



הפינה המקצועית של ירדן פלד

TECH עדיף

החשבון שאף אחד לא הראה לכם: כלכלת הטוקנים של פרויקט AI בסוכנות

לפני כמה שבועות הצגתי כאן כיצד כל סוכן הפך פתאום למפתח, ומדוע הקלות שבה בונים מערכת היא גם הסכנה הגדולה ביותר. הפעם אציג סכנה אחרת, שקטה יותר, שלא מופיעה על מסך מלוטש ולא נראית בדמו: העלות. כי בין הרגע שבו המערכת "עובדת" לבין הרגע שבו היא משרתת לקוחות בפועל, מסתתר משתנה אחד שיכול להפוך פרויקט מבריק לבור ללא תחתית, או לחלופין להפוך כלי יקר לכה שמשלם את עצמו פי עשרה. המשתנה הזה נקרא טוקן. בעולם הפרויקטים של ה-AI מתחילים להתייחס ל"כלכלת טוקנים". העלייה והירידה המהירה של המודל של קלוד - Claude Fable 5 בעל היכולות החזקות ביותר, עד שהממשל האמריקאי הנחה להוריד אותו, ממחישים היטב מצד אחד את היכולות הגבוהות של המודלים לעומת העלויות שלהם בפועל. אפשר להתלהב ולבנות במהירות, אבל יש לזה גם מחיר.

מודל	קלט (input)	פלט (output)	מדוע הוא קיים
Claude Haiku 4.5	1 דולר	5 דולר	מהיר וזול לסיווג, חילוף מידע, מענה לשאלות חוזרות
Claude Sonnet 4.6	3 דולר	15 דולר	האיזון בין מחיר לאיכות - רוב המשימות העסקיות
Claude Opus 4.8	5 דולר	25 דולר	המודל החזק - שיפוט מורכב, החלטות, קוד

* המחירים והמודלים משתנים ויש לוודא לפני ביצוע
למחשבון לבדיקת עלויות טוקנים לסוכנויות ביטוח - [לחצו כאן](#)

מה זה בעצם טוקן ומדוע אתם משלמים עליו פעמיים

ההיגיון פשוט אבל קריטי: אין שום סיבה להפעיל את המודל החזק והיקר ביותר על משימה שמודל זול מבצע אותה באיכות זהה. לשלוח שאלה כמו "מה שעות הפעילות של הסוכנות" למודל Opus זה כמו לשכור אקטואר בכיר כדי לענות לטלפון. הוא יענה מצוין ויעלה לכם פי חמישה ממה שצריך. עיקרון הזהב בכלכלת הטוקנים הוא התאמת המודל למשימה, לא הפעלה של הכי חזק "ליתר ביטחון".

דוגמה מהשטח: הצ'אטבוט ששרף תקציב

ניקח סוכנות בינונית שהקימה צ'אטבוט למענה ללקוחות. נניח 4,000 שיחות בחודש, כל שיחה כחמישה סבבי הודעות, וההקשר המצטבר מביא כל פנייה לממוצע של כ-30,000 טוקני קלט ו-2,500 טוקני פלט. אם הסוכנות בנתה את הצ'אטבוט על Opus 4.8 "כי זה הכי טוב", החשבון החודשי נראה כך: כ-600 דולר על הקלט וכ-250 דולר על הפלט, סך הכל כ-850 דולר בחודש על משימה של מענה לשאלות חוזרות.

כל מודל שפה לא "קורא" מילים, הוא מפרק אותן ליחידות קטנות שנקראות טוקנים. בעברית, טוקן הוא בערך חצי מילה עד מילה. כשאתם שולחים למודל הוראה, מסמך או שאלה, אתם משלמים על טוקני הקלט (input). כשהמודל עונה, אתם משלמים על טוקני הפלט (output) - וכאן הנקודה הראשונה שרוב הסוכנים מפספסים: הפלט יקר פי חמישה מהקלט, בכל המודלים. כלומר, מערכת שמייצרת תשובות ארוכות ומפורטות "סתם", שורפת כסף על כל מילה מיותרת.

עד היום זה נשמע זניח. כמה אגורות פה, כמה אגורות שם. הבעיה מתחילה כשמכפילים את השקל הזה באלפי פניות בחודש ושוכחים שבכל שיחה ההקשר מצטבר. כל הודעה חדשה בשיחה נושאת איתה את כל מה שנאמר קודם ואתם משלמים עליו מחדש, שוב ושוב.

ההבדל בין מודל יקר למודל זול, ולמה הוא קריטי

ספקיות ה-AI לא מוכרות מודל אחד אלא משפחה שלמה, ובכוונה.

המשך בעמוד הבא <<<

המשך מהעמוד הקודם <<<

נפח חוזרת ופשוטה, חוסכים בלי לחשוב פעמיים. במשימת שיפוט ששגיאה בה יקרה, משלמים על האיות בלב שלם.

איך זה משנה את הדרך שבה מתכננים פרויקט AI?

הטעות הנפוצה היא להתחיל מהשאלה "איזה מודל הכי חזק", במקום "מה המשימה ומה היא שווה". סוכנות שמתכננת נכון מתחילה אחרת. ראשית, היא ממפה כל תהליך לשתי שאלות: כמה פעמים זה רץ בחודש וכמה יקרה שגיאה. תהליך נפח גבוה עם סיכון נמוך הולך למודל הזול. תהליך נפח נמוך עם סיכון גבוה הולך למודל החזק. שנית, היא מודדת לפני שהיא מרחיבה. כל ספקיות ה-AI מאפשרות לראות צריכת טוקנים בפועל. הריצו פיילוט על מאה פניות, חישוב את העלות הממוצעת לפנייה והכפילו בנפח החודשי הריאלי. המספר הזה, ולא הדמו היפה, הוא מה שצריך לאשר את התקציב.

שלישית, היא בונה תקרות. הגדרת תקרת הוצאה חודשית והתראות הן לא פיננק אלא חובה ניהולית. הסיפור של החשבון המפתיע קורה כמעט תמיד כשאין מי ששם לב שהצריכה זינקה פי שלושה אחרי שינוי קטן בקוד.

שורה תחתונה

בטור הקודם כתבתי "לאמץ באומץ, לבנות באחריות". הפעם אוסיף נדבך שלישי: לתמחר בתבונה. מערכת AI אינה רכישה חד פעמית אלא הוצאה תפעולית שוטפת, וכמו כל הוצאה תפעולית בסוכנות, היא דורשת ניהול. הסוכנים שיובילו את העשור הבא לא יהיו רק אלה שבנו כלים נכון, אלא אלה שידעו בדיוק כמה כל כלי עולה להם ולמה. המסך היפה לא מראה את החשבון. אתם צריכים להראות אותו לעצמכם, לפני שהספק יראה אותו לכם.

הכותב הינו מנהל עדיףTech, בעלים של חברת InsurAI, מומחה להטמעת בינה מלאכותית ואוטומציה בענף הביטוח

אותה מערכת בדיוק, על Haiku 4.5, תעלה כ-120 דולר קלט וכ-50 דולר פלט, סך הכל כ-170 דולר בחודש. הפרש של פי חמישה, בלי שום פגיעה בחוויית הלקוח, מהסיבה שהמשימה לא דרשה את המודל היקר. על פני שנה זה ההבדל בין 10,200 דולר ל-2,040 דולר. 8,000 דולר שנשרפו על "ליתר ביטחון".

ויש כאן מנוף נוסף שכמעט אף אחד לא מנצל: prompt caching. ההוראות הקבועות של המערכת, אותו "מי אתה ואיך אתה מתנהג" שנשלח בכל פנייה, יכולות להישמר במטמון ולחסוך עד 90% מעלות הקלט החוזר. בציטבוט עם הקשר קבוע גדול, זה לבדו יכול לחתוך את החשבון בחצי.

דוגמה נוספת: תביעה עם עיבוד תמונה, וכאן ההיגיון מתהפך

עכשיו ניקח תהליך אחר לגמרי: טיפול בתביעת רכב הכולל ניתוח תמונות נזק וקבלת המלצת המשך. כל תמונה "שווה" כ-1,500 טוקנים. תביעה טיפוסית כוללת ארבע תמונות ועוד הקשר של תנאי הפוליסה והנחיות, ובסך הכל נגיע לכ-10,000 טוקני קלט ו-1,500 טוקני פלט לתביעה.

ב-500 תביעות בחודש, Opus 4.8 יעלה כ-44 דולר בחודש. Sonnet 4.6 יעלה כ-26 דולר. כאן ההפרש קטן בהרבה, ובמשימה הזו דווקא לא הייתי ממליץ לחסוך: ניתוח נזק וקבלת החלטה שמשיפיעה על תשלום תביעה הם בדיוק המקום שבו אתם רוצים את השיפוט הטוב ביותר. שגיאה כאן של המודל עולה הרבה יותר מהפרש של 18 דולר בחודש.

וזו בדיוק התובנה המרכזית: כלכלת טוקנים אינה "לחסוך תמיד". היא להתאים את ההשקעה לערך ולסיכון של המשימה. במשימת

עדיף

02.07.2026
ביתן 10, אקספו ת"א

לפרטים ולהרשמה <

סוכני המחר
הוועידה השנתית 2026
לסוכני ביטוח, יועצים פנסיונים ומתכננים פיננסיים

המסלקה הפנסיונית:
הדור הבא - לאן פנינו מועדות?
צמיר טל
מנכ"ל המסלקה הפנסיונית