به نام خدا

یک پیشنهاد سریع برای حرکتِ اول بودن

برای ساخت محصول کاربردی AI بومی، چه منابع و زیرساخت‌هایی برای اجرای این پروژه لازم است؟

- پاسخ: **(۱) تأمین GPU و سخت‌افزار، (۲) مدیریت کلاد، (۳) مدل هوش مصنوعی، (٤) نرم‌افزار پایدار و دسترس‌پذیر با پشتیبانی قوی، (٥) کلان‌داده‌ی فارسی، (٦) دسترسی به بازار عمومی.**

۱. مدل هوش مصنوعی، از معادله‌ی ساده تا مدل‌های چندمیلیارد پارامتری

مدل هوش مصنوعی را می‌توان به معادله‌ی «y = ax + b» تشبیه کرد؛ مدلی که به‌جای دو پارامتر ساده، گاه تا صدها میلیارد پارامتر دارد، مانند DeepSeek با ۷۰۰ میلیارد پارامتر. در فرآیند آموزش(Training) ، داده‌های زیادی (X و Y) به مدل داده می‌شود تا بهترین پارامترها برای تقریب دقیق تعیین شوند. پس از آموزش، مدل قادر است با دریافت هر X جدید، Y متناظر را پیش‌بینی کند که به این مرحله «استنتاج» (Inference) میگویند؛ این همان فرآیندی است که وقتی سؤال وارد ChatGPT می‌شود، پاسخ مناسب آن ارائه می‌شود. به معادله، مدل می‌گویند؛ به a و b، پارامتر و به اجرای مدل، استنتاج گفته می‌شود.

۲. ضرورت تفکیک نقش‌ها میان تیم‌های هوش مصنوعی و نرم‌افزار

حالا فرض کنید که قصد داریم این مدل را برای ده میلیون کاربر ایرانی اجرا کنیم و هر کدام روزانه حدود ۲۰ سؤال را مطرح می‌کنند. در این مرحله، دیگر نیازی به تیم هوش مصنوعی نداریم زیرا مدل، پیش‌تر آموزش دیده و آماده شده است. آنچه اکنون نیاز داریم، «یک تیم نرم‌افزاری» است که بتواند در مقیاس بالا، این مدل را در قالب یک نرم‌افزار پایدار و سریع در اختیار کاربران قرار دهد. اما کدام نهادها در کشور، تجربه‌ی چنین مقیاسی را دارند؟ پاسخ مشخص است: کافه‌بازار، دیجی‌کالا، پیام‌رسان‌ها و برخی اپلیکیشن‌هایی که میلیون‌ها کاربر دارند. حتی همراه اول و ایرانسل با وجود زیرساخت ارتباطی گسترده، چنین تجربه‌ای را ندارد چراکه اینها صرفاً اپراتورهای مخابراتی هستند اما در مقابل، اسنپ، تپسی و سایر آپ‌هایی که با عموم مردم در تعامل‌اند، تجربه‌ی عملیاتی لازم را دارند؛ بنابراین ما باید «نقش‌ها» را تفکیک کنیم: از تیم هوش مصنوعی می‌خواهیم مدل را تِرین کند و از تیم نرم‌افزاری می‌خواهیم که مدل آموزش‌داده را در اختیار مردم قرار دهد. **این دو فرآیند، به دو تیمِ جداگانه احتیاج دارند.**

۳. بازاریابی، توزیع و استفاده از ظرفیت برندهای موجود

از منظر بازاریابی، حتی اگر نرم‌افزاری را به‌طور کامل از گوگل خریداری کرده باشید، چند سال زمان می‌برد تا مردم، نام آن را بشنوند و استفاده از آن را بپذیرند. به عنوان نمونه، برند «با سلام» با صرف هزینه‌های بالا توانست فقط تا حدی نام خود را مطرح کند بنابراین چرا باید این مسیر زمان‌بر را مجدداً طی کنیم؟ در حالی‌که هم‌اکنون اپلیکیشن‌ها و برندهایی چون بله، ایتا، روبیکا، تپسی، اسنپ، باسلام و دیجی‌کالا

را در دسترس داریم. باید با یک ــ یا کنسرسیومی از آن‌ها ــ شراکت کنیم و بگوییم: شما این سرویس را ارائه دهید. این دقیقاً همان فرآیند «بازاریابی» است؛ تیم نرم‌افزاری‌ای که مثلاً در باسلام، سرویس دسترسی و خدمات نرم‌افزاری را ارائه می‌کند، در عمل کار تبلیغ و بازاریابی را هم به‌صورت رایگان یا بسیار کم‌هزینه انجام می‌دهد. کافی است باسلام یا دیجی‌کالا در وب‌سایت خود یک بخش جدید با عنوان‌هایی مانند «چت‌بات» یا «هوش مصنوعی» ایجاد کنند؛ کاربران وارد آن می‌شوند، سؤال می‌پرسند و پاسخ دریافت می‌کنند؛ همان کاربرانی که هر روز در سایت و اپ حضور دارند. همین الگو می‌تواند در اسنپ‌فود، بله، روبیکا، ایتا و سایر اپ‌ها نیز پیاده‌سازی شود. به این ترتیب، مسئله‌ی بازاریابی و ارائه‌ی سرویس، با انتخاب یک شریک قدرتمند، هم‌زمان حل می‌شود.

٤. تولید مدل بومی یا مدل آماده؟

پیشنهاد این است که در فاز نخست، «تولید مدل بومی و الگوی زبانی بزرگ» کنار گذاشته شود و به‌جای آن، از یک مدل آماده (عمدتاً Open Source) استفاده شود که تنها با داده‌های ایرانی بازآموزی می‌شود.

راهبرد پیشنهادی (۱): مشارکت چندجانبه

بر این اساس، ما به شراکت و تعامل میان دو یا چند بازیگر نیاز جدی داریم.

- در گام اول و در لایه‌ی سخت‌افزار، GPU باید تهیه شود. این منابع می‌تواند توسط خودمان یا با مشارکت سایر سرمایه‌گذاران تأمین شود.
- گام بعدی، مدیریت این سخت‌افزار در لایه‌ی ابری (Cloud) است.
- گام بعدی، تِرین کردن مدل است که یکی از تیم‌ها، تیم موجود و در مرحله‌ی قرارداد (آقای عباسی) است (از نظر فنی، این تیم دارای نمره‌ی قبولی است). راهکار پیشنهادی، استفاده از مدل‌های ازپیش‌آموزش‌دیده (Pretrained) و تنظیم آن‌ها با داده‌های ایرانی است که نیاز به تحقیق و توسعه‌ی گسترده را کاهش می‌دهد؛ این رویکرد، هزینه‌ها را کنترل کرده، زمان عرضه به بازار را بسیار کوتاه می‌کند (برای مدل AI بومی، پیشنهاد بعدی را طرح کرده‌ایم).
- گام بعدی و در لایه‌ی نرم‌افزار نیز به‌نظر می‌رسد «**پیام‌رسان‌ها**» بهترین گزینه هستند؛ زیرا «**کلان‌داده**»، «**بازار**» و «**بستر اجرایی و نرم‌افزاری**» را به‌طور هم‌زمان در اختیار دارند. اسنپ و تپسی و باسلام، اگرچه زیرساخت فنی و قدرت بازاریابی دارند اما فاقد کلان‌داده‌ی فارسی هستند؛ در حالی‌که بله، ایتا و روبیکا این داده‌ها را در اختیار دارند. برخی نهادهای رسمی نیز وعده داده‌اند داده‌های خود را ارائه دهند، مانند کتابخانه‌ی ملی، ایران‌داک، سازمان ثبت اسناد و قوه‌ی قضاییه اما در عمل، این داده‌ها واگذار نمی‌شوند؛ در نتیجه، تکیه بر پیام‌رسان‌ها واقع‌بینانه‌ترین تصمیم برای یک عملکرد سریع است.

راهبرد پیشنهادی (۲): الهام گرفتن از تجربه‌ی یاندکس (Yandex)

- در زمینه‌ی تولید مدل‌های هوش مصنوعی (بومی و تِرین‌شده)، تیم‌هایی هستند که چند سال است در این حوزه فعالیت می‌کنند. تیم‌های «همراه اول» و «همکاران سیستم» فعال هستند؛ هر تیم، نهایتاً ۵ نفر عضو اصلی و کلیدی دارد. این تیم‌ها با توجه به نیازهای سازمانی، در حوزه‌ی پشتیبانی، پاسخ‌گویی به پرسش‌ها و توسعه‌ی چت‌بات‌ها فعالیت دارند. تیم دیگری به نام «ژرفا» نیز وجود دارد که احتمالاً زیرمجموعه‌ی واجا. شده است و عملاً از اکوسیستم بازار، خود را حذف کرده است.
- این تیم‌ها و چند تیم دیگر، بخشی از مسیر موردنظر ما را رفته‌اند و سرویس هم ارائه کرده‌اند اما کمتر کسی نام سایت و محصول اینها را می‌شناسد! چرا؟ چون گمان میکرده‌اند با الگوهای بازاریابی معمول و نیز خرید کلان‌داده‌ها، میتوانند به کاربران میلیونی دست پیدا کنند (این مسیر، همان پیشنهادی است که در طرح آقای عباسی آمده است)؛ بنابراین ما نباید این مسیر اشتباه را دوباره تکرار کنیم.
- در نهایت، در زمینه‌ی تیم‌های هوش مصنوعی، تیم فعلی را باید وارد کار کرد اما از آنها باید تِرین کردن مدل را توقع داشت و پیشنهاد میشود حلقه‌های پیشین و پسین را با الگوی مشارکتی، پیش ببریم.
- تجربه‌ی هوش مصنوعی یاندکس (بومی روسیه) هم همین را میگوید. شرکت یاندکس از سال ۲۰۱۶ و همزمان با معرفی GPT-3 و در شرایطی که این فناوری برای بسیاری ناشناخته بود، سرمایه‌گذاری بلندمدت خود را در حوزه مدل‌های زبانی آغاز کرد. این شرکت طی ۱۸ ماه، موفق شد مدل بومی ۳۵ میلیارد پارامتری خود را توسعه دهد که از زبان‌های روسی، ازبکی، عربی، ترکی و انگلیسی پشتیبانی می‌کند (مقایسه کنید با مدل ۷۰۰ میلیارد پارامتری دیپ‌سیک). یاندکس در این مدت، برای این محصول از چهار سرور مجهز به ۸ کارت گرافیک و با بودجه‌ای در حدود دو میلیون دلار بهره برد و اخیراً بخشی از دستاوردهای خود (مدل ۸ میلیارد پارامتری) را به صورت اوپن‌سورس منتشر کرده است.
- این تجربه‌ی موفق، ثابت کرد که با برنامه‌ریزی استراتژیک و مدیریت منابع، حتی با امکانات محدود نیز می‌توان به دستاوردهای قابل توجهی در توسعه‌ی مدل‌های زبانی کاملاً بومی دست یافت.
- در مذاکرات اخیر، نمایندگان یاندکس اعلام کردند در سطوح مختلفی برای همکاری و تعامل با دانشگاه، آماده هستند، اعمّ از سرمایه‌گذاری مشترک، توسعه‌ی زیرساخت، توسعه‌ی مدل زبانی بر اساس زبان فارسی و مشاوره برای راه‌اندازی مدل بومی ایرانی.