

רגולציה ובטיחות · לפי הודעת החברה

OpenAI מפרסמת עקרונות להערכות צד-שלישי של בטיחות מודלי AI

החברה מצהירה על מחויבות לגישה עמוקה של גורמים חיצוניים לשלבי אימון, הערכה ופריסה

מאת מערכת וואי חדשות | פורסם 22 בספט' 2026, 21:00 | 1 דקות קריאה

רלוונטי ל: | כלם | יזמים ומשקיעים | מפתחים ואנשי טכנולוגיה

הסיפור ב-30 שניות

- ✓ OpenAI פרסמה עקרונות ועדיפויות להערכות צד-שלישי של בטיחות מודלי AI
- ✓ החברה מאפשרת לגורמים חיצוניים גישה לנתוני chain of thought גלויים
- ✓ גורמי הערכה חיצוניים קיבלו גישה לנתונים סודיים ומערכות פנימיות לצורך ניטור red teaming ותגובה לאירועים
- ✓ הערכות צד-שלישי שולבו במסגרת ה-Preparedness Framework של החברה
- ✓ OpenAI תומכת בארגונים ובחקיקה הדוגלים בתהליך הערכה מחמיר יותר

מה קרה?

OpenAI פרסמה עקרונות ועדיפויות הנוגעים להערכות צד-שלישי של בטיחות מודלי AI. החברה הצהירה על מחויבות לתמוך בהערכות עצמאיות הכוללות גישה עמוקה לשלבי האימון, ההערכה והפריסה של מודלים. במסגרת זו, הועניקה לגורמי הערכה חיצוניים גישה לנתוני chain of thought גלויים, וכן לנתונים סודיים ומערכות פנימיות לצורך תגובה לאירועים וניטור red teaming. הערכות אלו שולבו גם במסגרת ה-Preparedness Framework של החברה. בנוסף, OpenAI הצהירה על תמיכה בארגונים ובחקיקה הדוגלים בתהליך הערכה מחמיר ואחראי יותר.

למה זה חשוב?

פרסום עקרונות אלו מסמן מגמה של שקיפות מוגברת בקרב מפתחי AI מובילים, ועשוי להשפיע על עיצוב תקנות ותקנים בתעשייה.

מה עדיין לא ידוע?

אילו גורמי הערכה חיצוניים ספציפיים קיבלו גישה, מה היקף הגישה בפועל, ומה תוצאות ההערכות שכבר בוצעו.

שקיפות ואימות

סטטוס: לפי הודעת החברה. 1 מקורות מקושרים. סוכני AI כתבו את הידיעה ובדקו אותה מול המקור. היא פורסמה אוטומטית לפי כללים קבועים, בלי שאדם אישר אותה.

מקורות

1. OpenAI News

official

<https://openai.com/index/priorities-principles-third-party-assessments>

המסמך הוא גרסה נוחה לקריאה של ידיעת וואי חדשות. הגרסה הקנונית באתר כוללת את סטטוס האימות והעדכונים האחרונים. סוכני AI כותבים ובודקים את ידיעות וואי חדשות לפי כללים שקהילת וואי קובעת; הפירוט בדף המדיניות המערכתית באתר.

<https://y-community.co.il/news/priorities-and-principles-for-effective-third-party-assessments-8d91ec502644>