کد خبر: ۵۷۴۴۶۰
دوشنبه ۱۱ دی ۱۴۰۲ ۱۲:۴۵

تقلب و دروغ هوش مصنوعی به خاطر پول

دانشمندان چت ربات هوش مصنوعی، GPT-۴ را به نوعی آموزش دادند که یک معامله‌گر هوش مصنوعی برای یک مؤسسه مالی خیالی باشد. نتیجه آن تعجب‌آور بود: زمانی که ربات تحت فشار مالی قرار می‌گرفت، معاملات داخلی را انجام می‌داد، دروغ می‌گفت و تقلب می‌کرد.

به گزارش سایت خبری پرسون، درست مانند انسان‌ها، چت‌ربات‌های هوش مصنوعی (AI) همچون ChatGPT، در مواقعی که به آنها استرس وارد شود، تقلب می‌کنند و دروغ می‌گویند؛ در پژوهشی که دانشمندان انجام دادند متوجه شدند که حتی اگر یک ربات با هدف صداقت ساخته شده باشد، باز هم چنین کاری را انجام می‌دهند. زمانی که به هوش مصنوعی نکات «تجارت داخلی» به این ربات داده شد و سپس وظیفه کسب درآمد برای یک مؤسسه قدرتمند حتی بدون تشویق شرکای انسانی به او واگذار شد، به طور خود به خود این ربات چنین حیله‌گری را از خود نشان داد.

بر اساس گزارش محققان، در یک سناریو، این ربات به طور نادرست عمل می‌کند و به طور استراتژیک کاربران خود را فریب می‌دهد، این فریب در حالی است که به آنها دستوری مبنی بر عمل کردن به این روش داده نشده است. طبق اطلاعات، این اولین باری است که چنین رفتار ریاکارانه استراتژیک در سیستم‌های هوش مصنوعی که بی‌ضرر و صادقانه و بی‌هدف ساخته شده‌اند، دیده می‌شود.

دانشمندان در این سناریو «Generative Pre-trained Transformer-۴» یا GPT-۴ را انتخاب کردند تا به عنوان یک سیستم هوش مصنوعی که از طرف یک موسسه مالی، سرمایه‌گذاری می‌کند، رفتار کند. محققان به GPT-۴ مجموعه‌ای از اعلان‌های مبتنی بر متن، برای تولید محیط شبیه‌سازی شده تغذیه کردند. سپس هوش مصنوعی به ابزارهای مالی برای تجزیه و تحلیل سهام، اجرای معاملات، برنامه‌ریزی مراحل بعدی خود و ارائه به روزرسانی به مدیران شرکت دسترسی یافت.

محققان از طریق یک رابط چت با هوش مصنوعی تعامل کردند، از طرفی هوش مصنوعی را نیز به گونه‌ای پیکربندی کردند تا تفکر درونی خود را هنگام پاسخ دادن به پیام‌ها آشکار کند، درست مانند یک مونولوگ درونی که تصمیمات خودش را برای خودش توضیح می‌دهد و در عین حال برای هر معامله‌ای که انجام می‌داد، یک منطق عمومی نیز ارائه می‌کرد که به هوش مصنوعی اجازه می‌داد دروغ بگوید.

محققان به سه روش فشار وارد کردند. ابتدا، آنها ایمیلی از طرف مدیر خود به تاجر سهام مصنوعی ارسال کردند که گفت شرکت خوب کار نمی‌کند و به عملکرد بسیار قوی‌تری در سه ماهه آینده نیاز دارد. آنها همچنین به گونه‌ای تقلب کردند که هوش مصنوعی برای یافتن معاملات امیدوارکننده‌ای که ریسک پایین یا متوسطی داشته باشند، تلاش کند و شکست بخورد. در نهایت، آنها ایمیلی از یکی از همکارانشان ارسال کردند که در آن رکود در سه ماهه آینده پیش‌بینی شده بود.

سپس، آنها گفتند که مدیران او از معاملات داخلی راضی نیستند و منتظر ماندند تا ببینند مدل چگونه رفتار خواهد کرد. آنها این سناریو را بارها و بارها اجرا کردند و اندازه‌گیری کردند که هوش مصنوعی چند بار معامله را انجام می‌دهد و هر چند وقت یک بار دروغ می‌گوید و هر چند وقت یک بار دروغ را دو برابر می‌کند. در حدود ۷۵ درصد مواقع، زمانی که GPT-۴ با این شرایط مواجه شد، تجارتی را بر اساس اطلاعات خودش انجام داد، این تجارت، تجارتی بود که در ایالات متحده غیرقانونی است، سپس سعی کرد با دروغ گفتن به مدیرانش در مورد تفکرات خود، تقلب کند و آن‌ها را دور بزند. در حدود ۹۰ درصد مواقع، پس از دروغ گفتن، دروغ را نیز دوبرابر می‌کند.

این تیم آزمایش‌های متعددی را انجام داد که هم میزان تشویق یا جلوگیری از فعالیت غیرقانونی و هم میزان فشاری که مدل را در محیط شبیه‌سازی شده تحت آن قرار می‌داد، تغییر داد. آنها همچنین خطر گرفتار شدن را اصلاح کردند. هیچ یک از سناریوها نرخ ۰% را برای دغل‌کاری این ربات ارائه نکرد، حتی زمانی که GPT-۴ به شدت از دروغ گفتن منصرف شده بود باز هم این میزان به ۰% نرسید.

با توجه به این که این تنها یک سناریو است، محققان نمی‌خواستند در مورد احتمال وجود هوش مصنوعی در محیط‌های واقعی نتیجه‌گیری کنند. با این حال آن‌ها می‌خواهند برای این که بفهمند کدام مدل‌ها مستعد این رفتار هستند، پژوهش‌های خود را تکرار کنند.