

در مرز تازه نوشتن ایستاده‌ایم



قادر باستانی‌تبریزی

پژوهشگر علوم ارتباطات

ایران امروز در میانه موجی از تحولات فناورانه قرار گرفته که فراتر از مرزهای جغرافیا، شیوه نوشتن، خواندن و حتی اعتمادکردن ما را زیرورو کرده است. آنچه زمانی نشانه اصالت و خلاقیت انسانی بود، اکنون در برابر ماشین‌هایی قرار گرفته که می‌توانند نویسنند، تحلیل کنند و حتی احساسات را تقلید کنند. در چنین فضایی، ما به‌عنوان جامعه‌ای با رسانه‌های مسئله‌مند و اعتماد فرسوده، چه باید بکنیم؟۱۲وقتی ناوومی آل‌درمن، نویسنده مشهور آمریکایی، چندی پیش نوشت که دیگر هیچ جمله‌ای را نمی‌توان با اطمینان کامل به انسان نسبت داد، بسیاری این جمله را اغراق آمیز دانستند، اما واقعیت این است که جهان، و ازجمله ایران، درست در آستانه همین نقطه قرار گرفته است. مدل‌های زبانی امروز می‌توانند چنان روان، دقیق و کم‌خطا بنویسند که تشخیص منشأ متن به مهارتی تازه تبدیل شده است؛ مهارتی که به‌ویژه در جامعه‌ای با گردش خبر پُرشتاب و بسترهای ارتباطی غیرشفاف، ارزشی حیاتی یافته است. آل‌درمن در نوشته خود هفت نشانه برمی‌شمارد که می‌تواند میان متن انسانی و متن ماشینی فاصله‌ای ظریف، اما قابل تشخیص ایجاد کند. نخستین آن، کمبود انسجام در مقیاس بزرگ است. جملات هوش مصنوعی معمولاً درست و روان‌اند، اما وقتی قرار است روایتی طولانی یا تحلیلی پیچیده شکل بگیرد، رشته‌منطق ازهمی گسلد. ذهن انسان، ایده‌ها را می‌پرورد، ما ماشین آنها را کنار هم می‌چیند. نشانه دوم، تکرار الگوهای زبانی است. مدل‌ها از برخی عبارات‌ها مانند «نه‌تنها... بلکه» و نیز جملات مستقل متصل با علامت نقطه ویرگول (؛) بی‌اندازه استفاده می‌کنند. گویی اینبار واژگان‌شان در نقاط خاصی گیرکرده است. سومین نشانه، افت معنا در مسیر طولانی متن است؛ جایی که نوشته ماشینی ناگهان بی دلیل آرام می‌شود و به کلی گویی بازمی‌گردد. نشانه‌های دیگر نیز کم‌پوشش همین الگو و دنبال می‌کنند؛ گریز از تناقض‌ها، نبود تجربه زیسته، گرایش به کلیشه‌ها و اختلال در منطق و دقت. مجموعه اینها تصویری می‌سازد از نویسنده‌ای که همه‌چیز را می‌داند، اما هیچ‌چیز را لمس نکرده است. اما برای ایران، موضوع فقط یک بحث ادبی یا تکنیکی نیست. ما در جامعه‌ای زندگی می‌کنیم که اعتماد عمومی به رسانه‌ها، در سال‌های اخیر بهشدت آسیب دیده است. حال تصور کنید به این بستر، موجی از متن‌ها، تصاویر و ویدئوهای مصنوعی را اضافه کنیم؛ محتواهایی که «درست‌به‌نظر می‌رسند»، اما معلوم نیست چه دستی و با چه نیتی آنها را تولید کرده است. این مواد خام بی‌اصل و نسب امروز، در شبکه‌های اجتماعی فارسی‌زبان جولان می‌دهند؛ خبرهای ساختگی با امضای خبرنگاران خیالی، تحلیل‌های سیاسی تولیدشده‌با هوش مصنوعی، روایت‌های احساسی دربراره رویدادهای جعلی و حتی شعرهایی که با نام شاعران مشهور منتشر می‌شود اما یک‌جای کارشان می‌لنگد. ما خواسته یا ناخواسته، به بخشی از «اقتصاد محتوا»یی تبدیل شده‌ایم که تولیدکننده‌اش معلوم نیست، اما تأثیرش واقعی است. اینجاست که اهمیت سخن آل‌درمن دوچندان می‌شود که سواد رسانه‌ای در عصر هوش مصنوعی، صرفاً توانایی «خواندن» نیست، توانایی «تشخیص منشأ» است. یعنی اینکه پیش از یازنشر یک متن، از خود بپرسیم که آیا این صدا، صدای انسان است؟ آیا کسی پشت این نوشته ایستاده که مسئولیت معنایش را بپذیرد؟ هوش مصنوعی در ذات خود تهدید نیست، تهدید زمانی شکل می‌گیرد که جامعه، فاقد مهارت تشخیص باشد. وقتی کاربران ندانند که چگونه نشانه‌های متن ساختگی را بشناسند، محتواهای تقلبی بی‌صدا تکثیر می‌شوند و حقیقت در ازدحام واژه‌ها گم می‌شود. این مسئله در ایران، که همواره با چالش‌های ارتباطی مواجه بوده، بحرانی مهم محسوب می‌شود. اما ازسوی دیگر، همین فناوری فرصتی نیز فراهم کرده است. هوش مصنوعی می‌تواند ابزار توانمندسازی برای خبرنگاران، پژوهشگران، دانشجویان، معلمان و حتی کسب‌وکارهای کوچک باشد. اگر سواد رسانه‌ای ارتقا یابد و چارچوب‌های اخلاقی شفاف شکل بگیرد، این فناوری می‌تواند به‌گسترش مشارکت، افزایش دسترسی به دانش و بهبود کیفیت اطلاع‌رسانی کمک کند. باید بپذیریم که دوران «خودبستگی انسانی» در تولید متن پایان یافته است. اکنون هم نویسنده هستیم، هم سربیر و باید نه‌فقط به محتوا، بلکه به منشأ نگاه کنیم. مرز میان تخیل و تقلب هرروز باریک‌تر می‌شود، اما حساسیت انسانی هنوز ارزشمندترین ابزار ماست. در جهانی که واژه‌ها نیز مصنوعی شده‌اند، صداقت انسانی در گفتار و شفافیت در منشأ پیام، هم‌یک فضیلت اخلاقی، هم ضرورتی برای زیستن آگاهانه است. آینده ارتباطات حذف ماشین‌هانیست، بلکه به‌بلوغ‌مادر کنار آنها وابسته است؛ آینده‌ای که اگر هوشمندانه با آن روبه‌رو شویم، می‌تواند اعتماد ازدست‌رفته را از نو بسازد.

به این نکته زمانی پی بردم که با یک مدیر ارشد که کارش توسعه چت‌بات هوش مصنوعی است، گفت‌وگو کردم. سیستمی که به بیش از ۸ میلیون کشاورز در آسیا و آفریقا، مشاوره‌های کشاورزی ارائه می‌دهد. مشاوره‌هایی که عمدتاً متکی به پایگاه‌های داده مشاوره‌های دولتی و سازمان‌های بین‌المللی توسعه‌ای است که ماهیت علمی و پژوهشی دارند. مدیر یادشده اذعان کرد که بسیاری از شیوه‌های محلی اثربخش چون مستند نشده‌اند، در پاسخ‌های چت‌بات لحاظ نمی‌شوند، دلیل‌اش هم این نیست که توصیه‌های پژوهش‌محور همیشه درست یا بدون ریسک هستند، بلکه این است که تکیه بر منابع شناخته‌شده به‌عنوان گزینه‌ای ایمن‌تر دیده می‌شود و سازمان را از مسئولیت احتمالی محافظت می‌کند. به‌عبارتی دیگر، دانش محلی که از کانال‌های نهادی تأییدنشده، کنار گذاشته می‌شود. بنابراین این تصمیم صرفاً فنی نیست؛ بلکه یک مصالحه است که تحت‌تأثیر ساختارهای نهادی اتخاذ می‌شود، نه صرفاً براساس آنچه مفید یا درست است. یافتن راه‌هایی برای تأیید دانش بومی در سیستم‌هایی که تاریخی طولانی از نادیده‌نگاری دارند، دشوار است. هرچند هوش مصنوعی مولد ممکن است به روند حذف دانش محلی شتاب می‌بخشد، ریشه این وضع جای دیگری است. واقعیت این است ساختارهای قدرت تثبیت‌شده مدت‌هاست روند نادیده‌نگاری و به‌حاشیابردن دانش محلی و بومی را پیش می‌برند.

▼ ترازدی جهانی نابودی دانش بومی و محلی

هوش مصنوعی مولد نیز این روند را به‌طرز چشمگیری تشدید می‌کند. اغلب ما از بین‌فوتن دانش بومی را به‌عنوان یک ترازدی برای جوامع محلی می‌شناسیم ولی حقیقت این است که مخاطب و قربانی آن تمام جهانیان هستند. ناپدیدشدن دانش محلی، فقدانِ کم‌اهمیت نیست؛ این یک اختلال در شبکه گسترده‌ای از فهم است که رفاه انسانی و اکولوژیکی را پشتیبانی می‌کند. درست همان‌طور که گونه‌های زیستی برای بقا در محیط‌های محلی خاص تکامل یافته‌اند، نظام‌های دانش انسانی نیز با ویژگی‌های خاص هر مکان سازگار می‌شوند و در صورت بروز اختلال در آن پیامدهایی فراتر از نقطه مبدأ دارد. دود آتش‌سوزی‌ها از مرزها می‌گذرند، آب آلوده در خطوط ایالتی متوقف نمی‌شود، تغییرات اقلیمی و افزایش دما به مرزهای ملی توجهی ندارند، میکروب‌های عفونی در انتظار ویرا نمی‌نشینند و... چه بخواهیم و چه نخواهیم، ما در نظام بوم‌شناختی مشترکی گرفتار شده‌ایم که زخم‌های محلی ناگزیر به دردهای جهانی تبدیل می‌شوند. بزرگ‌ترین تضاد برای من در نوشتن این مقاله این است که در حال متقاعدسازی خوانندگان به مشروعیت و اهمیت نظام‌های دانش بومی و محلی هستم، درحالی‌که خودم هنوز درباره معجز گباهی پدرم مطمئن نیستم. من درباره اینکه آیا دانش بومی در هر مورد واقعاً همان‌طور که ادعا می‌شود، عمل می‌کند یا خیر، شک دارم؛ به‌ویژه وقتی اینفلونسر ها وسایستمداران از این دانش به‌سطحی‌ترین شکل ممکن برای جلب‌توجه یا اهداف خاص بهره‌برداری می‌کنند. به همان اندازه هم نگرانم که اجازه دهیم این دانش ناپدید شود. ممکن است چیزی ارزشمند را از دست بدهیم و بعدها به ارزش واقعی آن پی ببریم. خسارت جانی این روند چیست؟ فروپاشی اکولوژیکی که می‌توانستیم از آن جلوگیری کنیم؟

بحران اقلیمی شکاف‌های موجود در پارادایم‌های مسلط دانش را آشکار می‌کند. بااین حال توسعه‌دهندگان هوش مصنوعی قاطعانه معتقدند که فناوری آن‌ها پیشرفت علمی را تسریع می‌کند و بزرگ‌ترین چالش‌های ما را حل خواهد کرد. واقعاً می‌خواهم باور کنم که آن‌ها درست می‌گویند. اما چند پرسش هم باقی می‌ماند آیا می‌توانیم به این آینده فناورانه حرکت کنیم و درعین حال با سیستم‌های دانشی که پیش‌تر رد کرده‌ایم، به‌طور واقعی و با کنجکاوی فراتر از نمادگرایی تعامل داشته باشیم؟ یا اینکه همچنان با سلسله‌مراتب ساخته‌شده، فهم‌های مختلف را حذف خواهیم کرد و درنهایت خود را در تلاش برای استعمار مرخ می‌یابیم، چون هرگز یاد نگرفته‌ایم به کسانی گوش دهیم که می‌دانستند چگونه به‌طور پایدار روی زمین زندگی کنند؟ شاید هوشمندی‌ای که بیش از همه نیاز داریم، توانایی دیدن فراتر از سلسله‌مراتب‌هایی است که تعیین می‌کنند کدام دانش ارزش دارد. بدون این پایه، صرف‌نظر از صدها میلیارد دلار که برای توسعه ابرهوش سر‌مایه‌گذاری می‌کنیم، همچنان سیستم‌های دانشی را حذف خواهیم کرد که شکل‌گیری‌شان نسل‌ها طول کشیده است.

من نمی‌دانم معجون‌های گیاهی پدرم واقعاً مؤثر بوده‌اند یا خیر، اما در حال آموختن این واقعیت هستم که به‌رسمیت شناختن چیزی که نمی‌دانم، شاید صادقانه‌ترین نقطه برای شروع باشد.

مترجم: **سمابابی**
 منبع: **گاردین**

بپرسید، پاسخ‌دیپلماتیکی دریافت خواهید کرد که انگار توسط هیئتی از وکلاد و کارشناسان منابع انسانی تهیه‌شده است. پاسخی که به‌شکل غیر معمولی این لحن را دارد که رضایت‌تان را جلب کند. اگر همان سؤال را از گروک، چت‌بات هوش مصنوعی شرکت اکس، بپرسید ممکن است با یک جمله طعنه‌آمیز پاسخ‌تان را بدهد، سپس برداشت سیاسی ارائه کند که مناسب ضیافت میلیاردرهای فناوری باشد. فشارهای تجاری نیز لایه دیگری به این فرآیند اضافه می‌کنند. پردرآمدترین کاربران، حرفه‌ای‌های انگلیسی‌زبان که حاضرند ماهانه ۲۰ تا ۲۰۰ دلار برای اشتراک‌های پرمیموم AI بپردازند، به الگوی ضمنی «ابرهوش» تبدیل می‌شوند. این مدل‌ها در تولید گزارش‌های سه‌ماهه، کندیوسی به زبان‌های مورد علاقه جامعه بزرگ سیلیکون‌ولی و نوشتن ایمیل‌هایی که به هرمد‌های شرکتی غربی احترام می‌گذارند، عملکردشان عالی است. درعین حال آنها به زمینه‌های فرهنگی که به کسب‌درآمد فصلی بیشتر مرتبط نباشد اهمیت نمی‌دهند.

▼ فروپاشی دانش نزدیک است

با این حساب جای شگفتی ندارد که اغلب مطالعات نشان می‌دهند که مدل‌های زبانی بزرگ ارزش‌ها و رویکردهای معرفت‌شناختی غربی را با‌زتآب می‌هند. این مدل‌ها، اکثریت را در خروجی‌های خود بیش از حد نمایان می‌کنند، سوگیری‌هایشان را تقویت و برجسته می‌کنند و دقت‌شان در موضوعات مربوط به آمریکای شمالی و اروپا بیشتر است. حتی در مواردی مانند توصیه‌های سفر یا داستان‌سرایی، مدل‌های زبانی بزرگ تمایل دارند محتوای غنی‌تر و جزئیات بیشتری برای کشورهای ثروتمند تولید کنند و در مقایسه با کشورهای فقیر، محتوا کمتر و سطحی‌تر است. علاوه بر این، هوش مصنوعی مولد قدرت و ظرفیت زیادی برای تقویت بازتاب سلسله‌مراتب دانشی شکل گرفته دارد و برهمین اساس نیز بر رفتار انسان‌ها تغییر می‌گذارند. ادغام خلاصه‌های AI در موتورهای جست‌وجو، همراه با رشد محبوبیت موتورهای جست‌وجوی مبتنی بر AI مانند پرپلکسیتی، این تغییر را نشان می‌دهد. با بردن اینترنت از محتوای تولیدی هوش مصنوعی، لایه‌ای از ایده‌های پرطرفدار آنلاین ایجاد می‌شود و چرخه‌ای شکل می‌گیرد که فضای اینترنت تحت‌تأثیر خروجی‌های همین مدل‌ها قرار می‌گیرد. یعنی درحالی‌که اینترنت خود منبع اصلی دانش برای مدل‌های هوش مصنوعی است، در هر چرخه آموزش، مدل‌هایشتر به محتوای تولیدشده توسط هوش مصنوعی تکیه می‌کنند. این وضعیت خطر ایجاد چرخه‌بازخوردی را دارد که در آن ایده‌های غالب دائماً تقویت می‌شوند و دانش کمتر رایج یا تخصصی به حاشیه می‌رود. پژوهشگر هوش مصنوعی، اندرو پیترسون، این پدیده را «فروپاشی دانش» توصیف می‌کند؛ محدودشدن تدریجی اطلاعات در دسترس و کاهش سطح آگاهی از دیدگاه‌های جایگزین یا کمتر شناخته‌شده. هنگامی‌که مدل‌ها براساس داده‌هایی آموزش می‌بینند که براساس خروجی‌های قبلی هوش مصنوعی شکل گرفته‌اند، دانش کم‌بازنمایی‌شده نیز کمتر مشاهده می‌شود. دلیل این وضع نه کم ارزش بودن دانش، بلکه به این دلیل که کمتر به آنها ارجاع داده شده است. پیترسون همچنین درباره وضعیت «اثر چراغ خیابانی» هشدار می‌دهد. مشابه وضعیتی که در آن شخصی به‌دنبال کلید گمشده‌اش زیر چراغ خیابان را جست‌وجو می‌کند چون نور آنجا بیشتر است. در زمینه هوش مصنوعی، این یعنی انسان‌ها جست‌وجو را جایی انجام می‌دهند که آسان‌تر است، نه جایی که معنا دارتر باشد. با گذر زمان، شکی نیست که چنین روندی به محدودشدن تدریجی و کاهش کیفیت پایه‌دانش عمومی می‌انجامد. در سراسر جهان، هوش مصنوعی مولد به بخشی از نظام آموزش رسمی تبدیل شده‌اند. کشورها برای تولید محتوای یادگیری و پشتیبانی از آموزش خودآموز از هوش مصنوعی استفاده می‌کنند. تمام این‌ها به این معناست که در دنیایی که هوش مصنوعی به‌طور فزاینده‌ای میانجی دسترسی به دانش شده، نسل‌های آینده ممکن است ارتباط خود با گستره وسیعی از تجربه، بینش و خرد انسانی را از دست بدهند.

▼ مصالحه با ساختارهای نهادی مرسوم

توسعه‌دهندگان هوش مصنوعی ممکن است استدلال کنند که چنین موضوعاتی صرفاً یک مسئله داده‌ای است و با واردکردن منابع متنوع‌تر به مجموعه داده‌های آموزشی قابل حل است. از نظر فنی ممکن است بتوان چنین کرد، اما چالش‌های مربوط به تأمین داده، اولویت‌بندی و بازنمایی بسیار پیچیده‌تر از راه‌حل ساده است.

مناطق دارای تابش شدید آفتاب، نماهای شیشه‌ای باعث بالارفتن دمای داخلی و نارضایتی حرارتی می‌شوند. حتی اگر از شیشه‌های مدرن نیز در این ساختمان‌ها استفاده کنیم، فایده‌ای ندارد. چراکه به‌جای صرفه‌جویی در انرژی، نیاز به مصرف بیشتر انرژی برای خنک نگه‌داشتن فضا دارند. با این حال نماهای شیشه‌ای بدون توجه به اقلیم یا بافت فرهنگی به‌چهره مدرنیزته شهری تبدیل شده‌اند. فرقی هم ندارد چه در سافرانسیسکو باشند، چه در جاکارتا یا لاگوس.

▼ استعمار همچنان ادامه دارد



اندرو پیترسون، پژوهشگر هوش مصنوعی، این پدیده را «فروپاشی دانش» توصیف می‌کند؛ محدودشدن تدریجی اطلاعات در دسترس و کاهش سطح آگاهی از دیدگاه‌های جایگزین یا کمتر شناخته‌شده. هنگامی‌که مدل‌ها براساس داده‌هایی آموزش می‌بینند که براساس خروجی‌های قبلی هوش مصنوعی شکل گرفته‌اند، دانش کم‌بازنمایی‌شده نیز کمتر مشاهده می‌شود. دلیل این وضع نه کم ارزش بودن دانش، بلکه به این دلیل که کمتر به آنها ارجاع داده شده است

اگر چه تمام این مثال‌ها مربوط به هند است و دلیل آن نیز تجربه شخصی و آشنایی خودم است، به تاریخ جهانی امپریالیسم و استعمار که‌نگاه کنید، سلسله‌مراتب‌هایی گسترده‌تر وریشه‌دار را هم می‌بینید. لیندا توی‌وایی اسمیت، دانشمند مائوری در کتاب «استعمارزدایی از روش» تأکید می‌کند که استعمار با قطع ارتباط مردم با زمین، زبان، تاریخ و ساختارهای اجتماعی، نظام‌های دانش محلی و بنیان‌های فرهنگی و فکری آن‌ها را به‌شدت مختل کرده است، اسمیت بیان می‌کند که فرآیندهای استعماری محدود به یک منطقه نیستند، بلکه بخشی از میراث گسترده‌ای هستند که همچنان شکل‌دهنده تولید و ارزش‌گذاری دانش است. سیستم‌های دیجیتال و هوش مصنوعی مولد امروزین نیز براساس همین بنیاد تحریف‌شده ساخته شده‌اند. اخیراً بابخش پژوهشی شرکت مایکروسافت به‌نام Microsoft Research روی چند پروژه هوش مصنوعی مولد که برای جمعیت‌های غیرغربی طراحی‌شده، کار کردم. آنچه مشاهده کردم این بود که مدل‌های هوش مصنوعی اغلب زمینه‌های فرهنگی را نادیده می‌گیرند، دانش محلی را از قلم می‌اندازند و خیلی کم پیش می‌آید با نیازهای جامعه هدف، هم‌راستا باشند. تحقیقات ما به‌وضوح نشان داد که مدل‌های هوش مصنوعی هم‌زمان با رمزگذاری سوگیری‌های موجود، دانش به‌حاشیه رانده‌شده را حذف می‌کنند. این تجربه همچنین مرا به درک دلایل فنی شکل‌گیری چنین نابرابری‌هایی در مدل‌ها نزدیک‌تر کرد. مشکل فراتر از کمبود داده‌های آموزشی است. ماهیت مدل‌های زبانی بزرگ به‌گونه‌ای است که ایده‌های آماری غالب را بازتولید و تقویت و یک حلقه بازخورد ایجاد کنند که دامنه دانش بشری قابل دسترس را محدود می‌کند. می‌پرسید چرا چنین اتفاقی رخ می‌دهد؟ دلیل‌اش ساده است؛ بازنمایی داخلی دانش در یک مدل‌های زبانی بزرگ یکنواخت نیست. مفاهیمی که در داده‌های آموزشی بر جسته‌تر یا وسیع‌تر ی زمینه‌ها ظاهر می‌شوند با شدت بیشتری رمزگذاری می‌شوند. به‌عنوان مثال، اگر «پیتزا» به‌عنوان غذایی محبوب در مجموعه بزرگی از متون آموزشی بارها ذکر شده باشد، وقتی از مدل پرسیده می‌شود «غذای مورد علاقه‌ات چیست؟»، احتمال پاسخ «پیتزا» بالاتر است، زیرا این ارتباط از نظر آماری برجسته‌تر است. در مدل‌های زبانی بزرگ غالباً الگوها یا ایده‌های غالب به‌گونه‌ای تقویت می‌شوند که نسبت‌های اصلی آن‌ها را تحریف می‌کند. پدیده‌ای که به «تقویت الگو» شناخته می‌شود و بر خلاف دانش‌های دیگر است که توزیع خروجی مدل لزوماً بازتاب‌دهنده فراوانی ایده‌ها در داده‌های آموزشی نیست.

▼ کمتر همیشه حذف می‌شود

فرض کنید داده‌های آموزشی مربوط به علاقه به غذا شامل ۶۰ درصد اشاره به پیتزا، ۲۰ درصد اشاره به پاستا و ۱۰ درصد اشاره به بریانی باشد. ممکن است انتظار داشته باشیم که مدل در پاسخ به همان سؤال ۱۰۰ بار، دقیقاً همین توزیع را بازتولید کند. اما مدل‌های زبانی بزرگ، چنین نمی‌کنند. آنها تمایل دارند پاسخ رایج‌تر را بیش از حد تولید کنند. ممکن است پیتزا بیش از ۶۰ بار ظاهر شود، درحالی‌که مثلاً بریانی، کمتر یا حتی حذف شود. این اتفاق به این دلیل رخ می‌دهد که مدل‌ها زبانی بزرگ برای پیش‌بینی محتمل‌ترین توکن بعدی (در اینجا به‌معنی کلمه یا بخش کلمه در یک توالی) بهینه‌سازی شده‌اند و این باعث می‌شود پاسخ‌های با احتمال بالا بیش از حد برجسته شوند. این رمزگذاری نامتوازن با استفاده از یادگیری تقویتی مبتنی بر بازخورد انسانی (RLHF) بیشتر تحریف می‌شود؛ در این فرآیند، مدل‌های هوش مصنوعی مولد براساس ترجیحات انسانی تنظیم می‌شوند. نکته این است که ارزش‌ها و جهان‌بینی سازندگان به‌طور اجتناب‌ناپذیری در درون مدل‌ها نشانده می‌شود. اگر از چت‌جی‌پی‌تی درباره یک موضوع بحث‌برانگیز

