

گشودن راه‌های جدید: هوش مصنوعی مولد مسئولیت‌پذیر و اتخاذ خلاقانه یک مدل زبانی بزرگ در دلویت کانادا

صفحه ۱

گشودن راه‌های جدید: هوش مصنوعی مولد مسئولیت‌پذیر و اتخاذ خلاقانه یک مدل زبانی بزرگ در دلویت کانادا

مورد تهیه شده توسط ماریون کوروسک-سرفاتی 11 و لوک لسپرانس 22

در اواخر دسامبر ۲۰۲۲، سامیر بوهر، مدیر ملی فناوری اطلاعات (IT) در دلویت کانادا (دلویت)، در دفتر خود در تورنتو نشسته بود و به تفکر فرو رفته بود. منظره چراغ‌های درخشان و ترنینات کریسمس از پنجره‌اش چشم‌نواز بود و تعطیلات نزدیک بود، اما ذهن بوهر جای دیگری بود. او بیش از یک دهه بود که تیم فناوری اطلاعات دلویت را رهبری می‌کرد و بر بسیاری از ابتکارات تحول‌آفرین نظارت داشت، اما اکنون با چالشی دلهره‌آور روبرو بود. یک فناوری نوظهور در حال تغییر رفتار کاربران بود، بر نحوه انجام وظایف و تصمیم‌گیری‌ها تأثیر می‌گذاشت و نگرانی‌های مهمی درباره اخلاق کسب‌وکار، حریم خصوصی داده‌ها و امنیت ایجاد می‌کرد.

چند هفته قبل، یک ترانسفورماتور پیش‌آمورده مولد (GPT)، نوع پیشرفته‌ای از مدل‌های زبانی بزرگ (LLM) که در تولید متن شبیه به انسان و درک الگوهای پیچیده زبانی مهارت دارد، به صورت عمومی راه‌اندازی شده بود. پذیرش سریع و جهانی آن، نقطه عطفی برای مدل‌های زبانی بزرگ و ادغام گسترده‌تر راحل‌های مبتنی بر هوش مصنوعی در فرآیندهای سازمانی و شیوه‌های کاری محسوب می‌شد.

بوهر با تصمیمی بزرگ روبرو بود و باید به زودی تصمیم می‌گرفت. آیا باید به ۱۴۰۰۰ کارمند شرکت اجازه دهد از یک مدل زبانی بزرگ خارجی برای تعاملات داخلی و مشتری‌محور استفاده کنند؟ آیا باید این فناوری جدید را مسدود کند یا استفاده از آن را با دستورالعمل‌ها و بهترین شیوه‌ها محدود کند؟ یا باید فناوری مشابهی را به صورت داخلی مستقر کند تا کارمندان از قابلیت‌های جدید مدل زبانی بزرگ بهره‌مند شوند؟

سفر بوهر

سامیر بوهر در هند به دنیا آمد و در یک شهر کوچک در نیمه غربی این کشور به مدرسه رفت. علاقه او به کامپیوترها و فناوری زمانی آغاز شد که برای اولین بار در یک نمایشگاه علمی در اوایل دهه ۹۰ با یک کامپیوتر تعامل داشت. در حالی که بیشتر همسالانش به بازی‌های آرکید و کریکت علاقه داشتند، بوهر ساعت‌ها با یکی از سه کامپیوتر مدرسه‌اش وقت می‌گذراند، سیستم عامل آن را بررسی می‌کرد و برنامه‌نویسی را به خود یاد می‌داد. به زودی اتاق خواب او به یک آزمایشگاه کوچک فناوری پر از سیم‌ها، بردهای مدار و مجموعه‌ای از قطعات کامپیوتری تبدیل شد.

11 ماریون کوروسک-سرفاتی استاد فناوری اطلاعات در گروه تحلیل، عملیات و فناوری اطلاعات در مدرسه مدیریت UQAM و دانشجوی دکترای فناوری اطلاعات در HEC Montréal است.
22 لوک لسپرانس مدرس دانشکده در گروه فناوری‌های اطلاعات HEC Montréal است.

© HEC Montréal 2024

تمامی حقوق برای تمام کشورها محفوظ است. هرگونه ترجمه یا تغییر به هر شکل ممنوع است.
مجله بین‌المللی مطالعات موردی در مدیریت به صورت آنلاین منتشر می‌شود

این مورد به عنوان چارچوبی برای بحث آموزشی در نظر گرفته شده و قضاوتی در مورد موقعیت اداری ارائه شده ندارد. تحت شماره ۹ ۱۰۰ ۲۰۲۴ ۰۰۳ در مرکز مورد HEC Montréal، ۳۰۰۰، chemin de la Côte-Sainte-Catherine، Montréal (Québec) H3T 2A7 کانادا ثبت شده است.

این سند برای استفاده مرئوس تا مه ۲۰۲۴ توسط آنالیتا شاهنواز مجاز است. کپی یا ارسال نقض حق چاپ محسوب می‌شود Permissions@hbsp.harvard.edu یا ۶۱۷،۷۸۳،۷۸۶۰.

صفحه ۲

یک روز، در دوران دانشگاه، بوهر در یک مسابقه برنامه‌نویسی که از طریق معلم ریاضی‌اش شنیده بود شرکت کرد. با خوشحالی دریافت که می‌تواند بیشتر مشکلات را با به کارگیری مهارت‌های تفکر الگوریتمی خود حل کند، این امر شور و شوق او به فناوری را برانگیخت و او را به سمت مسیر شغلی فناوری با دلویت هدایت کرد.

به عنوان یک حرفه‌ای جوان فناوری اطلاعات، بوهر مهارت‌های فنی و رهبری را نشان داد که او را به سمت نقش‌های مدیریتی سوق داد. او به دلیل رهبری موفقیت‌آمیز بسیاری از پروژه‌های بزرگ، از جمله اجرای سیستم‌های برنامه‌ریزی منابع سازمانی (ERP) که فرآیندهای داخلی شرکت را ساده‌تر و کارایی را افزایش داد، شناخته شد. او همچنین پروژه‌های اتوماسیون فرآیندهای رباتیک (RPA) را آغاز و مدیریت کرد که هزینه‌های عملیاتی را کاهش داد و رقابت کلی شرکت را بهبود بخشید.

اخیراً، با آگاهی از پتانسیل هوش مصنوعی (AI) برای بهبود خدمات حرفه‌ای، بوهر و تیم‌های او یک سری آزمایش‌ها و پروژه‌های مهم را آغاز کردند که هدف آن بهره‌گیری از قدرت هوش مصنوعی برای بهبود تصمیم‌گیری، ساده‌سازی خدمات به مشتریان و کشف جریان‌های درآمدی جدید بود.

موفقیت بوهر در دلویت به درک جامع او از کسب‌وکار مرتبط بود. او معتقد بود که راحل‌های فناوری اطلاعات باید با اهداف استراتژیک شرکت همسو باشند، و توانایی بخش فناوری اطلاعات در پیوند فناوری با استراتژی شرکت برای حمایت از رشد سازمان و اطمینان از ماندن دلویت در خط مقدم صنعت خدمات حرفه‌ای کانادا حیاتی بود.

دلویت: یک شرکت خدمات حرفه‌ای

دلویت تاج توماتسو لیمیتد (دلویت) که در سال ۱۸۴۵ در لندن تأسیس شد، به بزرگترین شرکت خدمات حرفه‌ای جهان از نظر درآمد و تعداد کارمند تبدیل شده بود. در سال ۲۰۲۲، دلویت بیش از ۴۱۵'۰۰۰ کارمند در سراسر جهان داشت و درآمد سالانه آن به ۵۹ میلیارد دلار آمریکا می‌رسید.^۳ این شرکت خدمات حرفه‌ای را به چهار شرکت از پنج شرکت Fortune Global 500[®] از طریق شرکت‌های عضو خود در بیش از ۱۵۰ کشور، از جمله کانادا، ارائه می‌داد.^۴ شرکت عضو کانادایی، که به نوآوری معروف بود، حدود ۱۴'۰۰۰ نفر را استخدام می‌کرد و درآمد سالانه آن به ۳ میلیارد دلار آمریکا می‌رسید.^۵

خدمات حرفه‌ای طیف گسترده‌ای از فعالیت‌های تخصصی و مبتنی بر دانش را شامل می‌شود. در مورد دلویت، این خدمات شامل حسابرسی، اطمینان‌بخشی، مشاوره، مشاوره مالی، مشاوره ریسک، مالیات و خدمات مرتبطی است که به سازمان‌های دولتی و خصوصی ارائه می‌شود و به دنبال تخصص و راهنمایی متخصصان مورد اعتماد هستند. بنابراین، این شرکت کارکنان دانش‌محور را استخدام می‌کند - متخصصانی که عمدتاً با اطلاعات و دانش سروکار دارند - مانند وکلا، حسابداران، مشاوران و متخصصان در زمینه‌های خاص کسب‌وکار.

شرکت‌های خدمات حرفه‌ای به دلیل تخصص، حرفه‌ای‌گرایی، تحویل با کیفیت و پایبندی به استانداردهای اخلاقی شناخته می‌شوند. ارزش پیشنهادی جهانی دلویت "استاندارد تعالی بودن" بود.⁶ برای دستیابی به این هدف، شرکت فرهنگی با عملکرد بالا از یادگیری و توسعه مستمر را پرورش داد و به کارمندان خود توانایی ماندن در خط مقدم یک محیط کسب‌وکار به سرعت در حال تغییر را داد.

از منظر فناوری اطلاعات، ارزش پیشنهادی دلویت به معنای ارائه ابزارها و زیرساخت‌های مورد نیاز کارکنان برای افزایش بهره‌وری و خلاقیت آن‌ها بود. برای بوهرا، این به معنای نظارت مستمر بر نوآوری و اجرای سریع فناوری‌های جدید برای اطمینان از این بود که متخصصان شرکت در اوج بازی خود باقی بمانند.

ظهور مدل‌های زبانی بزرگ

مدل‌های زبانی بزرگ (LLM) ها (مدل‌های مبتنی بر هوش مصنوعی مولد هستند که بر روی مجموعه‌ای گسترده از داده‌های متنی از اینترنت، کتاب‌ها و مقالات آموزش دیده‌اند. با استفاده از تکنیک‌های پردازش زبان طبیعی (NLP)، بیشتر LLM ها -به ویژه GPT ها - محتمل‌ترین کلمه یا عبارت بعدی را در یک زمینه خاص پیش‌بینی می‌کنند. آن‌ها می‌توانند متن شبیه به انسان را درک و تولید کنند و اطلاعات مرتبط با زمینه را در مورد طیف گسترده‌ای از موضوعات ارائه دهند. کاربردهای LLM ها زمینه‌های متنوعی مانند ایجاد محتوا، کمک به برنامه‌نویسی و حل مسئله را شامل می‌شود و آن‌ها را به ابزاری قدرتمند برای متخصصان تبدیل می‌کند.

در سال ۲۰۲۲، ظهور مدل‌های هوش مصنوعی مولد قابل دسترسی عمومی، رایگان و قدرتمند، نشانگر رشد LLM ها بود. بیشتر این مدل‌ها دارای رابط‌های کاربرپسند چت‌بات بودند که به افراد با حداقل تخصص فنی امکان استفاده از قابلیت‌های آن‌ها را می‌داد. تا پایان سال، انتشار عمومی یک LLM مبتنی بر GPT موج بی‌سابقه‌ای از علاقه و اشتیاق را برانگیخت. مشخص شد که LLM ها به طور عمیقی بر رفتار کاربر تأثیر خواهند گذاشت و نحوه تعامل افراد با هوش مصنوعی در محیط‌های شخصی و حرفه‌ای را تغییر خواهند داد.

برای کارکنان دانش‌محور مانند متخصصان دلویت، LLM ها می‌توانستند تأثیر مثبت و سازنده‌ای بر اجزای مختلف کار آن‌ها، از جمله تحقیق، تحلیل داده‌ها، ایجاد محتوا و حل مسئله داشته باشند. LLM ها می‌توانستند تحقیقات را با تجزیه و تحلیل سریع حجم عظیمی از اسناد، پروژه‌های تکمیل شده و مطالعات بازار ساده‌تر کنند و زمان صرف شده برای خواندن و خلاصه‌سازی بینش‌های کلیدی را به میزان قابل توجهی کاهش دهند. آن‌ها می‌توانستند مجموعه داده‌های پیچیده را تفسیر کنند تا روندها و نتایج را پیش‌بینی کنند و به متخصصان کمک کنند تا تصمیمات آگاهانه‌تری بگیرند. این مدل‌ها همچنین می‌توانستند پیشنویس پیشنهادها، گزارش‌ها و تصاویر را تولید کنند، خلاقیت آن‌ها را افزایش دهند و سرعت تولید تحویل‌دانی‌ها را بالا ببرند. در نهایت، LLM ها نه تنها می‌توانستند دیدگاه‌های جدیدی در مورد چالش‌های کسب‌وکار ارائه دهند، بلکه می‌توانستند جستجو و سازماندهی دانش را تسهیل کنند، وظایف معمول را خودکار کنند و ارتباطات روزمره، کارایی و اثربخشی را بهبود بخشند.

در حالی که استفاده از LLM های عمومی وعده بهبود کارایی کارکنان دانش‌محور را می‌داد، چالش‌های عمده‌ای مانند نگرانی‌های اخلاقی، مسائل مربوط به حریم خصوصی داده‌های کاربر و تهدیدات امنیتی داده‌ها را نیز معرفی می‌کرد. به دلیل پیچیدگی ذاتی آن‌ها، سوگیری‌های موجود در داده‌های آموزشی LLM ها می‌توانستند عدالت در تصمیم‌گیری را به خطر بیندازند و پاسخگویی را مبهم کنند. مجموعه داده‌های گسترده‌ای که توسط LLM ها پردازش می‌شدند اغلب حاوی اطلاعات حساس بودند و به مدیریت دقیق داده‌ها و پایبندی شدید به قوانین حریم خصوصی برای جلوگیری از نقض داده‌ها نیاز داشتند. علاوه بر این، وابستگی فزاینده به LLM ها خطرات امنیتی را افزایش می‌داد و جذب تهدیدات سایبری را به همراه داشت و نیاز به اقدامات محافظتی قوی برای حفظ اطلاعات اختصاصی و محرمانه داشت.

بوهرا با آگاهی از چالش‌های مرتبط با پذیرش سریع هوش مصنوعی مولد، می‌دانست که باید با احتیاط عمل کند: ادغام موفقیت‌آمیز راحل LLM داخلی دلویت مبتنی بر ابزارهای هوش مصنوعی مولد شخص ثالث به تخصص فنی و یک استراتژی جامع حقوقی، اخلاقی و عملیاتی نیاز داشت.

پیاده‌سازی هوش مصنوعی مولد در دلویت (Gen-D)

راه‌اندازی پروژه

اگرچه مشتریان دلویت نگرانی فوری در مورد استفاده از LLM ها ابراز نکرده بودند، بوهرا و همکارانش خطرات بالقوه مرتبط با افشای داده‌های داخلی و مشتری حساس را تشخیص دادند.

قبل از ترک دفتر برای تعطیلات، بوهرا با بخش مدیریت ریسک دلویت، که مسئول مسائل حریم خصوصی و امنیت بود، تماس گرفت. همه موافق بودند که استفاده نامحدود از یک مدل هوش مصنوعی عمومی برای یک شرکت خدمات حرفه‌ای مانند دلویت غیرقابل قبول است، بنابراین بوهرا باید سریع عمل می‌کرد. او بنابراین رویکردی پیشگیرانه اتخاذ کرد و از فرآیند زمان‌بر معمول شناسایی حامیان، جلب حمایت، تأمین بودجه و جذب ذینفعان برای پشتیبانی از یک ایده جدید صرف‌نظر کرد.

این رویکرد با هر ابتکار دیگری که بوهرا در آن مشارکت داشت تفاوت داشت. هنگامی که او به عنوان یک حرفه‌ای فناوری اطلاعات شروع به کار کرد و عمدتاً بر روی پروژه‌های ERP در مقیاس بزرگ کار می‌کرد، پروژه‌ها با آغاز و برنامه‌ریزی، ارزیابی نیازها و مدیریت فرآیندهای کسب‌وکار شروع می‌شد. اجرا معمولاً از یک فرآیند ساختاریافته و آشنایی پیروی می‌کرد که شامل برنامه‌ریزی اولیه بر اساس مشخصات دقیق و نیازهای مشتری بود. به طور کلی تکرارهای کمی وجود داشت و کاربران نهایی فناوری را پس از توسعه کامل با آزمایش قابلیت استفاده تأیید می‌کردند. در طول سال‌ها، بوهرا به تدریج رویکردهای چابک‌تری برای ساخت و ارائه راحل‌های فناوری اطلاعات را پذیرفت و ترویج داد. برخلاف رویکردهای آشنایی، رویکردهای چابک انعطاف‌پذیری، همکاری و توسعه تکراری را در بر می‌گیرند و انتشارهای کوچک‌تر و افزایشی با بازخورد مستمر کاربران نهایی را ترجیح می‌دهند. این پیاده‌سازی‌ها به فناوری اطلاعات امکان می‌دهد تا به طور نزدیک با کاربران نهایی کار کند و بر ارزش و پاسخ‌های سریع به نیازهای در حال تغییر تمرکز کند. آزمایش قابلیت استفاده به یک مرحله واحد محدود نمی‌شود، بلکه در طول پروژه ادغام می‌شود.

با این ابتکار جدید، هدف ساده‌تر بود: کارکنان دانش‌محور فقط می‌خواستند از قابلیت‌های هوش مصنوعی مولد استفاده کنند. برای بوهرا، هرچه تیم او سریع‌تر می‌توانست یک راحل LLM ایمن و قابل اعتماد را منتشر کند، بهتر بود. برای تسریع فرآیند، بخش فناوری اطلاعات نقش پیشرو و رهبری را بر عهده گرفت، از پروژه حمایت کرد و ارزش آن را نشان داد.

رویکرد فنی

تجزیه و تحلیل گزینه‌های مختلف نشان داد که سریع‌ترین و ایمن‌ترین راه برای به‌رمندی از قابلیت‌های یک LLM ، استفاده از مدل‌های موجود شخص ثالث است. این استراتژی فنی از طریق یک رابط برنامه‌نویسی کاربردی (API) امکان‌پذیر می‌شد که از نظر کاربر تقریباً شفاف می‌بود: کارمندان از یک رابط چت‌بات - با نام Gen-D برای "هوش مصنوعی مولد در دلویت" - استفاده می‌کردند و در عین حال از قابلیت‌های هوش مصنوعی مولد شخص ثالث در محیط دلویت بهره می‌بردند (شکل ۱ را ببینید).

شکل ۱ - طرح کلی سطح بالا از فرآیندهای Gen-D

این معماری سه مزیت اصلی داشت. اول، به کاربران اجازه می‌داد تا درخواست‌ها را به عنوان ورودی ارائه دهند و متن را به عنوان خروجی تولید کنند. دوم، درخواست‌ها را با داده‌های کاربر تقویت می‌کرد تا پاسخ‌های مناسب‌تری برای یک شرکت خدمات حرفه‌ای ارائه دهد. سوم، سازمان را قادر می‌ساخت تا پاسخ‌ها را با تقویت مدل‌ها با داده‌های خود تنظیم کند.

تیم شش نفره NLP بوهرا بر آزمایش با API متمرکز شد. این تیم با پیاده‌سازی و پیکربندی API و توسعه رابط‌های خود شروع کرد، سپس درخواست‌ها و سناریوهای گفتگویی همسو با هدف Gen-D را فرموله کرد. این شامل ایجاد طیفی از پرس‌وجوهای ورودی و توالی‌های گفتگو برای دریافت پاسخ‌های معنادار از مدل بود.

برای حداکثر کردن دقت و مفید بودن Gen-D، تیم NLP یک شخصیت خاص دلویت را که منعکس‌کننده هویت حرفه‌ای شرکت بود، ادغام کرد LLM. ها اغلب از شخصیت‌ها برای تقویت درخواست‌های کاربر با زمینه معنادار استفاده می‌کنند. به عنوان مثال، هنگامی که یک کاربر درخواستی مانند "آیا می‌توانید این مقاله را خلاصه کنید؟" می‌نوشت، درخواست یک دستور داخلی طولانی‌تر اضافه می‌کرد که اطلاعات زمینه‌ای مانند "شما یک حرفه‌ای دلویت هستید؛ پاسخ را به شیوه‌ای تجاری بنویسید [...]" را شامل می‌شد. این استراتژی درخواست تقویت شده به همسو کردن پاسخ‌های مدل با لحن، سبک و تخصص مورد نظر کمک می‌کرد و تعاملات را برای متخصصان دلویت مناسب‌تر و متناسب با زمینه می‌ساخت.

تیم NLP رامحل Gen-D را در تنها دو ماه توسعه داد. تجربه اعضای آن در این زمینه به آن‌ها امکان داد تا به سرعت قابلیت‌های رامحل هوش مصنوعی مولد شخص ثالث را درک و مهار کنند و چالش‌های فنی را در طول راه برطرف کنند. با این حال، همانطور که گفته می‌شود، "مسائل نرم، مسائل سخت هستند"، و بیشتر چالش‌هایی که با آن‌ها روبرو شدند فنی نبودند.

نگرانی‌های مربوط به حریم خصوصی داده‌ها، امنیت و دسترسی

برای بوهرا، مهم‌ترین نگرانی‌های اخلاقی و کسب‌وکار مربوط به خطرات حریم خصوصی و امنیت بود. به عنوان مثال، هنگام تعامل با یک LLM عمومی، متخصصان دلویت ممکن بود به طور ناخواسته داده‌های حساس، غیرناشناس یا محرمانه را افشا کنند. این داده‌ها در معرض خطر ذخیره و استفاده مجدد توسط LLM قرار داشتند و به طور بالقوه می‌توانستند به عنوان مواد آموزشی توسط مدل‌های پایه آن مورد استفاده قرار گیرند LLM. می‌توانست تمام یا بخشی از این پرس‌وجوها را برای اصلاح الگوریتم‌های خود ادغام کند و خروجی‌هایی تولید کند که می‌توانست این اطلاعات حساس را در پاسخ به پرس‌وجوهای سایر کاربران آشکار یا استنباط کند. امکان افشا و نشت بالقوه نگرانی‌های مهمی در مورد حریم خصوصی و امنیت داده‌ها برای شرکتی که داده‌های حساس و محرمانه مشتری را مدیریت می‌کرد، ایجاد می‌کرد.

از آنجا که استفاده از یک LLM عمومی خطرناک بود، تیم NLP بوهرا باید راهی برای اطمینان از حریم خصوصی و امنیت داده‌ها هنگام استفاده از یک مدل هوش مصنوعی خارجی پیدا می‌کرد.

در این زمینه، استفاده از API به ایجاد تعادل مناسب بین تجربه کاربر و حریم خصوصی و امنیت داده‌ها کمک کرد. متخصصان دلویت از یک رابط دلویت، که از طریق اینترنت شرکت قابل دسترسی بود، برای استفاده از قابلیت‌های پیشرفته هوش مصنوعی مولد در فعالیت‌های کاری خود استفاده می‌کردند. در پس‌زمینه، سرویس API داده‌ها را در حال انتقال بین Gen-D و رامحل هوش مصنوعی مولد شخص ثالث رمزگذاری می‌کرد تا از دسترسی غیرمجاز و رهگیری جلوگیری کند. برای تقویت امنیت، ارائه‌دهنده API به پروتکل‌های قوی، از جمله امنیت شبکه و محافظت در برابر تهدیدات، پایبند بود. علاوه بر این، توافق‌نامه‌های قراردادی با فروشنده تضمین می‌کرد که داده‌های ارائه شده به رامحل هوش مصنوعی مولد شخص ثالث با سایر کاربران، مشتریان یا شرکت شخص ثالث به اشتراک گذاشته نمی‌شود. به

عبارت دیگر، Gen-D کاملاً مستقل عمل می‌کرد و داده‌های داخلی و مشتری شرکت در زیرساخت ابری دلویت به طور ایمن ذخیره می‌شد.

اگرچه از نظر فنی، API بخشی از راحل را ارائه می‌داد، تیم NLP می‌دانست که این کافی نیست. راحل هوش مصنوعی مولد شخص ثالث خوب بود، اما کامل، قابل اعتماد یا بدون سوگیری نبود. بوهرا از خود می‌پرسید که آیا باید صبر کنند تا ارائه‌دهنده شخص ثالث این مسائل را حل کند یا به کارمندان خود در مورد خطرات و محدودیت‌های چنین مدل‌هایی آموزش دهد - به آن‌ها کمک کند تا "رفتارهای خوب" را هنگام استفاده از LLM ها اتخاذ کنند.

تیم NLP بوهرا به طور نزدیک با بخش مدیریت ریسک دلویت، دفتر داده و متخصصان مدیریت تغییر همکاری کرد تا خطرات و مزایا را ارزیابی کند و دستورالعمل‌های استفاده از Gen-D را تعیین کند. این همکاری منجر به ایجاد یک "توافقتنامه راهنمایی و محدودیت‌ها" شد که هر بار که یک کاربر دلویت به Gen-D دسترسی پیدا می‌کرد، به صورت پنجره پاپ‌آپ ظاهر می‌شد. این توافقنامه کاربران را در مورد قابلیت‌ها و محدودیت‌های راحل، مانند دانش محدود جهانی، حساسیت به سوگیری‌ها، عدم قابلیت اطمینان بالقوه و ناتوانی در تأیید منابع، آموزش می‌داد. این توافقنامه همچنین شامل نمونه‌هایی از انواع داده‌هایی بود که برای استفاده با Gen-D تأیید شده یا نشده بودند (نمایش ۱ را ببینید).

هر کاربر دلویت موظف بود قبل از استفاده از Gen-D به صراحت به "توافقتنامه راهنمایی و محدودیت‌ها" رضایت دهد. با این کار، کاربران مسئولیت و پاسخگویی تعاملات خود با Gen-D را می‌پذیرفتند. این فرآیند رضایت سه چیز را انجام می‌داد: ۱) کاربران را در قالب ورودی‌هایی که در درخواست‌های خود ارائه می‌دادند و نتایجی که از Gen-D دریافت می‌کردند مسئول می‌کرد، ۲) اهمیت استفاده متفکرانه و مناسب از Gen-D را برای جلوگیری از نتایج بالقوه مضر یا غیراخلاقی تأکید می‌کرد، و ۳) تعهد دلویت به شفافیت را تقویت می‌کرد و آگاهی کاربران از قابلیت‌های هوش مصنوعی و استانداردهایی که هنگام استفاده از Gen-D ملزم به رعایت آن‌ها بودند را تضمین می‌کرد.

برای اهداف حقوقی و انطباق، دلویت باید شفافیت و ردیابی را تضمین می‌کرد. از آنجا که یک درخواست یکسان می‌توانست هر بار پاسخ‌های متفاوتی ایجاد کند، تیم تمام تعاملات کاربران با Gen-D را در یک پایگاه داده ایمن و قوی ثبت می‌کرد. هر ورودی ارائه شده به مدل و خروجی متناظر آن به طور سیستماتیک در پایگاه داده ثبت و ذخیره می‌شد. این رویکرد به ثبت داده‌ها به شرکت امکان می‌داد تا با استانداردهای نظارتی و حسابرسی داخلی مطابقت داشته باشد.

شرکت‌های خدمات حرفه‌ای اغلب پروژه‌های خود را حسابرسی می‌کنند تا از انطباق با مقررات و همسویی با انتظارات باکیفیت و استانداردهای اخلاقی اطمینان حاصل کنند. در زمان نگارش این متن، تنها دو کارمند دلویت می‌توانستند به فعالیت‌های پایگاه داده دسترسی داشته و آن را نظارت کنند.

انسان در حلقه

تیم NLP انسان‌در-حلقه را برای بهبود درک متخصصان دلویت از پیامدهای اخلاقی استفاده از هوش مصنوعی و درگیر کردن آن‌ها در ارائه بازخورد درباره محتوای تولید شده توسط هوش مصنوعی برای همسویی با استانداردهای شرکت و انطباق نظارتی، ادغام کرد.

انسان‌در-حلقه به یک فرآیند بازخورد تعاملی و تکراری اشاره دارد که شامل مشارکت انسان‌ها در طراحی، توسعه و استفاده از سیستم‌های هوش مصنوعی برای بهره‌گیری از نقاط قوت و کاهش نقاط ضعف هوش مصنوعی و هوش انسانی برای دستیابی به نتایج بهینه است. این امر به ویژه در شرکت‌های خدمات حرفه‌ای مانند دلویت اهمیت دارد، جایی که خطر خطاها یا نقض اخلاقیات می‌تواند پیامدهای قابل توجهی داشته باشد و نیاز به ادغام دقیق قضاوت انسانی برای اطمینان از دقت، انطباق و قابل اعتماد بودن تصمیمات مبتنی بر هوش مصنوعی دارد.

انسان در حلقه می‌تواند در سطوح مختلف در طول چرخه عمر کامل سیستم هوش مصنوعی ادغام شود تا هر مرحله از طریق نظارت و تخصص انسانی بهینه شود. در مرحله راه‌اندازی اولیه، انسان‌ها مسئولیت‌ها، مرزهای تصمیم‌گیری، اجرای وظایف و نقش‌های نظارتی خود را تعریف می‌کنند. در مرحله آموزش، انسان‌ها داده‌ها را برچسب‌گذاری می‌کنند و بازخورد ارائه می‌دهند تا دقت وظایف هوش مصنوعی را افزایش دهند. در مرحله عملیاتی، انسان‌ها عملکرد هوش مصنوعی را نظارت می‌کنند و بهبودهای افزایشی را از طریق حلقه‌های بازخورد تسهیل می‌کنند. در مرحله بهبود مستمر، انسان‌ها فرصت‌هایی برای ارتقای هوش مصنوعی شناسایی می‌کنند و در جلسات آموزش مجدد برای بهبود دقت و ارتباط شرکت می‌کنند. کنترل کیفیت شامل مدیریت عدم قطعیت‌ها و انجام حسابرسی‌ها توسط انسان‌ها برای حفظ استانداردها است. در نهایت، در مرحله تعامل با مشتری، انسان‌ها مسائل مربوط به رابط کاربری را که می‌تواند منجر به تجربه کاربری ضعیف شود، حل می‌کنند. در این زمینه، و برای اطمینان از بهبود مستمر، تیم NLP بوهرا بر تأیید انسانی متمرکز شد و مکانیسم بازخوردی را برای تسهیل بهبود مستمر Gen-D ایجاد کرد.

برای این منظور، "توافقنامه راهنمایی و محدودیت‌ها" با افزایش پاسخگویی کاربران، آگاهی را افزایش داد؛ این توافقنامه تأیید انسانی پاسخ‌های تولید شده توسط هوش مصنوعی را برای اهداف اصالت و مالکیت فکری پیشنهاد می‌کرد. هنگام استفاده از Gen-D، از متخصصان دلویت تشویق می‌شد تا پاسخ‌های آن را به طور انتقادی تحلیل و ارزیابی کنند و آن‌ها را با اسناد موجود، پایگاه‌های داده و سایر منابع مرتبط مقایسه کنند تا از نقض احتمالی یا تکثیر تصادفی مواد موجود اطمینان حاصل کنند و همزمان مطمئن شوند که پاسخ‌ها با دستورالعمل‌های اخلاقی و اخلاقی شرکت همسو هستند. این تأیید انسانی با کد رفتار دلویت همسو بود که از تمام متخصصان دلویت می‌خواست به اصول صداقت، یکپارچگی و حرفه‌ای‌گرایی پایبند باشند و استانداردهای رفتار را مشخص می‌کرد.

تیم NLP همچنین یک ویژگی "انگشت شست بالا/پایین" را توسعه داد که با هر پاسخ Gen-D نمایش داده می‌شد و به متخصصان دلویت امکان می‌داد بازخوردی ارائه دهند که به تنظیم دقیق راحل پایه هوش مصنوعی مولد شخص ثالث کمک می‌کرد. (نمایش ۲ را ببینید.) به جای استفاده از یک مکانیسم بازخورد ساده دوتایی، هنگامی که کاربران روی نماد انگشت شست بالا یا پایین مرتبط با پاسخ کلیک می‌کردند، Gen-D فیلدی را نمایش می‌داد که می‌توانستند افکار، نگرانی‌ها یا پیشنهادات خود را بیان کنند. تیم NLP این پیام‌های کیفی را به عنوان بخشی از بهبود مستمر راحل ثبت و تحلیل می‌کرد.

هوش مصنوعی مولد مسئولیت‌پذیر

در سال ۲۰۲۰، دلویت چارچوب Trustworthy AI™ خود را معرفی کرد تا به مشتریان در استقرار مسئولانه و اخلاقی فناوری هوش مصنوعی کمک کند و به سازمان‌ها در برآورده کردن استانداردهای حکمرانی و نظارتی در طول چرخه عمر کامل هوش مصنوعی از توسعه مفهوم اولیه تا طراحی، توسعه، استقرار و بهبود مستمر یاری رساند. دلویت از این چارچوب چندبعدی هوش مصنوعی برای کمک به سازمان‌ها در توسعه محافظت‌های اخلاقی در هفت بعد کلیدی استفاده کرد: شفاف و قابل توضیح، منصفانه و بی‌طرف، قوی و قابل اعتماد، خصوصی، ایمن و امن، مسئولیت‌پذیر و پاسخگو (نمایش ۳ را ببینید).

چارچوب Trustworthy AI™ نقش مهمی در راهنمایی تیم NLP در پیاده‌سازی راحل Gen-D داشت. ابعاد شفافیت و قابل توضیح بودن، مسئولیت‌پذیری، پاسخگویی و حریم خصوصی، تصمیمات گرفته شده توسط تیم NLP در طول توسعه راحل و "توافقنامه راهنمایی و محدودیت‌ها" را هدایت می‌کرد. این ابعاد منجر به ایجاد رویه‌های استفاده واضح، محدودیت دسترسی به فعالیت‌های پایگاه داده و یادآوری‌های پاسخگویی به کاربران شد.

با این حال، استفاده از یک راحل هوش مصنوعی خارجی، از پیش موجود و غیرقابل تغییر، چالش‌های اضافی مانند عدم امکان توضیح، شفافیت و انصاف در خروجی‌های آن را به همراه داشت. معماری LLM ها شامل ساختارهای شبکه بسیار پیچیده با میلیاردها پارامتر بود که درک چگونگی تولید خروجی‌های خاص توسط این مدل‌ها را برای کاربران دشوار

می‌ساخت LLM. های عمومی بر روی طیف گسترده‌ای از مجموعه داده‌ها، مانند کتاب‌ها و مقالات، آموزش دیده بودند که می‌توانستند سوگیری‌های اجتماعی مانند کلیشه‌ها و پیشداوری‌ها را در خود جای دهند و به طور بالقوه باعث می‌شد مدل‌ها پس از یادگیری از این الگوها و ارتباطات، آن سوگیری‌ها را تکرار کنند. بنابراین، کارکنان دانش‌محور حرفه‌ای باید هنگام استفاده از پاسخ‌های تولید شده توسط LLM ها محتاط می‌بودند تا از تداوم یا تشدید سوگیری‌ها، ایجاد مسائل اخلاقی یا آسیب به اعتبار شرکت خود جلوگیری کنند.

بوهرا می‌دانست که در حالی که چارچوب به هدایت و تنظیم دقیق توسعه Gen-D کمک می‌کند، تبدیل اصول گسترده به مراحل خاص و قابل اجرا برای پیاده‌سازی و استفاده از راحل چالش‌برانگیز خواهد بود. در حالی که تصمیم استراتژیک دلویت برای استفاده از یک LLM عمومی پیشرفت سریع و دسترسی ایمن به قابلیت‌های پیشرفته LLM را تسهیل می‌کرد، به معنای پذیرش مصالحه‌های ذاتی در استفاده از فناوری بود که خارج از کنترل یا تأثیر مستقیم شرکت قرار داشت.

ساختش کن، و آن‌ها... استفاده خواهند کرد؟

از آنجا که بوهرا هدفش انتشار یک راحل ایمن و قابل اعتماد در سریع‌ترین زمان ممکن بود، فرآیند پذیرش - و رویکرد مدیریت تغییر - کاملاً با ابتکارات قبلی فناوری اطلاعات متفاوت بود.

تیم NLP کاربران نهایی را در شناسایی موارد استفاده که تعاملات آن‌ها با Gen-D را توصیف می‌کرد، مشارکت نداد، و همچنین موارد استفاده از پیش تعریف شده را در راحل تحمیل نکرد. نیاز واضح بود: کارکنان دانش‌محور فقط می‌خواستند دسترسی سریع به قابلیت‌های هوش مصنوعی مولد داشته باشند. هدف تشویق پذیرش ارگانیک هوش مصنوعی بود در حالی که کاربران راحل را بررسی و آزمایش می‌کردند و استفاده از آن را در صورت لزوم تنظیم می‌کردند. به طور خلاصه، تیم می‌خواست رفتارهای فعلی کاربران با LLM های عمومی موجود را بازتولید کند.

برای بوهرا، فرض اساسی این بود که مردم فقط می‌خواستند از قابلیت‌های LLM ها بهره‌مند شوند. او این رویکرد "ساختش کن و آن‌ها خواهند آمد" را اینگونه خلاصه کرد: "ما نمی‌خواستیم کاربران را مجبور کنیم تا از Gen-D به روشی خاص استفاده کنند. فقط می‌خواستیم بگذاریم آن را بررسی کنند و با تکامل آن استفاده کنند".

در نتیجه، استفاده از Gen-D با ماهیت ظهور خود مشخص می‌شد. به جای از پیش تعیین شدن، استفاده از آن به صورت ارگانیک تکامل یافت زیرا کاربران در نقش‌ها و عملکردهای مختلف راه‌های جدیدی برای بهره‌گیری از قابلیت‌های LLM در کار خود کشف کردند و در نتیجه بهره‌وری و خلاقیت خود را افزایش دادند. این رویکرد به راحل اجازه داد تا به طور مؤثر طیف گسترده‌ای از وظایف را پوشش دهد و آن را به ابزاری همه‌کاره و ارزشمند برای کارکنان دانش‌محور دلویت تبدیل کند.

Gen-D در دو ماه اول پس از راه‌اندازی به طور گسترده مورد پذیرش قرار گرفت، با ۶٬۰۰۰ کاربر منحصر به فرد که با فناوری به روش‌های مختلف درگیر شدند. این کاربران ۱۷۰٬۰۰۰ درخواست ایجاد کردند که نشان‌دهنده تقاضای قوی برای قابلیت‌های LLM بود. علاوه بر این، 85 Gen-D میلیون توکن را پردازش کرد که نشان‌دهنده گستردگی استفاده از آن و قابلیت‌های پردازش زبان گسترده مدل هوش مصنوعی بود.

آینده چه بود؟

تقریباً ده ماه بعد، بوهرا دوباره در پشت میز خود نشسته بود و به تمام چیزهایی که از زمان تصمیم به استفاده از راحل هوش مصنوعی مولد شخص ثالث آموخته بود فکر می‌کرد. یکی از درس‌های کلیدی برای او و تیمش این بود که با فناوری موجود بهترین کار را انجام دهند. بخش‌های فناوری اطلاعات دیگر نمی‌توانستند منتظر بمانند تا فناوری‌ها بالغ شوند و بیشتر نقص‌های اولیه و مشکلات ذاتی آن‌ها قبل از پیاده‌سازی برطرف یا کاهش یابد. ابتکارات مرتبط با هوش

مصنوعی به سرعت در حال ظهور بودند و شرکت‌هایی مانند دلویت باید با کاوش، آزمایش و تست مداوم فناوری‌های جدید در صدر حرکت می‌ماندند.

در حالی که بوهرا به مراحل بعدی Gen-D فکر می‌کرد، نیاز به تنظیم دقیق بخشی از راحل هوش مصنوعی مولد شخص ثالث و توسعه موارد استفاده برای نیازهای خاص کسب‌وکار را تشخیص داد. تیم او قبلاً آزمایش یک چت‌بات اقتصادی را بر اساس همان مدل‌های مبتنی بر هوش مصنوعی آغاز کرده بود، اما با توجه به دیدگاه اقتصادی و انتشارات شرکت در دهه گذشته تنظیم شده بود.

او همچنین نیاز به گسترش Gen-D به سایر شرکت‌های عضو را تشخیص داد. موفقیت Gen-D توجه سایر شرکت‌های عضو را که با استفاده از LLM های عمومی دست و پنجه نرم می‌کردند، جلب کرده بود.

بوهرا می‌توانست پیش‌بینی کند که Gen-D به یک راحل جهانی برای تمام شرکت‌های گروه تبدیل شود. اگرچه این امر چالش‌های فنی به همراه داشت، یادگیری‌های کسب‌وکار شرکت کاندایی می‌توانست برای تسریع موفقیت جهانی دلویت مورد استفاده قرار گیرد.

نمایش‌ها

نمایش ۱

طبقه‌بندی داده‌ها در Gen-D

"توافقتنامه راهنمایی و محدودیت‌ها" نمونه‌هایی از انواع داده‌هایی را که برای استفاده با Gen-D تأیید شده یا نشده بودند، ارائه می‌داد.

جدول ۱: انواع داده‌های تأیید شده

مثال‌ها	انواع داده‌ها
اطلاعات موجود در وبسایت عمومی خارجی دلویت	داده‌های دلویت
اطلاعاتی که تحت حمایت حق چاپ یا سایر شرایط استفاده قرار ندارند	داده‌های عمومی
داده‌های خریداری شده از فروشندگان	داده‌های شخص ثالث خریداری شده از دلان داده‌ها ¹¹
داده‌های تولید شده به عنوان جایگزین داده‌های زنده یا واقعی	داده‌های مصنوعی
اطلاعات شخصی تجمیع شده، ناشناس یا غیرقابل شناسایی	اطلاعات شخصی
N/A	اطلاعات مشتری تأیید شده برای استفاده با ابزار هوش مصنوعی ¹²

جدول ۲: انواع داده‌های تأیید نشده

مثال‌ها

انواع داده‌ها

داده‌های استعدادی در سطح فردی	داده‌های دلویت
اطلاعات (عمومی یا اختصاصی) که شرایط استفاده به صراحت استفاده از آن را ممنوع می‌کند (مانند داده‌های مجله آنلاین یا تحقیقات بازار)	داده‌های عمومی
داده‌های خریداری شده از فروشندگان	داده‌های شخص ثالث خریداری شده از دلالات داده‌ها ¹³
هرگونه اطلاعات شخصی در سطح فردی یا قابل شناسایی	اطلاعات شخصی
¹¹ تنها در صورتی که با مالک قرارداد شخص ثالث تأیید شده باشد که مورد استفاده داده‌های هوش مصنوعی مولد تحت شرایط آن قرارداد مجاز باشد. ¹² تنها در صورتی که اطلاعات مشتری مطابق با شرایط و ضوابط توافقنامه مشتری استفاده شود. ¹³ داده‌های شخص ثالث خریداری شده که برای استفاده دلویت خریداری یا تهیه نشده‌اند.	

نمایش ۲

مکانیسم بازخورد کاربر

ویژگی "انگشت شست بالا/پایین" به هر پاسخ ارائه شده توسط Gen-D اضافه شد، همانطور که در زیر نشان داده شده است.

آیا این پاسخ مفید/دقیق بود؟

بله خیر

نمایش ۳

چارچوب Trustworthy AI™ دلویت¹⁴

جدول ۳: توضیح هر بعد از چارچوب Trustworthy AI™ دلویت

تعاریف	ابعاد
حریم خصوصی کاربر محترم شمرده می‌شود و داده‌ها فراتر از استفاده و مدت زمان مورد نظر و اعلام شده استفاده یا ذخیره نمی‌شوند؛ کاربران می‌توانند در به اشتراک گذاری داده‌های خود انتخاب کنند.	خصوصی
کاربران می‌فهمند که چگونه از فناوری به ویژه در تصمیم‌گیری‌ها استفاده می‌شود؛ این تصمیم‌ها قابل درک، قابل حسابرسی و باز برای بازرسی هستند.	شفاف و قابل توضیح
فناوری به گونه‌ای طراحی و عملیاتی شده است که به دنبال کاربرد، دسترسی و نتایج عادلانه باشد.	منصفانه و بی‌طرف
فناوری به شیوه‌ای مسئولانه از نظر اجتماعی ایجاد و عملیاتی شده است.	مسئولیت‌پذیر

تعاریف

ابعاد

سیاست‌هایی برای تعیین مسئول تصمیمات گرفته شده یا مشتق شده از استفاده از فناوری وجود دارد.	پاسخگو
فناوری خروجی‌های سازگار و دقیق تولید می‌کند، در برابر خطاها مقاومت می‌کند و به سرعت از اختلالات و سوء استفاده‌های پیش‌بینی نشده بهبود می‌یابد.	قوی و قابل اعتماد
فناوری از خطراتی که ممکن است باعث آسیب فیزیکی، عاطفی، محیطی و/یا دیجیتالی فردی و/یا جمعی شود، محافظت می‌شود.	ایمن و امن

¹⁴دلویت-Trustworthy AI: <https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html>