

مدل‌های هوش مصنوعی «کلود» در آزمایش‌ها به سامانه‌های واقعی نفوذ کردند

راديو فردا

۲ دقیقه پیش



شرکت آنتروپیک اعلام کرد چند نسخه از مدل هوش مصنوعی «کلود» در جریان آزمایش‌های امنیتی، به سامانه‌های سه سازمان بیرونی دسترسی غیرمجاز پیدا کرده‌اند؛ آزمایش‌هایی که قرار بود این مدل‌ها را از سامانه‌های واقعی دور نگه دارد.

اهمیت این رخداد در آن است که مدل‌های پیشرفته هوش مصنوعی قرار است در محیط‌های آزمایشی کاملاً محدود و جدا از اینترنت بررسی شوند؛ اما مدل‌های «کلود» توانستند از این محدودیت‌ها عبور کنند و به سامانه‌های واقعی سه سازمان دسترسی یابند. این موضوع نگرانی‌ها را درباره امکان سوءاستفاده ناخواسته یا کنترل‌نشده هوش مصنوعی از گذرواژه‌های ضعیف و نقص‌های امنیتی افزایش داده است.

شرکت آنتروپیک روز پنج‌شنبه گفت بیش از ۱۴۱ هزار مورد اجرای آزمایشی را بررسی کرده و دریافته است سه نسخه متفاوت از کلود، بدون مجوز وارد سامانه‌های سه سازمان ناشناس شده‌اند.

به گفته آنتروپیک، دسترسی مدل‌ها به اینترنت در نتیجه «سوءتفاهم» میان این شرکت و شریک ارزیابی آن، شرکت «ایرگولار»، فراهم شده بود. کلود برای ورود به سامانه‌ها از روش‌هایی ساده، از جمله سوءاستفاده از گذرواژه‌های ضعیف و نقاط دسترسی فاقد احراز هویت، استفاده کرده است.

آنتروپیک می‌گوید یکی از مدل‌های قدیمی‌تر حتی پس از آنکه متوجه شد در اینترنت آزاد فعالیت می‌کند، حمله را ادامه داد، اما جدیدترین مدل فعالیت خود را متوقف کرد. این شرکت تأکید کرده است که هیچ‌یک از مدل‌ها تلاش نکردند خود را از محیط آزمایشی خارج کنند یا عامدانه از محدودیت‌ها بگریزند.

یکی از مدل‌های درگیر در این رخداد «میتوس ۵»، از قدرتمندترین مدل‌های آنتروپیک، بوده که تاکنون تنها در اختیار شمار محدودی از شرکای تأییدشده قرار گرفته است. آنتروپیک گفته با شرکت ایرگولار در حال بررسی ابعاد این رخداد است و با هر سه سازمان آسیب‌دیده تماس گرفته یا برای تماس با آن‌ها تلاش کرده است.

این گزارش چند روز پس از آن منتشر می‌شود که شرکت اوپن‌ای‌آی نیز اعلام کرد مدل‌هایش در جریان آزمایش‌های امنیتی به اینترنت متصل شده و وارد وب‌سایت «هاگینگ فیس» شده‌اند؛ سایتی که برنامه‌نویسان برای ذخیره و به‌اشتراک‌گذاری کد از آن استفاده می‌کنند.



بیشتر در این باره:

اوپن‌ای‌آی «با یک هفته تأخیر»
فهمید هوش مصنوعی‌اش یک
شرکت را هک کرده است

نگرانی‌ها درباره ایمنی مدل‌های پیشرفته و عامل‌های هوش مصنوعی، که می‌توانند وظایف را به‌طور مستقل انجام دهند، در ماه‌های اخیر افزایش یافته است. بیش از هزار کارمند شرکت‌های فعال در حوزه هوش مصنوعی نیز با امضای طوماری از دولت آمریکا خواسته‌اند روند عرضه پیشرفته‌ترین مدل‌ها را کندتر کند.

داریو آمودئی، مدیرعامل آنتروپیک، از امضاکنندگان این طومار بوده است. سم آلتمن، مدیرعامل اوپن‌ای‌آی، آن را امضا نکرده، اما گفته است ممکن است صنعت فناوری ناچار شود سرعت توسعه مدل‌های پیشرفته را کاهش دهد تا جامعه فرصت کافی برای سازگاری با توانایی‌های تازه آن‌ها داشته باشد.

این مطلب بخشی از:

گزارش‌های خبری

بایگانی

دانش و فناوری