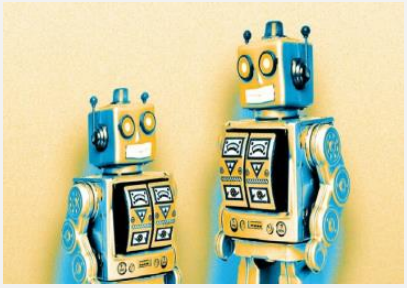

شنبه 06 مرداد 1403 - 06:16

شرکت‌های فناور به دنبال مدل‌های کوچک‌تر هوش مصنوعی هستند

بسیاری از توسعه‌دهندگان فناوری تمایل دارند تا از نمونه‌های کوچک‌تر مدل‌های هوش مصنوعی استفاده کنند. این گزارش به بررسی دلایل این تمایل می‌پردازد.

بسیاری از توسعه دهندگان فناوری تمایل دارند تا از نمونه های کوچک تر مدل های هوش مصنوعی استفاده کنند. این گزارش به بررسی دلایل این تمایل می پردازد.

به گزارش ایسنا، شرکت «اوپن ای آی»(OpenAI) روز پنجشنبه GPT-4o mini را معرفی کرد که نسخه کوچک تر و ارزان تر مدل هوش مصنوعی GPT-4o آن است.

به نقل از فست کمپانی، اوپن ای آی یکی از چندین شرکت هوش مصنوعی است که نسخه ای از بهترین مدل بنیادین خود را توسعه داده تا مقداری از هوش را با سرعت و مقرون به صرفه شدن عوض کند. چنین مبادله ای می تواند به توسعه دهندگان بیشتری امکان دهد تا برنامه های خود را با هوش مصنوعی تقویت کنند و شاید در آینده روزنه های جدیدی را برای برنامه های پیچیده تر باز کند.

بزرگ ترین مدل های زبانی بزرگ از میلیاردها یا تریلیون ها پارامتر استفاده می کنند تا طیف گسترده ای را از استدلال و وظایف مرتبط با پرس وجو انجام دهند. آنها با حجم بزرگی از داده ها آموزش دیده اند که موضوعات گوناگونی را پوشش می دهند. از سوی دیگر، مدل های زبانی کوچک فقط از میلیون ها یا ده ها میلیون پارامتر برای انجام دادن وظایف کمی استفاده می کنند و به قدرت محاسباتی کمتر و مجموعه کوچک تری از داده های آموزشی متمرکز نیاز دارند.

برای توسعه دهندگانی که برنامه های ساده تری دارند، مدل های زبانی کوچک ممکن است تنها گزینه قابل اجرا باشند. اوپن ای آی می گویدGPT-4o mini تا ۶۰ درصد ارزان تر از GPT-3.5 Turbo است که پیش از این مقرون به صرفه ترین مدل این شرکت برای توسعه دهندگان بود.

یک مورد دیگر، بحث سرعت است. بسیاری از کاربردهای هوش مصنوعی به دانش عمومی گسترده یک مدل زبانی بزرگ نیاز ندارند. آنها ممکن است به پاسخ های سریع تری برای پرسش های ساده تر نیاز داشته باشند. «مایک اینتراتور»(Mike Intrator) مدیرعامل «کورویو»(CoreWeave) که میزبان مدل های هوش مصنوعی در فضای ابری است، گفت: منظور از تأخیر، زمان مورد نیاز برای یک برنامه هوش مصنوعی به منظور دریافت پاسخ از یک مدل در فضای ابری است. اگر فرزند من مقاله ترم خود را با کمک هوش مصنوعی بنویسد، تأخیر مشکل بزرگی نیست اما اگر بخواهد از آن برای جراحی یا رانندگی خودران استفاده کند، تأخیر با تأثیر بسیار بیشتری همراه می شود.

اینتراتور خاطرنشان کرد: مدل های مورد استفاده در خودروهای خودران باید به اندازه ای کوچک باشند که روی یک تراشه رایانه ای اجرا شوند، نه در یک سرور فضای ابری.

مدل GPT-4o mini کوچک تر از مدل های دیگر است اما هنوز آن قدر کوچک نیست که روی دستگاهی مانند تلفن همراه یا کنسول بازی اجرا شود. بنابراین، باید مانند سایر مدل های شرکت اوپن ای آی روی یک سرور در فضای ابری اجرا شود.

کلید نسل بعدی برنامه های هوش مصنوعی

«رابرت نیشیهارا»(Robert Nishihara) یکی از بنیان گذاران و مدیرعامل شرکت «انی اسکیل» (Anyscale) گفت: امروزه بیشتر برنامه های کاربردی مبتنی بر هوش مصنوعی شامل یک پرس وجو یا چند پرس وجو برای مدلی هستند که در فضای ابری اجرا می شود اما برنامه های پیشرفته به پرس وجوهای زیادی برای مدل های گوناگون نیاز دارند. به عنوان مثال، برنامه ای که به شما کمک می کند یک اقامت گاه اجاره ای را انتخاب کنید، ممکن است از یک مدل برای ایجاد معیارهای انتخاب، یک مدل دیگر برای انتخاب گزینه های اجاره و یک مدل دیگر برای امتیاز دادن به هر یک از آن گزینه ها استفاده کند. کارگردانی و سازمان دهی همه این پرسش ها یک تجارت پیچیده است.

عملکرد مدل ها مهم است اما سرعت و هزینه آنها نیز به یک اندازه اهمیت دارد. اوپن ای آی این موضوع را می داند؛ همان طور که شرکت هایی مانند متا و گوگل می دانند.

نیشیهارا ادامه داد: تلاش های این شرکت ها به منظور کوچک سازی مدل های هوش مصنوعی برای برنامه های پیچیده تر مانند دستیارهای شخصی بسیار مهم هستند.

اوپن ای آی اندازه پارامتر مدل های خود را فاش نمی کند اما مدل جدید آن احتمالا به اندازه Claude 3 Haiku شرکت «آنتروپیک»(Anthropic) و Gemini ۱.۵ Flash گوگل است.

اوپن ای آی گفت: GPT-4o mini عملکرد بهتری نسبت به مدل های مشابه در آزمایش ها دارد.

اوپن ای آی خاطرنشان کرد که توسعه دهندگان اپلیکیشن می توانند از طریق یک واسط برنامه نویسی کاربردی به GPT-4o mini دسترسی داشته باشند و مدل های جدید نیز از برنامه ChatGPT پشتیبانی می کنند.