

איך להפסיק להיתקע על מגבלות השימוש של Claude

22 טריקים וטיפים שיעזרו להשתמש ב-Claude טוב יותר, ולא לבזבז יותר מדי כסף

22 הרגלים · מהכי פחות מוכרים עד הכי ברורים

המדריך הזה הוא רשימה של 22 הרגלים, מדורגים מהכי פחות מוכרים עד הכי ברורים מאליהם.

02 שלחו אותו לכל מי בצוות שלכם שממשיכ/ת להתלונן על מגבלות Claude.

01 שמרו את המדריך הזה. **בחרו 3 הרגלים לשבוע הזה.** עד סוף השבוע כבר תרגישו בהבדל.

ה ב ס י ס

Claude סופר טוקנים

טוקן הוא בערך מילה אחת.

שולחים הודעה אחת, ו-Claude קורא מחדש את כל השיחה מההתחלה. כל הודעה קודמת. כל תשובה קודמת. את הכול.

לכן הודעה מספר 1 עולה מעט מאוד. אבל הודעה מספר 30? Claude קורא מחדש 29 חילופים קודמים עוד לפני שהוא בכלל מתחיל לחשוב על השאלה החדשה שלך.

בגלל זה הקרדיטים שלנו נעלמים. ככל שהשיחה מתארכת, **כל הודעה נהיית יקרה יותר.**

כל הרגל במדריך הזה חוזר לאותו רעיון: **איך להפסיק לבזבז טוקנים, כדי שתוכלו להשקיע אותם במה שבאמת חשוב.**

ההרגלים שאת כנראה לא מכירה

להמיר קבצים לפני העלאה

01

עמוד PDF אחד עולה בין 1,500 ל-3,000 טוקנים. צילומי מסך אפילו גרועים יותר - תמונה מלאה בגודל 1000×1000 היא בערך 1,300 טוקנים. קבצי DOCX ו-PPTX כוללים נפח מיותר של מטא-דאטה שאת בכלל לא רואה.

לפני שמעלים, הוציאו את הטקסט. העתיקו והדביקו את החלקים הרלוונטיים לקובץ טקסט פשוט או Markdown. חיתוך הדוק יכול להוריד תמונה מ-1,300 טוקנים לפחות מ-100.

אם מעלים את אותו PDF של 15 עמודים ל-4 צ'אטים שונים, נשרפו יותר מ-180,000 טוקנים על מסמך שיכולתם להמיר ל-2,000 טוקנים של טקסט נקי.

תהליך העבודה המועדף:

WORKFLOW

פתחו Google Doc ← הקליד **doc.new** בשורת הכתובת
הדביקו את הטקסט שצריך להעלות ל-Claude
הורידו: File → Download → Markdown (.md)

Anthropic אישרו שיצירת קבצים, גיליונות, מסמכים ומצגות, צורכת יותר מהמכסה שלנו מאשר הודעות צ'אט רגילות.

אל תפתחו Cowork ותכתבו: "צור לי מודל פיננסי". במקום זה, פתחו Chat, תכננו את המבנה, סגרו את הסעיפים, דייקו את ההנחות. רק כשברור לכם בדיוק מה אתם רוצים, עברו ל-Cowork.

הכלל: את החשיבה עושים במוצר הזול. את הבנייה עושים במוצר היקר. העתיקו את התשובה מ-Claude Chat, אחרי שאתם מרוצים ממנה, והדביקו אותה ב-Cowork עם Opus 4.6 - Extended Thinking.

פרומפט של 500 מילים עולה 500 טוקנים בכל פעם ש-Claude קורא מחדש את השיחה. אבל אם תכתבו פרומפט של 15 מילים ותתנו ל-AskUserQuestion לעשות את העבודה, שאלות ההבהרה נוצרות פעם אחת, והתשובות הן לחיצות קצרות.

פרומפט שאפשר להשתמש בו **קבוע** - פחות מ-30 מילים:

"I want to [task] to [success criteria]. Read my folder. Ask me questions using AskUserQuestion before you start."

לחיצה על אפשרויות כמעט לא עולה כלום. להקליד פסקאות של הוראות עולה הרבה. תנו ל-Claude למשוך מכם את ההקשר, במקום שתדחפו אליו חומות של טקסט.

להשתמש ב-Wispr Flow כדי לתת תשובות עשירות יותר

04

Wispr Flow הוא כלי להמרת דיבור לטקסט.

זה נשמע הפוך מההיגיון: לדבר במקום להקליד, וככה להשתמש בפחות טוקנים? זו הסיבה שזה עובד: כשאנחנו מקלידים, אנחנו כותבים פרומפטים "עצלנים". "תשפר את זה." "תשנה את הטון." עמום. Claude מנחש לא נכון. ואז אנחנו ממשיכים לשלוח עוד ועוד הודעות.

כשאנחנו מדברים, אנחנו נוטים לתת יותר הקשר בשוט אחד:

"הטון קשיח מדי. אני רוצה שזה יישמע כמו הודעה לחבר שמנהל חברה של 200 איש. תשמור את הדאטה אבל תכתוב יותר קליל. תכתוב מחדש רק את סעיף 2."

פחות הודעות = פחות טעינות מחדש של ההקשר = חיסכון בטוקנים.



להפסיק לבקש מ-Claude לכתוב הכול מחדש

05

אם סעיף 3 בדוח לא טוב, אל תגידו "תכתוב מחדש את כל הדוח".

"תכתוב מחדש רק את סעיף 3. תשמור על כל השאר כדי לחסוך טוקנים."

כל כתיבה מחדש של כל הפלט אומרת ש-Claude מייצר שוב את כל התשובה. אם הדוח הוא 2,000 טוקנים, שרפתם שוב 2,000 טוקנים של פלט.

ועל הדרך, הוסיפו: **"בלי הערות. בלי הסברים. רק את הפלט."** כשברור לכם בדיוק מה אתם רוצים. כל משפט כמו "בשמחה, הנה מה שעשיתי..." הוא טוקנים שאנחנו משלמים עליהם.

שלושה פרומפטים נפרדים = שלוש טעינות מלאות של ההקשר. פרומפט אחד עם שלוש משימות = טעינה אחת.

במקום לשלוח "תסכם את המאמר הזה", ואז "תכתוב את הנקודות המרכזיות", ואז "תציע כותרת", כתבו:

"תסכם את המאמר הזה, תכתוב את הנקודות המרכזיות, ותציע כותרת."

בנוסף: בדרך כלל גם התשובה יוצאת טובה יותר. Claude צריך לראות את התמונה המלאה בבת אחת, בדיוק כמו בן אדם רגיל. ✨

Anthropic אישרו שפרומפטים דומים שמשתמשים בהם לעיתים קרובות נשמרים חלקית בקאש. המסקנה המעשית ברורה: כדאי להחזיק ספריית פרומפטים יציבה, ולהחליף רק את החלק המשתנה.

דוגמא לשימוש באותו מבנה של 30 מילים ב-80% מסשני ה-Cowork:

"I want to [task] to [success criteria]. Read my folder. Ask me questions using AskUserQuestion before you start."

לערוך את ההודעה במקום לשלוח הודעת המשך

08

ב-Chat, ניגוד ל-Cowork, אפשר ללחוץ **Edit** על ההודעה המקורית, לתקן אותה, ולייצר תשובה מחדש. החילוף הישן מוחלף - לא מצטבר.

כל פעם שאנחנו שולחים "לא, התכוונתי..." או "בעצם תשני את X ל-Y", אנחנו מוסיפים עוד היסטוריה לשיחה. כפתור העריכה מונע את זה לחלוטין.

בנוסף: זה מעולה גם כש-Claude פספס, כי אפשר פשוט לחזור אחורה.



לבחור את המוצר הנכון למשימה

09

לכל מוצר יש עלות טוקנים **אחרת** לכל אינטראקציה. כשמתאימים את הכלי למשימה, מפסיקים לשלם מחירי Cowork על עבודה שמתאימה ל-Chat.

מודל	מוצר	משימה
Haiku	Chat	שאלה מהירה
Opus	Cowork	כתיבת דוח על בסיס קבצים
Sonnet	Code	בניית גרף מנתונים
Sonnet/Opus	Chat → Research	מחקר מעמיק

אותו דבר נכון גם לפיצ'ר ה-Research בתוך Chat - שכמעט אף אחד לא משתמש בו.

הבסיסים שעדיין חשובים

לשמור קבצי ABOUT ME מתחת ל-2,000 מילים כל אחד

10

Cowork קורא את התיקיה שלנו לפני כל משימה. אם קובץ ה-**about-me** שלכם הוא 22,000 מילים, זה אומר שאלפי טוקנים נשרפים עוד לפני שהעבודה האמיתית מתחילה - בכל סשן ובכל משימה.

צמצמו אותו לפחות מ-2,000 מילים.

טיפ מקצועי: בסוף סשן Cowork, תנו את הפרומפט:

"*Write a session-notes.md with the key decisions and next steps*"

ובסשן הבא תתחילו עם: "*Read session-notes.md first*". ככה מעבירים את ההקשר קדימה בלי להסביר הכול מחדש.



להפעיל מחדש את השיחה במקום לשלוח הודעות המשך

11

כש-Cowork טועה, האינסטינקט הוא לכתוב "לא, התכוונתי..." ולשלוח עוד הודעה. כל הודעת המשך מצטברת על גבי כל היסטוריית השיחה. Claude קורא מחדש את הכול, שוב.

232K

טוקנים בסשן של 30 הודעות

105K

טוקנים בסשן של 20 הודעות

מכיוון שב-Cowork אי אפשר לערוך פרומפטים, עדיין אפשר לבחור "**Restart the conversation**" **from here** על הודעה קודמת. ככל שמאתחלים מנקודה מוקדמת יותר, חוסכים יותר טוקנים. אם כל הסשן התפזר - פתחו חדש.

לסכם ולהתחיל מחדש כל 15-20 הודעות

12

שיחות ארוכות הן תנור לשריפת טוקנים.

מפתח אחד עקב אחרי השימוש שלו וגילה ש-98.5% מהטוקנים הלכו על קריאה חוזרת של ההיסטוריה. רק 1.5% הוקדשו לפלט עצמו.

כשסשן Cowork מתארך, בקשו מ-Claude לסכם את כל מה שחשוב, העתיקו את הסיכום, פתחו סשן חדש, והדביקו אותו כהודעה הראשונה.

להשתמש ב-Sonnet או Haiku למשימות פשוטות

13

בדיקות דקדוק, סיעור מוחות, עיצוב מחדש של טקסט, תשובות קצרות. Sonnet מתמודד עם כל אלה בעלות נמוכה בהרבה.

Opus + Extended Thinking הם הצידוד הכבד שלכם. לא משתמשים בצידוד כבד כדי להזיז כיסא.

הכלל: אם ל-Claude לוקח פחות מ-30 שניות לענות, כנראה שלא צריך Opus. החליפו מודל לפני שמתחילים את הסשן. זה לוקח שתי לחיצות.

לא לזרוק את כל התיקייה שלכם לתוך Cowork

14

הרבה אנשים מכניסים 50 קבצים לתיקיית Cowork "למקרה שצריך". כל קובץ ש-Cowork קורא הוא טוקנים שנשרפים. ואם הקבצים גדולים מדי, Cowork מתחיל לסכם אותם בצורה רופפת במקום לקרוא אותם היטב.

אם Claude לא צריך את זה למשימה הזאת, הוא לא אמור לקרוא את זה.

 **במשימות שבכלל לא צריכות את הקבצים שלכם** - למשל טיוטת מייל מהירה עם connector - בחרו **אפס תיקיות** כשמתחילים את הסשן. אפס תיקיות = אפס הקשר מקומי = חיסכון לפני שהקלדתם משהו.

לפתוח צ'אט חדש כשמחליפים נושא

15

ביקשתם מ-Claude לעזור עם פוסט ללינקדאין. אחר כך שאלתם על הצעת מחיר ללקוח. אחר כך על מתכון. והכול באותו צ'אט.

Claude עדיין קורא מחדש את שיחת הלינקדאין ואת הצעת המחיר כל פעם שהוא חושב על ארוחת הערב שלך. כל ההודעות הישנות האלה הן משקל מת.

נושא חדש = צ'אט חדש. תמיד.

NEW

חיפוש ברשת, connectors ומצב Explore - כולם מוסיפים טוקנים לכל תשובה, גם כשלא צריך אותם.

כותבים תוכן משלכם? כבו Search ו-Tools. עושים בדיקת דקדוק פשוטה? כבו Extended Thinking.

ברירת המחדל: הכול כבוי. להדליק פיצ'רים לפי משימה, לא לפי חשבון. וכשמשתמשים ב-connectors, היו ספציפים: "חפשו ב-Slack מה-7 ימים האחרונים הודעות על השקת Q2" הרבה יותר זול מ"חפש ב-Slack כל דבר על השקות".



אם מעלים את אותו PDF לחמישה צ'אטים שונים, Claude עושה טוקניזציה למסמך הזה מחדש בכל פעם. חמישה צ'אטים, חמש קריאות מלאות.

במקום זה, השתמשו ב-**Projects**. מעלים את הקובץ פעם אחת, והוא נשמר בקאש. כל הצ'אטים העתידיים בפרויקט כבר מכירים אותו.

Anthropic אישרו שתוכן של פרויקט שנעשה בו שימוש חוזר לא נספר באותה צורה כמו העלאות חדשות. בתוכניות בתשלום, Projects עובדים גם עם RAG - כלומר Claude שולף רק את החלקים הרלוונטיים במקום לטעון את כל המסמך.



לכבות Memory ולהוסיף User Preferences

18

כל צ'אט חדש בלי הקשר שמור שורף 3-5 הודעות רק על הגדרות פתיחה: "אני משווקת, אני כותבת בצורה קלילה, אני מעדיפה פסקאות קצרות..."

אז תעשו את זה:

Settings → General → Personal preferences

כבו את Memory. הגדירו גם Styles דרך בורר המודלים. בחרו "Concise" או צרו סגנון מותאם. זה נשמר בין צ'אטים בלי לאכול את ההקשר. **הגדרה אחת, חיסכון קבוע.**

להשתמש במשימות מתוזמנות לעבודה חוזרת

19

אם מריצים את אותו דוח, דיג'סט או משימת מחקר כל שבוע, אל תעשו את זה ידנית בתוך סשן Cowork שהולך ומתארך.

השתמשו בתוסף `schedule/` כדי להגדיר משימות אוטומטיות שרצות לבד.

תנו ל-Claude Code מסגרת ברורה לפני שהוא מתחיל

20

סשנים ב-Code יכולים לשרוף טוקנים מהר יותר מכל דבר אחר, אם לא נזהרים. ל-Code יש נטייה להתרחב - הוא בודק קבצים, קורא תיקיות, מריץ בדיקות. אם לא אומרים לו בדיוק מה צריך, הוא יתחיל לחקור כל דבר שנמצא מולו.

היו ספציפיים:

"Create a bar chart from this CSV showing monthly revenue for 2025.
Save it as chart.png."

אל תשאירו מקום ל-Claude להתחיל לשוטט.

להשתמש בקובץ CLAUDE.md כדי להגדיר הקשר קבוע

21

Code קורא קובץ בשם **CLAUDE.md**, אם הוא קיים, לפני כל משימה. שימו שם את ההוראות הקבועות שלכם: באיזו תיקייה לעבוד, באיזו שפה להשתמש, מה מוסכמות השמות שלכם. אותו היגיון כמו Global Instructions ב-Cowork. כותבים פעם אחת, לא חוזרים על זה שוב.

Anthropic מזהירים שקבצי CLAUDE.md מנופחים גורמים ל-Claude להתעלם מההוראות האמיתיות שלכם. שמרו עליו קצר. אם יש תהליכי עבודה שמשמשים בהם רק לפעמים - העבירו אותם ל-Skills. Skills נטענים לפי צורך. CLAUDE.md נטען בכל פעם.



לפזר את העבודה לאורך היום

22

Claude משתמש בחלון מתגלגל של 5 שעות למגבלות שימוש. אם שורפים את כל המכסה בסשן אחד בבוקר, רוב הקיבולת היומית פשוט לא מנוצלת.

עדיף לפצל ל-2 עד 3 ששנים: בוקר, צהריים, ערב. עד שתחזרו, השימוש הקודם שלכם כבר ירד מהחלון.



shaniburshtein.com



קהילת AI Marketing Hub



SHANI BURSHTein

GenAI Marketing Transformation Expert

.LESS TOOLS, MORE SKILLS