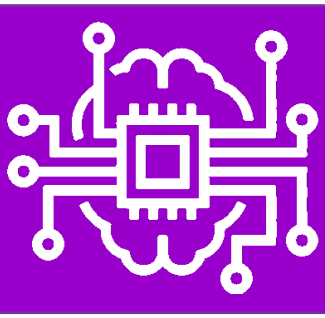


عنوان مقاله: مدل زبانی بزرگ (LLM) چیست؟ آشنایی با نحوه کار، ویژگی ها و کاربردها

نویسنده مقاله: تیم فنی نیک‌آموز

تاریخ انتشار: ۱۷ خرداد ۱۴۰۳

منبع: https://nikamooz.com/what-is-a-large-language-model/

مدل زبانی بزرگ (LLM) جادوی عصر ماست؛ مثل گوی بلورین به همه سؤالات شما جواب می‌دهد و همچون آینه جادویی، هرچه بخواهید را در لحظه برایتان به تصویر می‌کشد. این مدل‌ها با مجموعه عظیمی از داده آموزش‌دیده و با تکیه بر الگوریتم‌های هوش مصنوعی، توانایی درک متون و پاسخ به کاربر را دارند. در ادامه مطلب به شما می‌گوییم LLM چیست، چگونه کار می‌کند و بهترین نمونه‌های آن کدامند.

مدل زبانی بزرگ (LLM) چیست؟

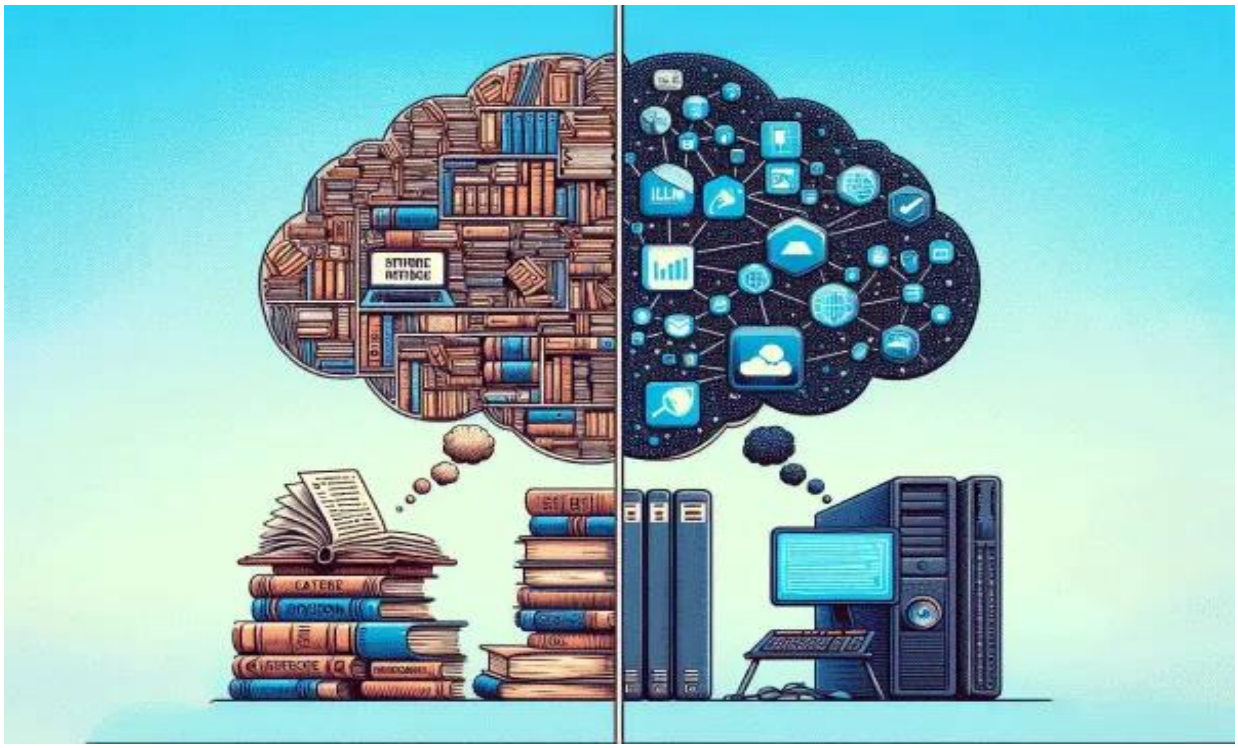
مدل زبانی بزرگ (Large Language Model) نوعی برنامه هوش مصنوعی است که می‌تواند زبان انسان یا دیگر مجموعه داده‌های پیچیده را تفسیر کند. این مدل‌ها بر پایه یادگیری ماشین (Machine Learning) و نوعی شبکه عصبی به نام ترنسفورمر (Transformer) توسعه می‌یابند. مدل‌های زبان بزرگ برای درک ارتباط بین حروف، کلمات و جملات، از نوعی یادگیری ماشین به نام یادگیری عمیق (Deep Learning) استفاده می‌کنند. یادگیری عمیق به مدل امکان می‌دهد که ازطریق تجزیه‌وتحلیل داده‌های حجیم، به‌طور خودکار الگوها و روابط پیچیده را در داده‌ها کشف کنند. با این فرآیند می‌توان پردازش زبان طبیعی، ترجمه، تشخیص تصویر و مواردی از این دست را با سرعت و دقت بالا توسط هوش مصنوعی انجام داد.

اکثر LLM ها روی چند پتابایت داده متنی شامل صفحات سایت، کتاب، مجلات و... در سطح اینترنت آموزش داده می‌شوند. کیفیت نمونه‌ها بر یادگیری مدل تأثیر می‌گذارد و از این رو، توسعه‌دهندگان، مجموعه‌های داده را پالاش کرده و محتوای نادرست، نژادپرستانه یا آزارگرانه را از آن حذف می‌کنند.

تاریخچه و تکامل مدل های زبانی

رؤیای ساخت ماشین‌هایی که زبان انسان را درک کنند با پیشرفت در زمینه پردازش زبان طبیعی (NLP) شکل گرفت. مدل‌های اولیه در درک جملات پیچیده مشکل داشتند اما در دهه ۹۰ میلادی ظهور یادگیری عمیق انقلابی در این زمینه ایجاد کرد. در دهه ۲۰۰۰ توسعه الگوریتم‌های بهینه‌تر، به‌ویژه شبکه‌های عصبی بازگشتی (RNN) و شبکه‌های حافظه بلندمدت و کوتاه‌مدت (LSTM)، عملکرد مدل‌ها را در درک جملات بهبود بخشید. عصر جدید هوش مصنوعی از حدود یک دهه قبل و به لطف افزایش قدرت محاسباتی و دسترسی به حجم عظیمی از داده‌های متنی ممکن شد. مدل‌های فعلی به آسانی الگوها و روابط بین کلمات را تشخیص می‌دهند و انتظار می‌رود شکل تعامل انسان و رایانه را متحول کنند .

هزینه پیاده‌سازی و به‌کارگیری LLM در فرآیندهای مختلف به‌طور پیوسته رو به کاهش است. در سال ۲۰۲۰، ارزیابی محصولات با استفاده از GPT-2 حدود ۱۰,۰۰۰ دلار هزینه داشت. امروزه، GPT-4 نتایج مشابهی را با هزینه‌ای حدود ۳,۰۰۰ دلار ارائه می‌دهد. این روند باعث شده است تا LLM ها برای کسب‌وکارها مقرون‌به‌صرفه و ازنظر اقتصادی توجیه‌پذیر شوند.



مدل زبانی بزرگ (LLM) چگونه کار می کند؟

مدل‌های زبانی بزرگ بر سه پایه یادگیری عمیق، شبکه‌های عصبی و مدل‌های ترنسفورمر متکی هستند که در ادامه آن‌ها را توضیح می‌دهیم.

یادگیری عمیق

مجموعه داده‌ای که مدل زبانی با آن آموزش می‌یابد، چند هزار ترابایت است و کورپوس (Corpus) نام دارد. مرحله اول، آموزش یادگیری بدون نظارت با داده‌های بدون برچسب است که طی آن، مدل شروع به کشف روابط بین کلمات و مفاهیم مختلف می‌کند. گام بعدی، تنظیم دقیق مدل با یادگیری خودنظارتی است. در این سطح، بخشی از داده‌ها برچسب‌گذاری شده‌اند که به مدل برای شناسایی دقیق‌تر مفاهیم مختلف کمک می‌کند.

مدل زبانی بزرگ به لطف یادگیری عمیق و آموزش با مجموعه دیتا قدرت استنتاج پیدا می‌کند. بدین ترتیب، مدل پس از تحلیل میلیاردها جمله درک می‌کند منظور از کلمه «دایی» در متنی که درباره فوتبال است با متنی که درباره روابط فامیلی است، فرق می‌کند.

شبکه های عصبی

برای دستیابی به این نوع یادگیری عمیق، مدل‌های زبانی بزرگ بر پایه شبکه‌های عصبی ساخته می‌شوند. همانطور که مغز انسان از نورون‌های به هم پیوسته تشکیل شده که سیگنال به یکدیگر ارسال می‌کنند، شبکه عصبی مصنوعی نیز شامل شبکه‌ای از گره‌های به هم متصل است. این شبکه‌ها از چندین «لایه» تشکیل شده‌اند: یک لایه ورودی، یک لایه خروجی و یک یا چند لایه میانی. این لایه‌ها تنها در صورتی اطلاعات را به هم منتقل می‌کنند که دقت خروجی خودشان از یک آستانه مشخص عبور کند.

مدل های ترنسفورمر

شبکه‌های عصبی به‌کاررفته در مدل‌های زبانی بزرگ، ترنسفورمر نامیده می‌شوند. این مدل‌ها توان درک و یادگیری بافت متن را دارند؛ قابلیتی که برای درک زبان انسان حیاتی است. مدل ترنسفورمر بر پایه مکانیزم ریاضی خودتوجهی (Self-Attention) روابط ظریف بین عناصر موجود در یک توالی (مثل کلمات در یک متن) را درک می‌کند.

ترنسفورمر از دو لایه انکودر و دیکودر تشکیل شده که در آن، هر آیتم خاص (مثلا کلمه) توکن (Token) نامیده شده و وزنی به آن اختصاص پیدا می‌کند. لایه انکودر ورودی‌ها را دریافت کرده و آن‌ها را مرحله‌به‌مرحله به کد داخلی تبدیل می‌کند. لایه دیکودر نیز که از شبکه عصبی تشکیل شده، کد دریافتی را براساس دستورالعمل‌ها به خروجی نهایی تبدیل می‌کند .

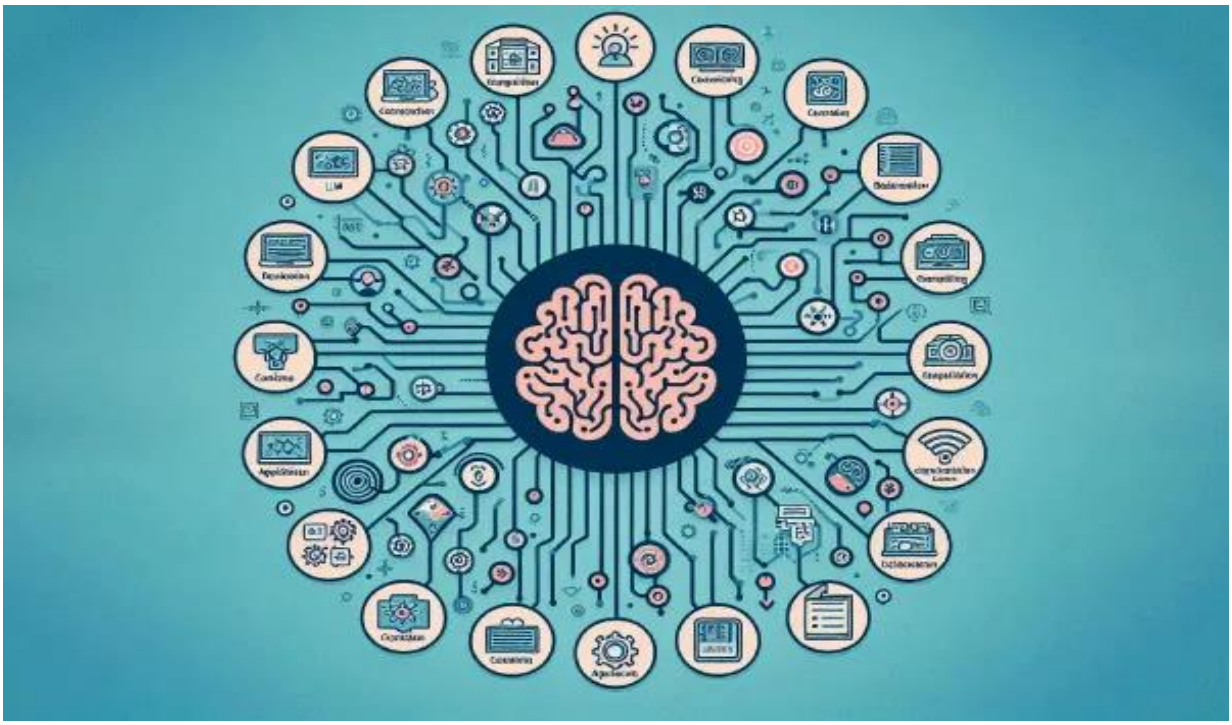
برای مثال، اگر هدف ما ترجمه فارسی به انگلیسی باشد، انکودر متن فارسی را به نمایش برداری از توکن‌ها تبدیل کرده و دیکودر از بردار برای تولید متن انگلیسی بهره می‌برد. در این بین، هر دو بخش از مکانیزم توجه برای یافتن ارتباط بین کلمات استفاده می‌کنند. این رویکرد، مدل‌های زبان بزرگ را قادر می‌سازد تا زبان انسان را حتی زمانی که مبهم بیان شده یا برایشان تازگی دارد، تفسیر کنند.

کاربردهای مدل زبانی بزرگ (LLM)

مدل‌های زبان بزرگ را می‌توان برای انجام وظایف مختلفی، از تولید محتوا گرفته تا تحلیل داده و برنامه‌نویسی، آموزش داد. یکی از شناخته‌شده‌ترین کاربردهای آن‌ها، استفاده به‌عنوان هوش مصنوعی مولد (Generative AI) در مدل‌های جمینای گوگل یا **ChatGPT** است. در این مدل‌ها هنگامی که یک پرامپت به سیستم داده شود، متنی را به‌عنوان پاسخ تولید می‌کنند. برای مثال، ChatGPT می‌تواند در پاسخ به ورودی‌های کاربر، مقاله، شعر، نمودار و سایر اشکال متنی را تولید کند. از LLM ها می‌توان در موارد زیر بهره برد:

- **تولید محتوا:** توانایی تولید متن در هر موضوعی که LLM روی آن آموزش دیده، یکی از کاربردهای اصلی آن‌هاست. پاسخ بسته به درخواست کاربر می‌تواند در سطح یک کودک ۵ ساله یا از دید یک استاد دانشگاه نوشته شود.
- **ترجمه:** برای مدل‌های زبانی بزرگی که روی چندین زبان آموزش دیده‌اند، قابلیت ترجمه از یک زبان به زبان دیگر، یک ویژگی رایج است. جمینای گوگل توانایی ترجمه متن به فارسی از هر زبانی را در سطحی قابل قبول دارد.

- **خلاصه‌سازی محتوا:** می‌توانید از مدل‌های زبانی بخواهید نکات مهم یک متن طولانی یا حتی کتابی کامل را در چند پاراگراف برای شما خلاصه کنند.
- **بازنویسی محتوا:** بازنویسی بخشی از متن قابلیت دیگری است که LLM ها ارائه می‌دهند. برای مثال، اگر درحال نوشتن مقاله انگلیسی هستید، ChatGPT می‌تواند ایرادات گرامری و نگارشی را در متن شما تشخیص داده و رفع کند.
- **طبقه‌بندی و دسته‌بندی:** LLM داده‌ها را براساس دسته‌ها، موضوعات یا سایر معیارها طبقه‌بندی می‌کند. این امر برای وظایفی مانند طبقه‌بندی ایمیل‌ها، شناسایی اسپم و دسته‌بندی اسناد مفید خواهد بود.
- **تحلیل داده:** از LLM های آموزش‌دیده می‌توان برای تحلیل اطلاعات اقتصادی، اجتماعی، پزشکی و... بهره برد. برای مثال، در معاملات مالی و پیش‌بینی نوسانات بازار، تجزیه‌وتحلیل پرونده‌های پزشکی، تحلیل رسانه‌های اجتماعی و درک احساسات عمومی نسبت به محصولات و خدمات خود، از آن‌ها استفاده می‌شود.
- **چت‌بات‌ها:** LLM ها می‌توانند مکالمه با کاربر را به روشی امکان‌پذیر سازند که به مراتب از نسل‌های قبلی هوش مصنوعی طبیعی‌تر است. در نسخه جدید ChatGPT، از روی لحن صدای کاربر، شرایط روحی و عواطف را تشخیص داده و با آن همدلی می‌کند.
- **پرسش و پاسخ:** مدل زبانی بزرگ حتی اگر سؤالات تخصصی، چالش‌برانگیز یا نیازمند استدلال باشد، قادر به ارائه پاسخ خواهد بود. این مورد برای وظایفی مانند پاسخگویی به سؤالات مشتریان و ارائه گزارش‌های تحلیلی، یک مزیت است.



ویژگی ها و مزایای مدل زبانی بزرگ (LLM)

مدل‌های زبانی بزرگ به‌عنوان پدیده دوران ما، مزایای متعددی دارند که شامل این موارد می‌شود:

- **توسعه و انطباق‌پذیری:** مدل‌های زبانی بزرگ قابلیت سفارشی‌سازی بالایی دارند و درصورت تغذیه با داده‌های کافی، می‌توان برای موارد خاص مدنظر شرکت یا سازمان از آن‌ها بهره برد.
- **انعطاف‌پذیری:** یک LLM واحد را می‌توان در سازمان‌ها و برنامه‌های کاربردی برای وظایف مختلف و در اشکال متنوع به کار گرفت.
- **یادگیری:** مدل زبانی بزرگ با قرارگرفتن در معرض داده‌ها و بازخوردهای جدید، نکات جدیدی یاد گرفته و خود را سازگار می‌کنند. این امر آن‌ها را در وظایف نیازمند انعطاف‌پذیری و تطبیق‌پذیری، به ابزاری کارآمد تبدیل می‌کند.
- **عملکرد سریع:** مدل‌های زبان بزرگ عموماً عملکرد بالایی دارند و قادر به تولید پاسخ‌های سریع در لحظه هستند. برای مثال، کاری که برای برنامه‌نویس یک روز زمان می‌برد، هوش مصنوعی در ثانیه تحویل می‌دهد.
- **دقت بالا:** هنوز در ابتدای راه هوش مصنوعی هستیم اما نسخه‌های فعلی با افزایش تعداد پارامترها، دقت بسیار بالایی را در پاسخ به درخواست‌های کاربران ارائه می‌کنند.
- **سهولت آموزش:** بسیاری از مدل‌های زبانی بزرگ روی داده‌های برچسب‌گذاری‌نشده، آموزش داده می‌شوند که به تسریع فرآیند آموزش کمک می‌کند.
- **بهره‌وری عالی:** مدل‌های زبان بزرگ با خودکارسازی وظایف روتین، زمان موردنیاز برای انجام کارها را به‌شدت کاهش می‌دهند .

چالش ها و محدودیت های مدل زبانی بزرگ (LLM)

در کنار مزایای متعدد، مدل زبانی بزرگ (LLM) درگیر برخی چالش‌ها و محدودیت‌ها هستند که شامل موارد زیر می‌شود:

- **هزینه بالا:** اجرای مدل‌های هوش مصنوعی در ابعاد ChatGPT ، به توان پردازش سرسام‌آوری نیاز دارد که هزینه زیادی را به شرکت‌ها تحمیل می‌کند.
- **هزینه‌های عملیاتی:** پس از دوره آموزش و توسعه، هزینه عملیاتی مدل نیز می‌تواند برای سازمان‌ها بسیار بالا باشد.
- **سوگیری (Bias):** یکی از خطرات مدل‌های آموزش‌دیده با داده‌های بدون برچسب، سوگیری است. برای مثال، ممکن است هوش مصنوعی، افراد آفریقایی‌تبار را بیشتر از سفیدپوست‌ها در معرض جرم تلقی کند.
- **توهم:** توهم هوش مصنوعی (Hallucination) زمانی رخ می‌دهد که یک LLM ، پاسخی نادرست ارائه دهد که مبتنی‌بر داده‌های آموزشی نباشد. برای مثال، هوش مصنوعی ممکن است در یک مقاله، به مرجعی استناد کند که درواقع وجود خارجی ندارد.
- **پیچیدگی:** LLM های مدرن با میلیاردها پارامتر، فناوری‌های فوق‌العاده پیچیده‌ای هستند و عیب‌یابی آن‌ها می‌تواند بسیار دشوار باشد.
- **توکن‌های مخرب:** از سال ۲۰۲۲ ایجاد توکن‌های مخرب با هدف اختلال در عملکرد مدل‌های زبانی به یک روند نوظهور تبدیل شده است.

- **خطرات امنیتی:** کاربران ممکن است برای افزایش بهره‌وری خود، داده‌های امن و محرمانه را در مدل‌های زبانی بارگذاری کنند. از آنجا که مدل زبانی بزرگ از ورودی‌ها برای آموزش خود استفاده می‌کند، ممکن است در پاسخ به پرسش‌های کاربران دیگر، داده‌های محرمانه را فاش کند. همچنین از LLM ها می‌توان برای طراحی حملات فیشینگ علیه سازمان‌ها استفاده کرد.

آینده مدل زبانی بزرگ

سرمایه‌گذاری‌های هنگفت برروی حوزه هوش مصنوعی انجام شده که به تحول و پیشرفت‌های قابل توجهی در این زمینه منجر خواهد شد. انتظار می‌رود که در زمینه مدل‌های زبانی، شاهد این دستاوردها باشیم:

- **بی‌نیازی از داده‌های جدید:** LLM ها به‌زودی داده‌های آموزشی موردنیازشان را خود تولید می‌کنند که آن‌ها را از وابستگی به داده‌های جدید برای بهبود عملکرد رها می‌سازد. تکنیک‌هایی مانند تولید و پالایش پاسخ‌ها می‌توانند عملکرد آن‌ها را به‌طور قابل توجهی ارتقا داده و کمبود داده‌های آموزشی را جبران می‌کند.
- **راستی‌آزمایی خودکار:** مدل‌های زبانی برزگ فعلی، مستعد اشتباه هستند اما به‌زودی می‌توانند اطلاعات لحظه‌ای را از منابع خارجی معتبر بازیابی کنند که باعث شفافیت و اعتمادسازی بهتر آن‌ها می‌شود. مدل‌هایی مانند WebGPT و Sparrow از DeepMind، پیشگامان این دسته هستند.
- **معماری ساده‌تر:** برخلاف مدل‌های متراکم که تمام پارامترها را برای یک وظیفه فعال می‌کنند، مدل‌های پراکنده فقط مرتبط‌ترین پارامترها را فعال می‌کنند و به همین دلیل، ازنظر محاسباتی کارآمدتر هستند. این معماری که در مدل‌های GLaM گوگل و Mixture of Experts متا به کار رفته، نویدبخش عملکرد بهتر نسبت به مدل‌های متراکم سنتی با استفاده از منابع کمتر است.
- **استدلال قوی‌تر:** توان مدل‌های زبانی در استدلال منطقی، کاهش سوگیری و استدلال چندمدله (شامل صدا، تصویر، ویدیو، متن و کد) به‌طور قابل توجهی بهبود می‌یابد. مدل‌هایی مانند GPT-5 ، LLAMA 3 و Gemini Ultra به استدلال منطقی دست می‌یابند که دسترسی به پلتفرم‌های شخصی را برای کسب‌وکارها تسریع می‌کند.
- **تولید محتوای سفارشی:** انتظار می‌رود مدل زبانی بزرگ با درنظرگرفتن جزئیاتی مانند رفتار کاربر، اهداف بازاریابی و مواردی از این دست، امکان تولید محتوای شخصی‌سازی‌شده را فراهم آورند. این محتوا می‌تواند شامل هر چیزی، از مقالات خبری و خوراک رسانه‌ای گرفته تا محتوای تبلیغاتی هدفمند باشد.

نمونه هایی از مدل زبانی بزرگ (LLM)

ChatGPT احتمالاً شناخته‌شده‌ترین مدل زبانی بزرگ است اما در کنار آن، مدل‌های کارآمد دیگری نیز برای اهداف مختلف توسعه داده شده که در ادامه به معرفی آن‌ها می‌پردازیم.

گوگل جمینای (Gemini)

جمینای (Gemini) خانواده‌ای از مدل‌های زبانی بزرگ (LLM) گوگل است که با زبان فارسی سازگاری کامل دارد و در اکثر ارزیابی‌ها، از GPT-4 عملکرد بهتری داشته است. مدل‌های جمینای چندرسانه‌ای هستند؛ به این معنی که علاوه‌بر متن، تصاویر، صدا و ویدئو را نیز پردازش می‌کند. این مدل در بسیاری از برنامه‌ها و محصولات گوگل ادغام شده است.

جمینای در سه نسخه بزرگ (Ultra)، حرفه‌ای (Pro) و کوچک (Nano) ارائه می‌شود. Ultra بزرگترین و توانمندترین مدل است، مدل Pro میان‌رده و Nano کوچک‌ترین مدل است که برای اجرای وظایف روی دستگاه طراحی شده است.

OpenAI چت جی پی تی (ChatGPT)

OpenAI با مدل زبانی ChatGPT نگاه‌ها را به‌سوی هوش مصنوعی خیره کرد. جدیدترین نسخه این خانواده، GPT-4 Omni (با نام اختصاری GPT-4o) است که بهبودهای قابل توجهی نسبت به مدل قبلی ارائه می‌دهد. GPT-4o تعامل طبیعی‌تری با انسان برای ChatGPT ایجاد می‌کند و یک مدل چند حالته بزرگ است که ورودی‌های مختلفی ازجمله صدا، تصویر و متن را می‌پذیرد. مدل مثل یک مخاطب عادی با کاربر به صحبت می‌نشیند و حتی عواطف و احساسات را از روی لحن و حرف‌های کاربر درک می‌کند.

GPT-4o می‌تواند در طول تعامل، تصاویر یا صفحه نمایش را ببیند و در مورد آن‌ها سؤال بپرسد یا به سؤالات پاسخ دهد. پاسخگویی GPT-4o در ۲۳۲ میلی‌ثانیه انجام می‌شود که مشابه زمان پاسخگویی انسان و سریع‌تر از GPT-4 Turbo است. مدل GPT-4o رایگان بوده و برای محصولات توسعه‌دهندگان و مشتریان در دسترس خواهد بود.

متا Llama

متا یا همان فیسبوک سابق با مدل Llama وارد میدان رقابت شده که در سال ۲۰۲۳ منتشر شد. Llamaدر ابتدا تنها برای محققان و توسعه‌دهندگان در دسترس بود، اما اکنون به‌صورت متن‌باز منتشر شده است. Llama در ابعاد کوچک‌تری نیز ارائه می‌شود که برای استفاده و به‌کارگیری آن، به قدرت محاسباتی کمتری نیاز است. بزرگ‌ترین نسخه‌ی آن دارای ۶۵ میلیارد پارامتر است، از معماری ترنسفورمر استفاده می‌کند و با استفاده از منابع داده عمومی، ازجمله صفحات وب، آموزش دیده است.

کلود (Claude)

هوش مصنوعی Claude یک چت‌بات مبتنی‌بر هوش مصنوعی است که توسط شرکت Anthropic در سال ۲۰۲۲ معرفی شد. جدیدترین نسخه آن ۳.۰ Claude است که روی هوش مصنوعی قانون‌مدار (Constitutional AI) تمرکز دارد و خروجی را براساس مجموعه‌ای از اصول شکل می‌دهد تا مفید، بی‌خطر و دقیق باشد. این چت‌بات در زمینه‌های مختلف مانند تجزیه‌وتحلیل داده‌ها، پاسخ به سؤالات، حل مسائل ریاضی، کدنویسی، برنامه‌نویسی و موارد دیگر به کاربران سرویس می‌دهد. ازطریق سرویس جستجوی گوگل، به داده‌ها دسترسی دارد و پاسخ‌های جامع و به‌روز به سؤالات کاربران می‌دهد. دیگر مزیت آن، توانایی تولید فرمت‌های مختلف متن خلاقانه مثل شعر، کد، فیلمنامه، قطعات موسیقی، ایمیل، نامه و... است.

فالکون (Falcon 40B)

فالکون B۴۰ یک مدل مبتنی‌بر ترنسفورمر است که توسط مؤسسه نوآوری فناوری (Technology Innovation Institute) توسعه یافته و از ۴۰ میلیارد پارامتر برخوردار است. این مدل متن‌باز در دو نسخه کوچک‌تر با نام‌های فالکون ۱ Bو فالکون ۷ B (به ترتیب با یک میلیارد و هفت میلیارد پارامتر) نیز در دسترس است. شرکت آمازون مدل فالکن ۴۰ B را در سرویس SageMaker ارائه کرده است. این مدل همچنین به‌صورت رایگان در وبسایت GitHub در دسترس است.

جمع بندی: مدل زبانی بزرگ (LLM) چیست؟

مدل زبانی بزرگ با یک مجموعه داده عظیم تغذیه شده و به لطف یادگیری عمیق، توانایی تشخیص ارتباط بین ارکان متن و تولید محتوا را دارد. بهترین مدل‌های زبانی شامل جمینای گوگل، GPT-4o از OpenAI، مدل Claude و متا Llama است. کاربردهای مدل زبانی بسیار گسترده است و از تولید محتوا و برنامه‌نویسی تا تحلیل داده و چت‌بات طبیعی را شامل می‌شود. هرچند مدل‌های زبانی با مشکلاتی مثل محدودیت منابع مواجه هستند اما در آینده نزدیک ازنظر قدرت استدلال، شفافیت و دقت، بهبود قابل توجهی پیدا خواهند کرد.