציפיות

5. הדמו: ניתוח נתונים ו-Claude Docs



שקף One Project, Three Demos: חברת ריטייל שהצמיחה שלה יורדת, Amelia בתפקיד המנהלת של Nathan, ושלושה חלקים: Data analysis, Multiplayer work in Docs ו-Opus 5.5 vs. Opus 5: Creating slides Head to head

Nathan פותח session ב-Cowork עם Opus 5.5 ב-effort medium. ל-session הועלו מראש קבצי CSV של חנויות ומכירות, קובצי PDF מה-offsite של ההנהלה ועוד מקורות, חלקם מובנים וחלקם לא. הוא מסביר שבעבודה אמיתית ה-context מגיע דרך connectors כמו SharePoint, Google Drive או BigQuery. הפרומפט נכתב בכוונה בלי "think really hard" ובלי לפרט שלבים: לבדוק את ארבע ההשערות של ההנהלה מול הנתונים ולומר אילו

מחזיקות [18:40]. אחר כך הוא שואל שאלת המשך על הנחות ומבצעים, ומבקש להפוך את התוצאה לדוח ב-house style.

כשהוא מנסה ליצור את ה-Doc בשידור חי, זה נתקע. מתברר שצריך להפעיל את Claude Docs בלשונית connectors-, ובסוף הוא מדגל לפלט שהוכן מראש [23:31]. ב-Doc יש תוכן עניינים, סטטיסטיקות, אנימציה לכל השערה, regression של המרחק בין חנות חדשה לחנות קיימת, ו-slider אינטראקטיבי שמראה איך המכירות משתנות לפי המרחק. ההנחות והמבצעים נפסלו, והשערה אחת לא נשללה לגמרי אבל גם לא נמצא לה ביסוס בנתונים.

The screenshot shows a video call window on the left with three participants: Amelia Pedone, Nicholas Lin, and Nathan Wu. On the right is the Claude AI interface. The main window displays an executive summary report titled "Why did growth fall from 13% to 8%?". The report includes an "Executive summary" section with bullet points and a "Key Metrics" section with a table.

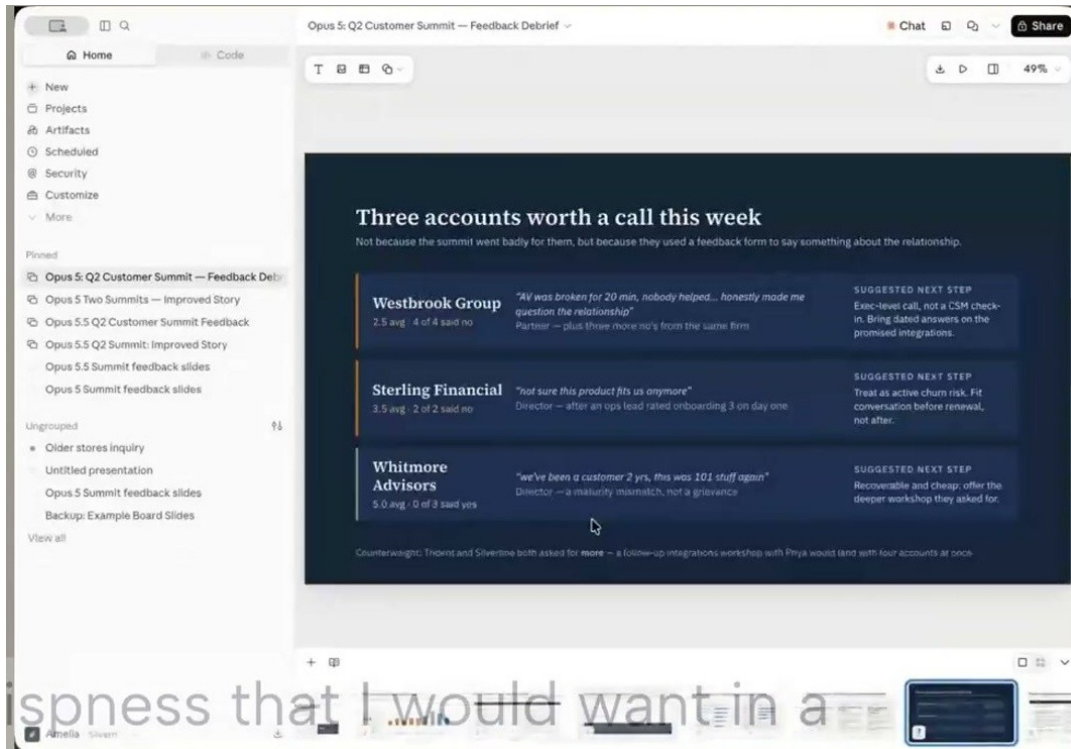
KEY METRICS	
Sales growth, January to August	+7.7%
Older stores within 5 miles of a new store	-8.4%

Below the table, the text reads: "In this report and key metrics. Seven sections follow. Hover a number to see how it is worked out." The report also includes a section titled "Testing the explanations" with the subtext "Four explanations checked against this data".

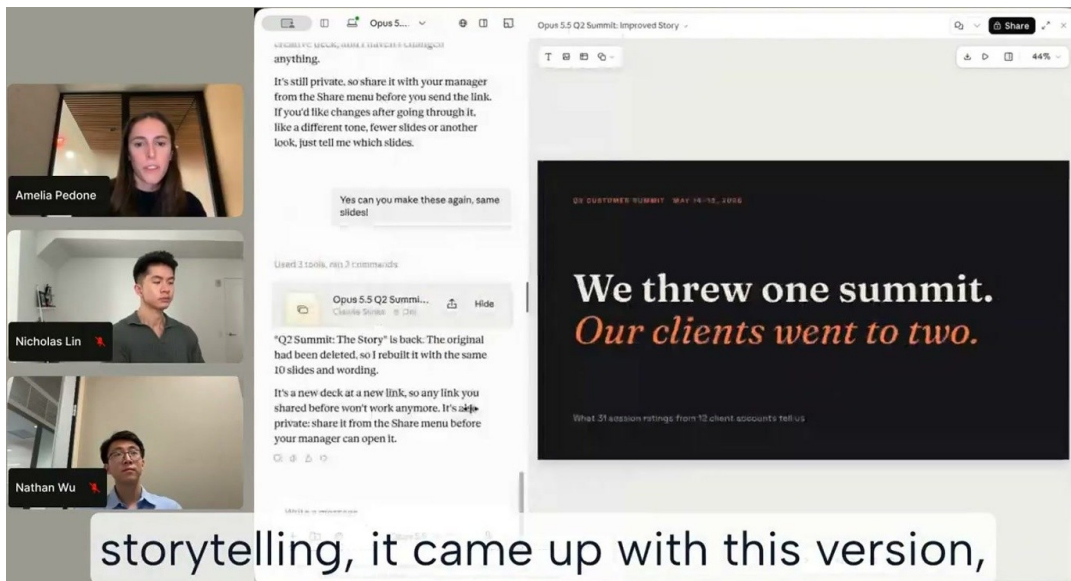
ה-Doc שנוצר ב-Cowork: Executive summary עם תבליטים, משפטים מסומנים בהדגשה, והערות של משתתפים בעמודה הימנית

Nathan Wu [25:01] Claude is proposing that the reason sales have declined is because the new — stores that we open are actually too close to our existing stores

Opus 5.5 מול Slides: Opus 5.6



פלט של Opus 5 על סקר משוב מכנס לקוחות: שקף כזה בשם Three accounts worth a call this week עם טבלה עמוסה בטקסט



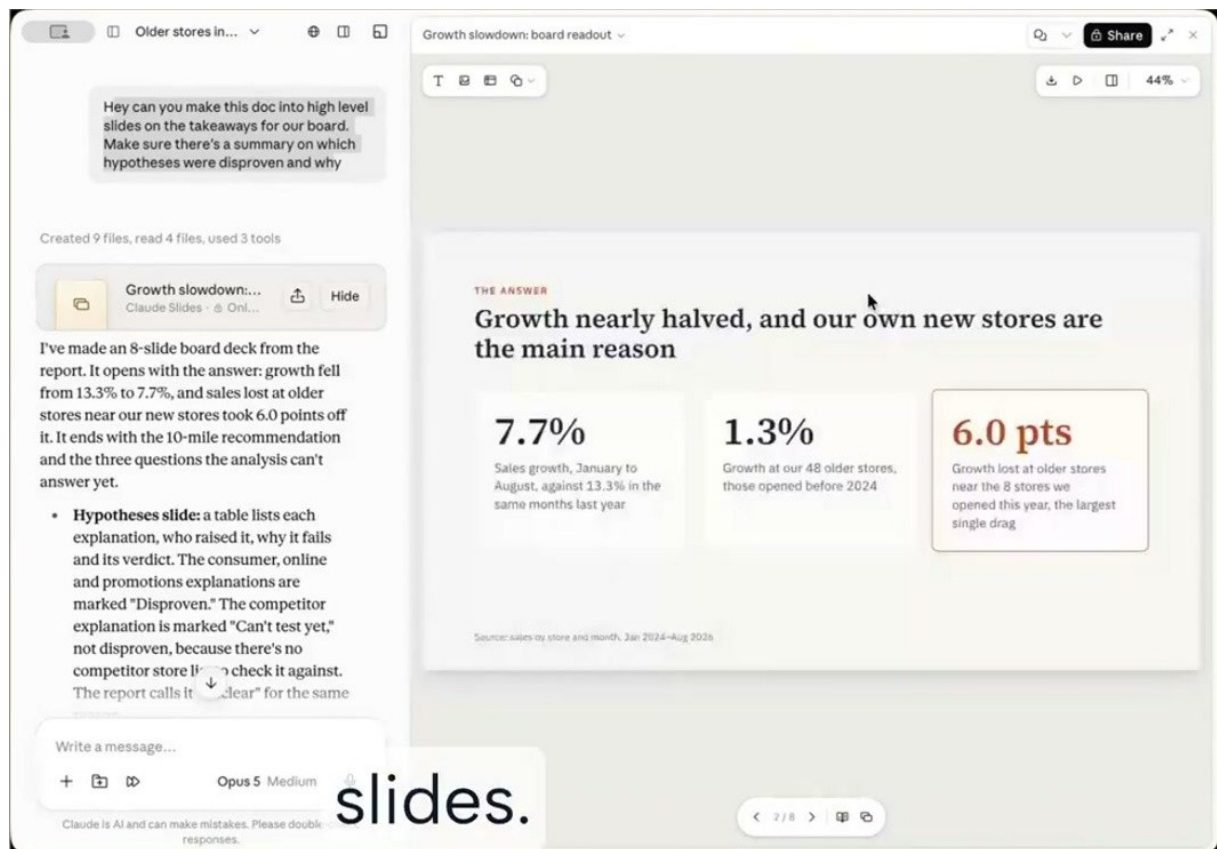
הגרסה היצירתית של Opus 5.5 לאותו משוב: שקף פתיחה עם התזה We threw one summit. Our clients went to two.

Amelia חוזרת לפני כן ל-Doc כדי להציג עבודה משותפת. Nathan משתף איתה את ה-Doc, היא נכנסת כ-viewer, מבקשת הרשאת edit ומקבלת אותה. מכאן כל אחד מהם יכול לתייג את Claude בתגובה, ו-Claude מבצע. בדוגמה Nathan ביקש להחליף את הריבועים בגרף בסמיילים. אחר כך היא מבקשת להפוך את ה-Doc לשקפים לדירקטוריון, עם סיכום של ההשערות שהופרכו והסיבה לכך.

בזמן שהשקפים נוצרים היא מציגה השוואה על סקר משוב מכנס לקוחות שנמשך מ-Google Sheet. לדבריה, הגרסה של Opus 5 "a little bit corporate": עמוסה בטקסט, עם רווחים לא אחידים וגדלי פונט שונים. גם אחרי שביקשה גרסה יצירתית היא נשארה ארוכה מדי. הגרסה של Opus 5.5 נקייה, והכותרות שלה מעבירות את המסר בעצמן: משתמשים חדשים נתנו ציון גבוה ולקוחות ותיקים לא. בגרסה היצירתית יש תזה בפתיחה וגרף התפלגות שמספר את הסיפור. במצגת לדירקטוריון היא נותנת תגובה על טבלה ("can you make this into columns instead of a")

export:-export: boring table") ומשתפת את Nathan כ- [39:08], [37:35] editor. בסיום מוצגות אפשרויות ה-export: PowerPoint (התוכן נשאר ניתן לעריכה), PDF, Word, Google Docs.

Amelia Pedone [37:35] it seems like Opus is looking, Opus 5.5 is looking at, like, the full story that — I need to tell, whereas Opus 5 attacks each slide, like, individually



המצגת לדירקטוריון שנוצרה מה-Doc: שקף בשם "Growth nearly halved, and our own new stores are the main reason", עם כרטיסי מדדים, ולצידו צ'אט ה-sidebar

7. שאלות ותשובות

- **Chat או Cowork:** אפשר ליצור פלטים גם מ-chat. הירידה במחיר cache reads נועדה לטפל בכך ש-Cowork היה יקר יותר [42:18].
- **לא רואים artifacts:** בתוכנית Enterprise האדמין צריך להפעיל אותם, ואפשר להפעיל רק לקבוצות משתמשים לניסיון [42:18].
- **Opus 5 על high מול Opus 5.5 על medium:** ברירת המחדל נקבעת לפי evals על כל משטח מוצר. "running" [43:50] "it on like max or extra max won't always like get you a much, much better output".
- **Brand guidelines:** אפשר להדביק הנחיות כטקסט, ואולי יידרשו כמה סבבים. דרך נוספת היא להפנות דרך connector לתיקייה ב-Drive עם המצגות הטובות ביותר של הארגון. בדמו הוחלו ה-brand guidelines של Anthropic [45:24].
- **Opus מול Sonnet:** ההמלצה היא לנסות קודם את Opus 5.5 כ-Sonnet. daily driver חדש בדרך, בלי תאריך [47:01].
- **Design artifact מול אפליקציית Design העצמאית:** הכול עדיין ב-beta. המטרה היא תאימות מלאה (one-to-one), וכל היכולות כבר הועברו [47:01].
- **איחוד chat ו-Cowork:** בהדרגה במהלך השבועות הקרובים. ב-Enterprise זה לא יופעל אוטומטית ותהיה תקופת מעבר [48:32].

8. מסקנות ופעולות

- לנקות את ספריית הפרומפטים: להוריד think hard ו-make no mistakes אם התשובות מתחילות מהר יותר.
- להעביר workflows ל-effort medium: ולבדוק מול high לפני שמניחים ש-max משתלם.
- לכתוב פרומפטים שמוגדרים לפי מטרה: הגדרה של done בהודעה הראשונה, בלי לפרק לשלבים.
- בעיצוב, לתת רשימת anti-patterns מפורשת: ולא רק avoid generic.
- לבקש מהאדמין להפעיל את Docs, Slides ו-Design: לקבוצת pilot. הם כבויים כברירת מחדל ב-Enterprise.
- להשקיע ב-connectors: לפי הדוברים זה "[50:04] the single highest leverage thing that you can do".
- למדוד עלות בעצמכם: ה-40% וה-90% הם מספרים של Anthropic. כדאי להשוות token usage על משימות שלכם.

שאלה פתוחה	למה זה חשוב
מה המחיר המוחלט לכל מיליון tokens?	בלי זה אי אפשר לחשב תקציב, רק להשוות לגרסה הקודמת.
על אילו משימות נמדדה הירידה של 40%?	הדמו הפיננסי הראה 90% במשימה אחת. הפער מראה שהחיסכון משתנה הרבה בין משימות.
מהו Fable 5.1 ומהו ה-benchmark הפיננסי?	השמות מגיעים מ-ASR. צריך לבדוק אותם לפני שמצטטים.
כמה פעמים העבודה עוברת למודל קודם בגלל safeguards?	המעבר השקט משפיע על איכות הפלט בלי שהמשתמש יודע.
מתי יגיע Sonnet החדש ובאיזה מחיר?	החלטה בין Sonnet ל-Opus לשימוש יומיומי בארגון.

9. מילון מונחים

מונח	משמעות
Cache reads	קריאה של context ששמור ב-cache. בכל תור בשיחה agentic קוראים מחדש את כל מה שקדם לו
Effort level	הגדרה שקובעת כמה המודל חושב לפני שהוא עונה (medium, high, max)
Prompt injection	הוראות נסתרות בתוכן שמנסות להשתלט על המודל
Daily driver	המודל שמשתמשים בו כברירת מחדל בעבודה היומיומית
AI fluency	היכולת לבחון מחדש מה אפשרי עם כל מודל חדש ולהתאים את דרך העבודה
Cowork	משטח העבודה ה-agentic של Claude שבו רץ הדמו
Artifacts	פליטים עשירים בתוך Claude: Docs, Slides ו-Design
Connectors	אינטגרציות למערכות הארגון (Drive, SharePoint, BigQuery)
Multiplayer	כמה משתמשים ו-Claude עובדים יחד על אותו מסמך
Context	המידע שניתן למודל כדי לבצע את המשימה
Merger model / Football field	מודל מיזוג פיננסי וגרף טווחי שווי שמקובלים ב-investment banking
House style	סגנון הכתיבה והעיצוב של הארגון