

מחקר · verified

OpenAI משיקה מטמון פרומפטים משופר ב-GPT-6 עם הנחות של עד 90%

לצד ההנחות, OpenAI השיקה לוח מחוונים וכלי אבחון חדשים לניטור ביצועי המטמון

מאת מערכת וואי חדשות | פורסם 22 בספט' 2026, 23:46 | 1 דקות קריאה

רלוונטי ל: מפתחים ואנשי טכנולוגיה | יזמים ומשקיעים | עסקים ועצמאים

הסיפור ב-30 שניות

- ✓ GPT-6 כולל מערכת שמירת-פרומפט במטמון עם שיעורי פגיעה גבוהים יותר כברירת מחדל
- ✓ מפתחים יכולים לקבל הנחות של עד 90% על טוקנים שנשמרו במטמון
- ✓ ההנחות חלות על פרפיקסים משותפים שנעשה בהם שימוש חוזר בתוך חלון של 30 דקות
- ✓ OpenAI השיקה לוח מחוונים חדש וכלי אבחון לזיהוי כשלי מטמון
- ✓ GPT-6 מאפשר לסוכנים לעבוד שעות על משימות מורכבות כגון שכתוב קוד והפקת מסמכים

מה קרה?

OpenAI הכריזה על שיפורים במערכת שמירת-הפרומפט במטמון של GPT-6, הכוללת שיעורי פגיעה גבוהים יותר כברירת מחדל. מפתחים המשתמשים ב-API יוכלו לקבל הנחות של עד 90% על טוקנים שנשמרו במטמון, בתנאי שהפרפיקסים המשותפים נעשה בהם שימוש חוזר בתוך חלון זמן של 30 דקות. במקביל, החברה השיקה לוח מחוונים חדש לניטור ביצועי המטמון וכלי אבחון לזיהוי סיבות לכשלי מטמון בבקשות GPT-6. API תומך גם בסוכנים המסוגלים לעבוד שעות על משימות מורכבות, כגון שכתוב קוד והפקת מסמכים.

למה זה חשוב?

הנחות של עד 90% על טוקנים חוזרים עשויות להפחית משמעותית את עלויות ה-API עבור מפתחים הבונים יישומים עם הקשרים ארוכים או חוזרים. כלי הניטור והאבחון החדשים מוסיפים שקיפות שעד כה הייתה מוגבלת.

מה עדיין לא ידוע?

לא ידוע מתי בדיוק הושקו השינויים, לאילו תוכניות תמחור הם חלים, ומהו שיעור ההנחה המדויק מעבר לתקרה של 90%.

שקיפות ואימות

סטטוס: 1. verified. מקורות מקושרים. סטטוס הבדיקה האנושית אינו מוצג.

מקורות

1. OpenAI News

official

<https://openai.com/index/better-prompt-caching-for-gpt-6>

המסמך הוא גרסה נוחה לקריאה של ידיעת וואי חדשות. הגרסה הקנונית באתר כוללת את סטטוס האימות והעדכונים האחרונים. כלי AI עשויים לסייע בגילוי, השוואה וניסוח; האחריות לפרסום כפופה למדיניות המערכת.

<https://y-community.co.il/news/better-prompt-caching-for-gpt-6-fd3cf06f8761>