چهار شنبه 27 خرداد 1405 - 06:10

هوش مصنوعی هنوز حریف ریاضیدانان برتر نمی‌شود

یک معیار جدید که هوش مصنوعی را در برابر مسائل ریاضی از پیش دیده نشده قرار می‌دهد، نشان می‌دهد که این سامانه‌ها هنوز تا رسیدن به سطح برترین متخصصان انسانی فاصله دارند.

یک معیار جدید که هوش مصنوعی را در برابر مسائل ریاضی از پیش دیده نشده قرار می دهد، نشان می دهد که این سامانه ها هنوز تا رسیدن به سطح برترین متخصصان انسانی فاصله دارند.

به گزارش ایسنا، هوش مصنوعی دقیق ترین آزمون ریاضی خود را تاکنون پشت سر گذاشته است. نتایج آماده است و مدل های هوش مصنوعی که در آن شرکت کردند، به مهارت های حل مسئله ریاضیدانان برتر نرسیدند.

به نقل از نیچر، این آزمون بخشی از پروژه ای به نام First Proof به معنای اثبات اول است که هدف از آن ارزیابی توانایی هوش مصنوعی در حل سوالات پیچیده در ریاضیات است. ۱۰ مسئله ریاضی در سطح تحقیق برای چهار سیستم هوش مصنوعی مطرح شد. سپس هیئت منصفه ای از متخصصان انسانی ناشناس در زمینه های ریاضی مربوطه، پاسخ های مدل ها را ارزیابی کردند.

این آزمون اولین آزمون از نوع خود بود که به طور همزمان سه شرط کلیدی را برآورده می کرد: اول، شامل سوالات ریاضی در سطح تحقیق بود؛ دوم، شامل مسائلی بود که در داده های آموزشی ظاهر نشده بودند و سوم، به طور رسمی توسط ریاضیدانان درجه بندی شده بود.

این یافته ها پس از پیشرفت های اخیر هوش مصنوعی در حل مسائل ریاضی رخ داده است. برای مثال، ماه گذشته، یک چت بات ساخته شده توسط شرکت فناوری اوپن ای آی در سانفرانسیسکو، کالیفرنیا، یک چالش ریاضی ۸۰ ساله را که توسط ریاضیدان فقید، پاول اردوش، مطرح شده بود، حل کرد. گروه فرست پروف می گوید که تکرارهای آینده این آزمون می تواند به محققان کمک کند تا قضاوت کنند که مدل های هوش مصنوعی چقدر می توانند به عنوان مثال، در حل خودکار مسائل، بررسی اثبات ها یا ایفای نقش دستیاران تحقیق برای ریاضیدانان مفید باشند.

یکی از نوآوری های مهم آزمون فرست پروف این بود که سوالات قبلا در هیچ کجای مقالات منتشر شده یا در اینترنت ذکر نشده بودند و این خطر را که مدل ها به سادگی اطلاعاتی را که در طول آموزش خود آموخته اند، تکرار کنند، از بین می برد. در عوض، ۱۰ محقق از طیف گسترده ای از تخصص های ریاضی، هر کدام سوالی را ارائه دادند که در جریان تحقیقات خود حل کرده بودند، اما هنوز منتشر نکرده بودند.

فرست پروف در ماه فوریه یک آزمون آزمایشی با دسته ای متفاوت از مسائل جدید برگزار کرد. در آن دور، هر کسی می توانست سیستم های هوش مصنوعی مورد علاقه خود را روی مسائل امتحان کند و بسیاری از گروه ها این کار را انجام دادند اما نتایج به طور رسمی تأیید نشد. همچنین هیچ راهی برای بررسی مستقل اینکه هوش مصنوعی کمکی از انسان ها دریافت نکرده است یا خیر، وجود نداشت. این بار، فرست پروف خودش آزمایش را اجرا کرد: گروه از مدل ها خواست تا مسائل را به روشی کاملاً خودکار حل کنند و گروهی متشکل از ۳۰ ریاضیدان برای بررسی پاسخ ها به کار گرفته شدند. جرمی آویگاد، ریاضیدان و رئیس موسسه استدلال به کمک رایانه در ریاضیات در دانشگاه کارنگی ملون در پیتسبورگ، پنسیلوانیا، می گوید: برگزارکنندگان به وضوح با دقت بیشتری به دسته دوم فکر کرده اند تا آن را کنترل شده تر و سیستماتیک تر کنند.

قانون دیگر این بود که مدل های شرکت کننده باید در دسترس عموم باشند. این بدان معنا بود که Aletheia گوگل که سیستمی است که به طور خاص برای حل مسائل ریاضی طراحی شده است و نسخه کامل و منتشر نشده Claude Mythos، مدلی که توسط آنتروپیک در سانفرانسیسکو، کالیفرنیا ساخته شده است، قابل استفاده نبودند. اوپن ای آی تنها شرکت بزرگی بود که با مدل ChatGPT ۵.۵ Pro خود در این آزمایش شرکت کرد.

سیستم های دیگر توسط سه گروه دانشگاهی، از دانشگاه کالیفرنیا، لس آنجلس (UCLA، دانشگاه پرینستون در نیوجرسی و موسسه فناوری فدرال سوئیس (ETH) در زوریخ ارائه شدند. هر سه دانشگاه، «مهارهایی» بر روی چت بات های موجود، مانند چت جی پی تی، جمینای گوگل و نسخه عمومی کلاود آنتروپیک ساختند. مهار، سیستمی خودکار است که از یک چت بات سوالی می پرسد و پاسخ آن توسط یک چت بات دیگر، اغلب با رفت و برگشت های مکرر، بررسی می شود.

نتایج ریاضی

مدل تیم موسسه فناوری فدرال سوئیس بهترین عملکرد را داشت و ۶ از ۱۰ مسئله را با سیستمی حل کرد که در آن پاسخ های چت جی پی تی توسط یک «شورای مشورتی» متشکل از هر سه چت بات اصلی بررسی شده یا بهبود یافته بود. تیم دانشگاه کالیفرنیا، لس آنجلس که یک مهار بر روی چت جی پی تی ساخته بود، دومین تیم برتر بود و پس از آن تیم چت جی پی تی اوپن ای آی بدون مهار و پرینستون مهاره ای که عمدتا از Gemini ۳.۱ Pro به عنوان پشتیبان خود استفاده می کند، قرار گرفتند.

یوهانس اشمیت، ریاضیدانی که عضوی از تیم موسسه فناوری فدرال سوئیس بود، می گوید که برای تنظیم دقیق سیستم خود قبل از مسابقه، او و همکارانش با جامعه ریاضی گسترده تر تماس گرفتند و از آنها مسائل را درخواست کردند. پاسخ شگفت انگیز بود: ظرف چند روز، ۳۰ مسئله ارسالی از حوزه های مختلف ریاضیات دریافت کردیم و مردم بسیار کنجکاو و روشنفکر بودند.

لورن ویلیامز، ریاضیدان دانشگاه هاروارد در کمبریج، ماساچوست و عضو تیم فرست پروف، می گوید: مشخص نیست که آیا مسائل حل نشده لزوما سخت تر از بقیه بوده اند یا خیر. او می افزاید: من فکر می کنم مسائلی که حل نشده بودند، چه از نظر موضوع و چه از نظر ایده های اثبات، از چیزهایی که قبلا در مقالات علمی آمده بودند، دورتر بودند.

مدل های استدلال همچنین مستعد توهم یا تولید خروجی های واقعا نادرست بودند، حتی زمانی که صریحا به آنها گفته می شد که منابع خود را بررسی کنند که مشکلی شناخته شده در مدل های زبانی بزرگ است.

ویلیامز می گوید که از کمبود «شدید» استناد در تمام پاسخ های مدل های هوش مصنوعی شگفت زده شده است به ویژه در

مورد مسئله ۲، که چندین مدل با اقتباس از روشی که یک مسئله مشابه در گذشته توسط انسان ها حل شده بود، آن را حل کردند. چندین راه حل، در برخی موارد، کپی کردن عبارات از مقاله قبلی به صورت خط به خط و استفاده مجدد از نمادگذاری ها و اصطلاحات دقیق بود، اما هرگز به آن مقاله در هیچ کجا استناد نکردند.
اکنون که مسائل فرست پروف منتشر شده اند، شرکت هایی که رسما در آن شرکت نکرده اند، احتمالا از آنها برای آزمایش غیررسمی سیستم های خود استفاده خواهند کرد. کوین بارتو، ریاضیدان دانشگاه کمبریج انگلستان که معیارهای ریاضی غیررسمی خود را برای هوش مصنوعی اجرا کرده است، می گوید: شخصا از دیدن مدل های داخلی آزمایش شده از سه آزمایشگاه لذت می بردم، فقط برای اینکه ببینم مرز واقعی در حال حاضر کجاست.