آیا هوش مصنوعی فوق هوشمند به ما حمله خواهند کرد؟ مطالعات آن را رد می‌نمایند.

بهبود عملکرد ابزارهایی مانند ChatGPT بیش از آن چه پیش‌بینی می‌شود است.

آیا یک ابرهوش مصنوعی[1] (AI) به طور ناگهانی ظاهر می‌شود یا دانشمندان تماشاگر ظهور آن خواهند بود و فرصتی برای هشدار به جهان خواهند داشت؟ این سوالی است که با ظهور مدل‌های زبان عظیم، مثل مدل‌های ChatGPT که با بزرگتر شدن اندازه‌شان به توانایی‌های جدید گسترده‌ای دست یافته‌اند، بسیار مورد توجه قرار گرفته است. برخی از یافته‌ها به «ظهور» اشاره می‌نمایند، پدیده‌ای که در آن مدل‌های هوش مصنوعی به روشی دقیق و غیرقابل پیش‌بینی هوش به دست می‌آورند. اما مطالعه‌ای که در نوامبر گذشته در کنفرانس یادگیری ماشین NeurIPS 2023 در نیواورلان-لوییزیانا ارائه شد، این موارد را امری خیالی نامید - مصنوعاتی که از نحوه آزمایش سیستم‌ها ناشی می‌شوند - و نشان می دهد که توانایی‌های نوآورانه در عوض به تدریج ایجاد می شوند.

دبورا راجی[2]، دانشمند کامپیوتر در بنیاد موزیلا در سانفرانسیسکو، کالیفرنیا، می‌گوید: «فکر می‌کنم آن‌ها کار خوبی کردند که گفتند «هیچ چیزِ جادویی اتفاق نیفتاده است». این یک "انتقاد واقعاً خوب، محکم و مبتنی بر اندازه گیری" است.»

هرچه بزرگتر بهتر

مدل‌های زبانی عظیم معمولاً با استفاده از حجم زیادی از متن یا اطلاعات دیگر در کنار هم قرار داده می‌شوند که از آن برای ایجاد پاسخ‌های واقع بینانه با پیش‌بینی موارد بعدی استفاده می‌کنند. به‌طور معمول، هر چه مدل بزرگتر باشد – برخی دارای بیش از صد میلیارد پارامتر قابل تنظیم هستند

[1] Artificial intelligence
[2] Deborah Raji

– عملکرد بهتری دارد. برخی از محققان گمان می‌کنند که این ابزارها در نهایت به هوش عمومی مصنوعی[1] (AGI) دست می‌یابند که در بیشتر کارها با انسان‌ها همخوانی دارند و حتی از آن‌ها فراتر می‌روند.

تحقیقات جدید ادعاهای ظهور را به روش‌های مختلفی مورد آزمایش قرار دادند. در یک رویکرد، دانشمندان توانایی های چهار سایز مدل GPT-3 را که توسط OpenAI در سانفرانسیسکو توسعه داده شده است، برای جمع‌آوری اعداد چهار رقمی مقایسه نمودند. با نگاهی به دقتِ مطلق، عملکرد بین سایز سوم و چهارم مدل از نزدیک به 0٪ تا نزدیک به 100٪ متفاوت است. اما اگر تعداد بیشتر در نظر گرفته شود، جواب بدست آمده قابل پیش‌بینی‌تر خواهد بود. این محققان همچنین دریافتند که می‌توانند منحنی را با دادن سؤالات تستی بیشتر به مدل‌ها کاهش دهند – در این مورد، مدل‌های کوچکتر گاهی اوقات به درستی پاسخ می‌دهند.

"مدل‌ها در حال بهبود هستند، اما هنوز به هوشیاری نزدیک نشده اند."

در مرحله بعد، محققان عملکرد مدل زبان LaMDA گوگل را در چندین کار بررسی نمودند. کارهایی که برای آن‌ها جهش ناگهانی در هوش ظاهری نشان داد، مانند تشخیص کنایه یا ترجمه ضرب المثل‌ها، اغلب کارهای چند گزینه‌ای بودند که پاسخ‌ها به طور مجزا به‌عنوان درست یا غلط نمره گذاری می‌شدند. در عوض، وقتی محققان احتمالاتی را که مدل‌ها روی هر پاسخ قرار می‌دهند – یک متریک پیوسته – بررسی نمودند، نشانه‌های ظهور ناپدید شدند. در نهایت، محققان به کامپیوتر بصری روی آوردند، حوزه ای که کمتر ادعای ظهور در آن وجود دارد. آن‌ها مدل‌هایی را برای فشرده سازی و سپس بازسازی تصاویر آموزش دادند. با تعیین آستانه‌ای دشوار برای صحت سنجی، آن‌ها می‌توانند ظهور ظاهری را القا نمایند. یجین چوی[2]، دانشمند کامپیوتر در دانشگاه واشنگتن در سیاتل می‌گوید: «آن‌ها در روشی که تحقیقات خود را طراحی نمودند خلاق بودند.

هیچ چیز منتفی نیست

یکی از نویسندگان مطالعه، ، سانمی کایجو[3] دانشمند کامپیوتر در دانشگاه استنفورد در پالو آلتو، کالیفرنیا،

[1] Artificial general intelligence
[2] Yejin Choi
[3] Sanmi Koyejo

می‌گوید که پذیرفتن ایده ظهور برای مردم غیر منطقی نبود، زیرا برخی از سیستم‌ها «تغییرات فاز» ناگهانی را نشان می‌دهند. ایشان همچنین خاطرنشان می‌نماید که این مطالعه نمی‌تواند به‌طور کامل ظهور در مدل‌های زبانی عظیم را رد کند - چه رسد به سیستم‌های آینده - اما اضافه می‌کند که "مطالعات علمی تا به امروز قویاً نشان می‌دهد که بیشتر جنبه‌های مدل‌های زبانی واقعاً قابل پیش بینی هستند".

راجی از اینکه می‌بیند جامعه هوش مصنوعی به جای توسعه معماری شبکه‌های عصبی، توجه بیشتری به محک‌ها می‌کند خوشحال است. ایشان مایل است که محققان حتی فراتر رفته و بپرسند که وظایف چقدر با استقرار آن در دنیای واقعی مرتبط است. به‌عنوان مثال، آیا آزمون LSAT که برای وکلای مشتاق است، اگر که GPT-4 انجامش دهد، می‌توان گفت که یک مدل می‌تواند به عنوان یک وکیل حقوقی عمل کند؟

این کار همچنین پیامدهایی برای ایمنی و سیاست هوش مصنوعی دارد. راجی می‌گوید: «جمعیت AGI از ادعای قابلیت‌های نوظهور استفاده کرده‌اند. ترس بی دلیل می تواند منجر به خفقان مقررات یا انحراف توجه از خطرهای دیگر شود. او می‌گوید: «مدل‌ها در حال بهبود هستند و این پیشرفت‌ها مفیدند. اما آن‌ها هنوز به هوشیاری نزدیک نشده‌اند.»

ترجمه و ویرایش: یاسمن باغبان

Reference

Nature | Vol 625 | 11 January 2024

https://www.nature.com/articles/d41586-023-04094-z

DOI: https://doi.org/10.1038/d41586-023-04094-z