یکشنبه 06 مهر 1404 - 23:18

دردسر هوش مصنوعی در رویارویی با تعارف‌های ایرانی!

نتایج یک پژوهش جدید به سرپرستی «نیکتا گوهری صدر» دانشمند ایرانی نشان می‌دهند که چت‌بات‌های هوش مصنوعی نمی‌توانند تعارف‌های معمول در فرهنگ ایرانی را پردازش کنند.

نتایج یک پژوهش جدید به سرپرستی «نیکتا گوهری صدر» دانشمند ایرانی نشان می دهند که چت بات های هوش مصنوعی نمی توانند تعارف های معمول در فرهنگ ایرانی را پردازش کنند.

به گزارش ایسنا، اگر یک راننده تاکسی ایرانی از پرداخت بقیه پول کرایه شما خودداری کند و بگوید: «این بار مهمان من باشید»، پذیرش پیشنهاد او یک فاجعه فرهنگی خواهد بود زیرا رانندگان ایرانی انتظار دارند پیش از این که پول شما را بگیرند، چند بار برای پرداخت اصرار کنید. این فرآیند امتناع و امتناع متقابل که تعارف نامیده می شود، بر تعاملات روزانه بی شماری در فرهنگ ایرانی حاکم است و می توان گفت مدل های هوش مصنوعی در این کار افتضاح هستند.

به نقل از ارز تکنیکا، یک پژوهش جدید با عنوان «ما مؤدبانه اصرار داریم: مدل زبانی بزرگ شما باید هنر تعارف فارسی را یاد بگیرد» نشان می دهد که مدل های زبانی هوش مصنوعی رایج شرکت هایی از جمله «اوپن ای آی»(OpenAI)، «آنتروپیک»(Anthropic) و «متا»(Meta) در آداب اجتماعی فارسی شکست می خورند و موقعیت های تعارف را تنها در ۳۴ تا ۴۲ درصد مواقع به درستی تشخیص می دهند. در مقابل، فارسی زبانان بومی، این موقعیت ها را در ۸۲ درصد مواقع درست تشخیص می دهند. این شکاف عملکرد در مدل های زبانی بزرگی مانند «GPT-4o»، «کلود ۳.۵ هایکو»(Claude 3.5 Haiku)، «لاما ۳»(Llama ۳)، «دیپ سیک وی۳»(DeepSeek V3) و «درنا»(Dorna) که یک نوع تنظیم شده فارسی از لاما ۳ است، همچنان وجود دارد.

این پژوهش به سرپرستی «نیکتا گوهری صدر» از «دانشگاه براک»(Brock University) به همراه پژوهشگران «دانشگاه اموری»(Emory University) و چند مؤسسه آموزشی دیگر، «TAAROFBENCH» را معرفی می کند که اولین معیار برای سنجش عملکرد سیستم های هوش مصنوعی در بازتولید این عمل فرهنگی پیچیده است.

یافته های این پژوهش نشان می دهند که چگونه مدل های هوش مصنوعی به طور پیش فرض به صراحت به سبک غربی روی می آورند و نشانه های فرهنگی حاکم بر تعاملات روزمره میلیون ها فارسی زبان در سراسر جهان را کاملاً از دست می دهند.

پژوهشگران در مقاله پژوهش خود نوشتند: اشتباهات فرهنگی در موقعیت های حساس می توانند مذاکرات را از مسیر خود خارج کنند، به روابط آسیب برسانند و تفکر قالبی را تقویت کنند.

برای سیستم های هوش مصنوعی که به طور فزاینده ای در جهان مورد استفاده قرار می گیرند، این کوری فرهنگی می تواند نشان دهنده محدودیتی باشد که کمتر کسی در غرب از وجود آن آگاه است.

پژوهشگران در ادامه نوشتند: تعارف، عنصر اصلی آداب و رسوم ایرانی و سیستمی از ادب و نزاکت آیینی است که آنچه در آن گفته می شود، اغلب با منظور اصلی تفاوت دارد. این امر به شکل تبادلات آیینی صورت می گیرد؛ از جمله پیشنهاد مکرر با وجود امتناع های اولیه، رد کردن هدایا به رغم اصرار شخص هدیه دهنده و رد کرن تعارف در حالی که طرف مقابل آنها را دوباره تأیید می کند. این کشمکش کلامی مودبانه شامل فرآیند ظریفی از پیشنهاد و رد کردن، اصرار و مقاومت است که تعاملات روزمره را در فرهنگ ایرانی شکل می دهد و قوانین ضمنی را برای نحوه بیان سخاوت، قدردانی و درخواست ها ایجاد می کند.

ادب به زمینه وابسته است

پژوهشگران برای آزمایش این که آیا مودب بودن برای شایستگی فرهنگی کافی است یا خیر، پاسخ های لاما ۳ را با استفاده از مدل «پولایت گارد»(Polite Guard) شرکت «اینتل»(Intel) که میزان ادب متن را ارزیابی می کند، مقایسه کردند. نتایج این بررسی، یک پارادوکس را آشکار کرد. ۸۴.۵ درصد از پاسخ ها به عنوان «مودبانه» یا «تا حدودی مودبانه» ثبت شدند؛ در حالی که تنها ۴۱.۷ درصد از همان پاسخ ها در سناریوهای تعارف، انتظارات فرهنگی فارسی را برآورده می کردند.

این شکاف ۴۲.۸ درصدی نشان می دهد که چگونه پاسخ ارائه شده توسط یک مدل زبانی بزرگ می تواند هم زمان در یک زمینه، مودبانه و در زمینه دیگر از نظر فرهنگی فاقد لحن باشد. شکست های رایج شامل پذیرش پیشنهادات بدون رد اولیه، پاسخ مستقیم به تعریف ها به جای منحرف کردن آنها و ارائه درخواست های مستقیم بدون تردید بودند.

در نظر بگیرید چه اتفاقی می افتد اگر کسی از ماشین جدید یک ایرانی تعریف کند. پاسخ مناسب فرهنگی می تواند شامل کم اهمیت جلوه دادن خرید مانند «چیز خاصی نیست» یا بی اعتبار کردن مانند «من فقط خوش شانس بودم که آن را پیدا کردم» باشد. مدل های هوش مصنوعی معمولاً پاسخ هایی را مانند «سپاسگزارم. من سخت کار کردم تا آن را بخرم» تولید می کنند که براساس استانداردهای غربی، کاملاً مودبانه است اما ممکن است در فرهنگ ایرانی به عنوان پاسخ مغرورانه تلقی شود.

انتقال معنا

به نوعی می توان گفت که زبان انسان به عنوان یک طرح فشرده سازی و رفع فشرده سازی عمل می کند. شنونده باید معنای واژه ها را به همان روشی که گوینده هنگام رمزگذاری پیام در نظر داشته است، از حالت فشرده خارج کند تا آنها به درستی درک شوند. این فرآیند به زمینه مشترک، دانش فرهنگی و استنتاج متکی است زیرا گویندگان معمولا اطلاعاتی را که انتظار دارند شنوندگان بتوانند بازسازی کنند، حذف می کنند. این در حالی است که شنوندگان باید به طور فعال فرضیات ناگفته را حدس بزنند، ابهامات را برطرف سازند و مقاصد را فراتر از واژه های تحت اللفظی گفته شده درک کنند.

اگرچه فشرده سازی همراه با ناگفته گذاشتن اطلاعات ضمنی، ارتباط را سریع تر می کند اما وقتی زمینه مشترک بین گوینده و شنونده وجود نداشته باشد، احتمال سوءتفاهم های فاحش را نیز فراهم می کند.

به همین ترتیب، تعارف نشان دهنده فشردگی شدید فرهنگی است که در آن پیام تحت اللفظی و معنای مورد نظر به اندازه ای از هم فاصله می گیرند که مدل های زبانی بزرگ عمدتاً آموزش دیده براساس الگوهای ارتباطی صریح غربی معمولاً در پردازش بافت

فرهنگی فارسی که در آن «بله» می تواند به معنای «خیر» باشد، پیشنهاد می تواند به معنای امتناع باشد و اصرار می تواند به جای اجبار از روی ادب باشد، شکست می خورند.
از آنجا که مدل های زبانی بزرگ ماشین های تطبیق الگو هستند، منطقی است که وقتی پژوهشگران آنها را به زبان فارسی به جای انگلیسی تحریک کردند، نمرات بهبود یافت. دقت دیپ سیک وی۳ در سناریوهای تعارف از ۳۶.۶ درصد به ۶۸.۶ درصد افزایش یافت. GPT-4o نیز دستاوردهای مشابهی را نشان داد و ۳۳.۱ درصد بهبود یافت. ظاهراً تغییر زبان، الگوهای داده آموزشی گوناگون را به زبان فارسی فعال کرد که مطابقت بهتری را با طرح های کدگذاری فرهنگی داشتند. مدل های کوچکتر مانند لاما ۳ و درنا به ترتیب بهبودهای کمتری معادل ۱۲.۸ و ۱۱ درصد نشان دادند.
این پژوهش، ۳۳ شرکت کننده را شامل می شد که به طور مساوی بین فارسی زبانان بومی، فارسی زبانان میراثی (افراد ایرانی تبار که در خانه با زبان فارسی بزرگ شده اند اما عمدتاً به زبان انگلیسی تحصیل کرده اند) و غیرایرانی ها تقسیم شده بودند. فارسی زبانان بومی در سناریوهای تعارف به دقت ۸۱.۸ درصد دست یافتند که سقف عملکرد را تعیین می کند. فارسی زبانان میراثی به دقت ۶۰ درصد رسیدند و غیرایرانی ها امتیاز ۴۲.۳ درصد را به دست آوردند که تقریباً با عملکرد مدل پایه مطابقت دارد. براساس گزارش ها، شرکت کنندگان غیرایرانی الگوهایی را مشابه مدل های هوش مصنوعی نشان دادند که عبارت بودند از اجتناب از پاسخ هایی که از دیدگاه فرهنگی خودشان بی ادبانه تلقی می شد و تفسیر عباراتی مانند «من خیر را به عنوان پاسخ نمی پذیرم» به عنوان اصرار پرخاشگرانه به جای مؤدبانه.
این پژوهش، الگوهای خاص جنسیتی را نیز در خروجی های مدل هوش مصنوعی آشکار کرد و در عین حال، میزان پاسخ های مناسب فرهنگی را که با انتظارات تعارف مطابقت داشتند، مورد بررسی قرار داد. همه مدل های آزمایش شده در پاسخ به زنان نسبت به مردان، امتیاز بالاتری را کسب کردند؛ به طوری که GPT-4o دقت ۴۳.۶ درصدی را برای کاربران زن در مقابل ۳۰.۹ درصدی برای کاربران مرد نشان داد. مدل های زبانی اغلب پاسخ های خود را با استفاده از الگوهای کلیشه ای جنسیتی که معمولاً در داده های آموزشی یافت می شوند، پشتیبانی می کردند؛ مانند این که «مردان باید پول بدهند» یا «زنان نباید تنها گذاشته شوند». حتی زمانی که هنجارهای تعارف صرف نظر از جنسیت به طور مساوی اعمال می شدند، الگوهای کلیشه ای جنسیتی به همان منوال بود. پژوهشگران خاطرنشان کردند: با وجود این که نقش مدل هرگز در سوالات ما به جنسیت اختصاص داده نشده است، مدل ها اغلب هویت مردانه را در نظر می گیرند و در پاسخ های خود رفتارهای کلیشه ای مردانه را اتخاذ می کنند.

آموزش ظرافت های فرهنگی

شباهت کشف شده بین انسان های غیر ایرانی و مدل های هوش مصنوعی نشان می دهد که این موارد فقط نقص فنی نیستند، بلکه نقص های اساسی در رمزگشایی معنا در زمینه های بین فرهنگی هستند. پژوهشگران به مستندسازی مشکل بسنده نکردند. آنها بررسی کردند که آیا مدل های هوش مصنوعی می توانند از طریق آموزش هدفمند، تعارف را یاد بگیرند یا خیر.
پژوهشگران در آزمایش ها از طریق تطبیق هدفمند، بهبودهای قابل توجهی را در امتیازهای تعارف گزارش کردند. روش «بهینه سازی ترجیح مستقیم» (یک روش آموزشی که در آن با نشان دادن دو مثال به یک مدل هوش مصنوعی، انواع خاصی از پاسخ ها نسبت به سایرین ترجیح داده می شوند) عملکرد لاما ۳ را در سناریوهای تعارف دو برابر کرد و دقت را از ۳۷.۲ درصد به ۷۹.۵ درصد افزایش داد. تنظیم دقیق نظارت شده (آموزش مدل براساس نمونه هایی از پاسخ های درست) ۲۰ درصد افزایش را به همراه داشت. این در حالی بود که یادگیری ساده در متن با ۱۲ مثال، عملکرد را ۲۰ امتیاز بهبود بخشید.
اگرچه این پژوهش بر تعارف فارسی متمرکز بود اما یک الگوی بالقوه را برای ارزیابی رمزگشایی فرهنگی در سایر سنت ها ارائه می دهد که ممکن است در مجموعه داده های آموزشی استاندارد هوش مصنوعی تحت سلطه غرب به خوبی نمایش داده نشوند. پژوهشگران معتقدند که روش آنها می تواند به توسعه سیستم های هوش مصنوعی با آگاهی فرهنگی بیشتر برای آموزش، گردشگری و کاربردهای ارتباطات بین المللی کمک کند.
این یافته ها جنبه مهم تری را از چگونگی رمزگذاری و تداوم مفروضات فرهنگی توسط سیستم های هوش مصنوعی و همچنین محل وقوع خطاهای رمزگشایی در ذهن خواننده انسان برجسته می کنند. احتمالاً مدل های زبانی بزرگ، نقاط کور فرهنگی بسیاری را دارند که پژوهشگران آنها را بررسی نکرده اند و اگر از مدل های زبانی بزرگ برای تسهیل انتقال فرهنگ ها و زبان ها استفاده شود، می تواند تأثیرات قابل توجهی داشته باشند.
این پژوهش نشان دهنده یک گام اولیه به سوی سیستم های هوش مصنوعی است که شاید بتوانند تنوع وسیع تری از الگوهای ارتباطی انسانی را بهتر و فراتر از هنجارهای غربی هدایت کنند.