پنجشنبه 25 آبان 1402 - 07:05

نگرانی دانشمندان از تعصب «چت جی‌پی‌تی»

مطالعات تأیید می‌کنند که هوش مصنوعی چت جی‌پی‌تی که محبوب‌ترین مدل زبانی هوش مصنوعی در جهان است در پاسخ‌های خود دارای تعصب است.

مطالعات تأیید می کنند که هوش مصنوعی چت جی پی تی که محبوب ترین مدل زبانی هوش مصنوعی در جهان است در پاسخ های خود دارای تعصب است.

به گزارش ایسنا و به نقل از آی ای، براساس مطالعه محققان دانشگاه کالیفرنیا، برکلی، چت جی پی تی متعلق به شرکت اُپن ای آی(OpenAI) که در حال حاضر محبوب ترین مدل زبان مبتنی بر هوش مصنوعی در جهان به حساب می آید، در اطلاعاتی که ارائه می کند بی طرف نیست. مقاله آنها نشان می دهد که از آنجایی که چت جی پی تی از حجم عظیمی از مقالات دارای حق نشر به عنوان مجموعه داده ای که بر اساس آن آموزش یافته و پاسخ های خود را پایه گذاری می کند، استفاده می کند، سوگیری ذاتی در آن داده ها به نتایج چت جی پی تی نیز راه پیدا می کند.

محققان در مطالعه خود نوشتند که مدل های اُپن ای آی تعداد زیادی از انواع مختلف مقالات دارای حق نشر را در خود جای داده اند و این فرآیند تحت تاثیر تعداد دفعاتی است که قسمت های آن مقالات در فضای وب ظاهر می شوند. به این ترتیب، اطلاعات مرتبط با مقالاتی که در فرآیند آموزش بیشتر مورد استفاده قرار گرفته اند، نسبت به اطلاعاتی که به طور مکرر مورد استفاده قرار نگرفته اند، بیشتر در پاسخ ها گنجانده می شوند.

همانطور که محققان در مورد این مطالعه توضیح دادند، دقت چنین مدل هایی به شدت به فرکانس مشاهده اطلاعات در داده های آموزشی وابسته است و این، توانایی آنها در تعمیم را زیر سوال می برد.

یکی از نمونه های آن این است که کتاب های علمی تخیلی و فانتزی ظاهرا در فهرست کتاب هایی که برای آموزش استفاده می شوند بیش از حد استفاده شده اند. این یک «سوگیری درونی» ایجاد می کند تا آنجا که می توان گفت از چت جی پی تی چه نوع پاسخ هایی را می توان انتظار داشت.

این مطالعه اشاره می کند که کتاب هایی که در فهرست داده های آموزشی چت جی پی تی به طور غالب حضور داشته اند، عناوین محبوبی مانند هری پاتر، ۱۹۸۴، ارباب حلقه ها، بازی های گرسنگی، راهنمای مسافران مجانی کهکشان، فارنهایت ۴۵۱، بازی تاج و تخت، و تپه های شنی را شامل می شوند.

آنچه در داده های آموزشی پیدا نمی کنید

در حالی که چت جی پی تی موارد زیادی در مورد مقالات موجود در حوزه عمومی می داند، آنچه در داده های آموزشی نمی بینید، عناوین کمتر شناخته شده ای مانند به اصطلاح ادبیات جهانی آنگلوافون هستند. این ها کتاب هایی هستند که برای و توسط مخاطبانی به جز کشورهای اصلی انگلیسی زبان نوشته شده اند.

چنین مناطقی شامل آفریقا، آسیا و کارائیب می شود.

برای اینکه نتایج تولید شده توسط چت جی پی تی معتبرتر باشد، محققان از مدل های هوش مصنوعی می خواهند که شرکت اُپن ای آی در مورد داده های مورد استفاده در آموزش مدل هایش شفاف تر باشد. به گفته آنها دانستن اینکه یک مدل از روی چه کتاب هایی آموزش دیده است، برای ارزیابی چنین منابع سوگیرانه ای بسیار مهم است.

یکی از محققان دانشگاه برکلی در توییتی به این سوگیری بالقوه پرداخت و نوشت که با توجه به ترجیح آنها نسبت به منابع علمی تخیلی و فانتزی، برنامه های محبوب هوش مصنوعی احتمالا معیارهای اندازه گیری کاملی برای عملکرد مدل ها نیستند و ما باید به این فکر کنیم که تجربیات روایت چه کسی در این مدل ها کدگذاری شده است و چگونه بر سایر رفتارها تاثیر می گذارد.

استفاده از مطالب دارای حق نشر در آموزش مدل ها نیز انبوهی از سوالات حقوقی را ایجاد می کند. چه کسی دارای حق نشر متنی را که توسط چت جی پی تی ایجاد شده دارد. چت جی پی تی که خود پیش از این از روی کتاب های دارای حق نشر آموزش داده شده است؟ آیا مفهوم «استفاده منصفانه» در چنین موردی می تواند کاربرد داشته باشد؟

اگر تعدادی از افراد سعی کنند برای خروجی های مشابه یا یکسان توسط مدلی مانند چت جی پی تی حق نشر بگیرند، چه خواهد شد؟ از آنجایی که ماشین، انسان نیست، آیا هر چیزی که توسط آن تولید می شود در وهله اول دارای حق نشر است؟

سوگیری که توسط مطالعات دیگر آشکار شد

علاوه بر تحقیقات برکلی، مطالعات دیگری نیز مواردی از سوگیری را در پاسخ های چت جی پی تی پیدا کرده اند.

مطالعه ای که توسط گروهی در دانشگاه آنگلیا شرقی در بریتانیا انجام شد، سوگیری سیاسی را در برخی از پاسخ های این مدل نشان داد. هنگامی که صدها سوال در مورد اعتقادات سیاسی از این هوش مصنوعی پرسیده شد، به نظر می رسید چت جی پی تی بیشتر به دموکرات ها در ایالات متحده، حزب کارگر در بریتانیا و رئیس جمهور لولا داسیلوا(Lula da Silva) از حزب کارگران در برزیل متمایل باشد.

نویسنده اصلی آن مطالعه، دکتر فابیو موتوکی(Fabio Motoki)، از دانشکده تجارت نورویچ در دانشگاه آنگلیا شرقی، می گوید: وجود سوگیری سیاسی می تواند بر دیدگاه های کاربران نیز تاثیر بگذارد و پیامدهای بالقوه ای برای فرآیندهای سیاسی و

انتخاباتی داشته باشد.
یافته های ما این نگرانی را تقویت می کند که سیستم های هوش مصنوعی می توانند چالش های موجود ناشی از اینترنت و رسانه های اجتماعی را تکرار یا حتی تقویت کنند.
در مطالعه دیگری، محققان دانشگاه واشنگتن، دانشگاه کارنگی ملون و دانشگاه شیان جیائوتنگ، ۱۴ مدل هوش مصنوعی را تحت یک آزمون سوگیری سیاسی قرار دادند و به پاسخ های هر مدل به انواع مختلف ۶۲ بیانیه سیاسی پرداختند. چیزی که آن ها دریافتند این بود که پاسخ های تولید شده توسط چت جی پی تی و نسخه جدیدتر آن موسوم به جی پی تی-۴(GPT-۴) چپ گرایانه و آزادیخواهانه بودند.
شرکت اُپن ای آی رویکرد خود را در یک پست شرکتی با عنوان چگونه سیستم ها باید رفتار کنند به اشتراک گذاشته است.
براساس این پست، در حالی که بسیاری نگران سوگیری ها در طراحی سیستم های هوش مصنوعی هستند، ما متعهد هستیم که به طور جدی به این موضوع رسیدگی کنیم و در مورد اهداف و پیشرفت خود شفاف باشیم.
این شرکت همچنین خاطرنشان می کند که دستورالعمل های ما صریح است که بازبینان نباید از هیچ گروه سیاسی حمایت کنند. آن ها اظهار داشتند که ممکن است سوگیری ها همچنان در این فرآیند ظاهر شوند، اما ادعا کردند که این ها اشکالات به حساب می آیند و جزو ویژگی های هوش مصنوعی آن ها نیستند.