



قطب کشوری شبکه سازی محیط های یادگیری و کد نویسی

آموزش سری یک مدل زبانی بزرگ مانند GPT

فصل اول - بخش اول

نویسنده کتاب: جناب آقای مسلم عسکری - عضو قطب کشوری کد نویسی



بهمن ماه ۱۴۰۳

فصل ۱

مدل های زبانی بزرگ

این فصل شامل موارد زیر است:

- توضیحات سطح بالا در مورد مفاهیم اساسی پشت مدل های زبانی بزرگ (LLM)
- بینش هایی در مورد معماری ترانسفورمر که از آن LLM های مشابه ChatGPT مشتق شده اند
- طرحی برای ساخت LLM از ابتدا

مدل های زبانی بزرگ (Large Language Models یا LLM)، نظیر آن هایی که در ChatGPT OpenAI به کار گرفته می شوند، مبتنی بر شبکه های عصبی عمیق هستند و در سال های اخیر تحولات چشمگیری در حوزه پردازش زبان طبیعی (Natural Language Processing یا NLP) ایجاد کرده اند. پیش از ظهور این مدل ها، روش های سنتی NLP عمدتاً در کاربردهای ساده ای مانند طبقه بندی هرزنامه های ایمیل یا تشخیص الگوهای مبتنی بر قوانین دست ساز موفق عمل می کردند. با این حال، این روش ها در وظایفی که نیازمند درک عمیق زبان، تولید متن یا تحلیل های پیچیده بودند، عملکرد محدودی داشتند. به عنوان مثال، مدل های قدیمی قادر به تولید یک ایمیل منسجم بر اساس فهرستی از کلمات کلیدی نبودند، در حالی که این وظیفه برای LLM های معاصر به سادگی قابل انجام است.

LLM ها توانایی قابل توجهی در پردازش، تولید و تفسیر زبان طبیعی از خود نشان می دهند. با این حال، باید تأکید کرد که اصطلاح «درک» در اینجا به معنای توانایی مدل در تولید متن های منسجم و مرتبط است، نه داشتن آگاهی یا درک انسانی. این مدل ها با بهره گیری از تکنیک های یادگیری عمیق (Deep Learning)، که زیرمجموعه ای از یادگیری ماشین و هوش مصنوعی است، بر روی حجم عظیمی از داده های متنی آموزش می بینند. این فرآیند آموزش گسترده به آن ها امکان می دهد تا اطلاعات زمینه ای و ظرافت های زبانی را به شکلی کارآمدتر از روش های پیشین درک کنند. در نتیجه، LLM ها عملکرد بهتری در طیف وسیعی از وظایف NLP، از جمله ترجمه ماشینی، تحلیل احساسات، پاسخ به سؤالات و تولید متن، از خود نشان داده اند.

یک تمایز کلیدی بین LLM های معاصر و مدل های سنتی NLP این است که مدل های قدیمی معمولاً برای کاربردهای خاصی طراحی می شدند، در حالی که LLM ها از قابلیت تعمیم پذیری بالاتری برخوردار هستند و می توانند در وظایف متنوع NLP عملکرد مطلوبی ارائه دهند. موفقیت این مدل ها تا حد زیادی مرهون معماری ترانسفورمر (Transformer) است که امکان یادگیری الگوهای پیچیده زبانی را فراهم می کند. علاوه بر این، حجم گسترده داده های آموزشی مورد استفاده در این مدل ها، به آن ها اجازه می دهد تا تفاوت های ظریف زبانی و زمینه های مختلف را به شکلی مؤثر رمزگذاری کنند. این ویژگی ها باعث شده اند که LLM ها به ابزاری قدرتمند در حوزه های مختلف NLP تبدیل شوند.

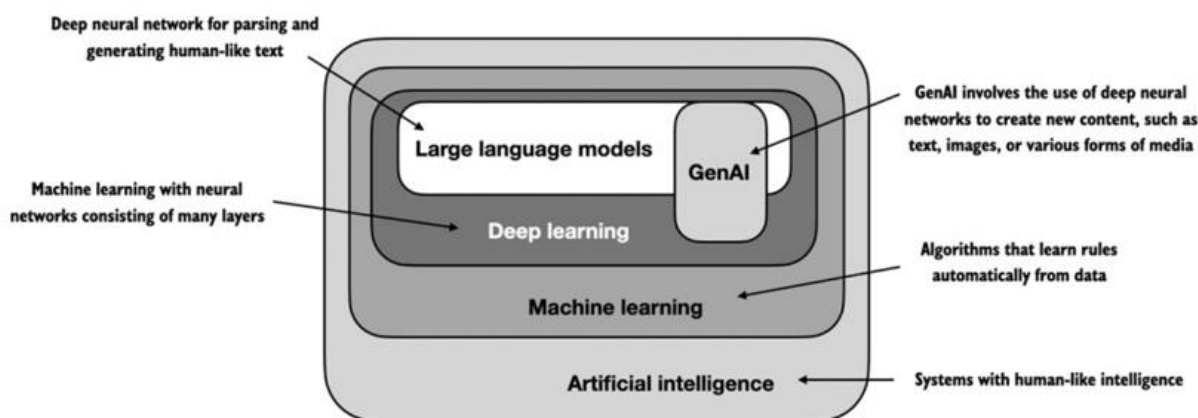
LLM چیست؟

مدل زبانی بزرگ (Large Language Model یا LLM) یک شبکه عصبی عمیق است که با هدف درک، تولید و پاسخ‌دهی به متن‌هایی شبیه به زبان انسان طراحی شده است. این مدل‌ها بر روی حجم گسترده‌ای از داده‌های متنی آموزش می‌بینند، که در برخی موارد شامل بخش‌های قابل توجهی از محتوای متنی موجود در اینترنت می‌شود. اصطلاح «بزرگ» در LLM به دو جنبه اشاره دارد: اولاً، به تعداد پارامترهای مدل، که ممکن است به ده‌ها یا حتی صدها میلیارد برسد؛ و ثانیاً، به حجم عظیم داده‌های آموزشی که برای آموزش مدل استفاده می‌شود. این پارامترها، که در واقع وزن‌های قابل تنظیم در شبکه عصبی هستند، در طول فرآیند آموزش بهینه می‌شوند تا بتوانند کلمه بعدی در یک دنباله متنی را پیش‌بینی کنند.

پیش‌بینی کلمه بعدی، اگرچه به ظاهر یک وظیفه ساده است، اما از ماهیت ترتیبی زبان بهره می‌برد تا مدل را در درک زمینه، ساختار و روابط موجود در متن آموزش دهد. این فرآیند ساده به طور شگفت‌انگیزی منجر به ایجاد مدل‌هایی با توانایی‌های چشمگیر شده است. در فصل‌های بعدی، مراحل آموزش این مدل‌ها را به صورت گام‌به‌گام مورد بررسی قرار خواهیم داد.

LLMها از معماری ترانسفورمر (Transformer) استفاده می‌کنند، که به آن‌ها امکان می‌دهد به بخش‌های مختلف ورودی توجه کنند و به همین دلیل در پردازش ظرافت‌ها و پیچیدگی‌های زبان انسانی بسیار مؤثر عمل می‌کنند. از آنجایی که این مدل‌ها قادر به تولید متن هستند، اغلب به عنوان نوعی از هوش مصنوعی مولد (Generative AI یا GenAI) نیز شناخته می‌شوند.

هوش مصنوعی مولد زیرمجموعه‌ای از حوزه گسترده‌تر هوش مصنوعی است، که شامل زیرشاخه‌هایی مانند یادگیری ماشین (Machine Learning) و یادگیری عمیق (Deep Learning) می‌شود. یادگیری عمیق، شاخه‌ای تخصصی از یادگیری ماشین است که بر استفاده از شبکه‌های عصبی چندلایه متمرکز است. هدف یادگیری ماشین و یادگیری عمیق، توسعه الگوریتم‌هایی است که می‌توانند از داده‌ها یاد بگیرند و بدون نیاز به برنامه‌ریزی صریح، پیش‌بینی‌ها یا تصمیم‌گیری‌های دقیقی انجام دهند. به عنوان مثال، فیلتر هرزنانه‌های ایمیل یک نمونه کاربردی از یادگیری ماشین است که در آن الگوریتم‌ها به طور خودکار از داده‌ها یاد می‌گیرند تا ایمیل‌های اسپم را تشخیص دهند.



شکل ۱-۱: همانطور که این تصویر سلسله مراتبی از رابطه بین زمینه‌های مختلف نشان می‌دهد، LLMها کاربرد خاصی از تکنیک‌های یادگیری عمیق را نشان می‌دهند و از توانایی آن‌ها برای پردازش و تولید متن، مانند انسان استفاده می‌کنند.

به جای اتکا به قوانین دستی برای شناسایی ایمیل‌های هرزنامه، الگوریتم‌های یادگیری ماشین از مجموعه‌ای از ایمیل‌های برچسب‌گذاری شده (هرزنامه و غیرهرزنامه) آموزش می‌بینند. با بهینه‌سازی پیش‌بینی‌های خود بر روی این داده‌های آموزشی، مدل قادر به یادگیری الگوها و ویژگی‌های متمایزکننده هرزنامه می‌شود و می‌تواند ایمیل‌های جدید را به درستی طبقه‌بندی کند.

یادگیری عمیق، زیرشاخه‌ای از یادگیری ماشین است که بر استفاده از شبکه‌های عصبی با سه یا چند لایه (شبکه‌های عصبی عمیق) متمرکز است. این شبکه‌ها توانایی مدل‌سازی الگوها و انتزاعات پیچیده در داده‌ها را دارند. در مقابل، یادگیری ماشین سنتی نیازمند استخراج دستی ویژگی‌ها است، به این معنا که متخصصان انسانی باید ویژگی‌های مرتبط را شناسایی و انتخاب کنند. برای مثال، در تشخیص هرزنامه، ممکن است ویژگی‌هایی مانند فراوانی کلمات خاص («جایزه»، «رایگان»)، تعداد علامت‌های تعجب یا وجود لینک‌های مشکوک به صورت دستی استخراج شوند. در حالی که در یادگیری عمیق، مدل به طور خودکار این ویژگی‌ها را از داده‌ها استخراج می‌کند.

با این حال، در هر دو رویکرد یادگیری ماشین سنتی و یادگیری عمیق، وجود داده‌های برچسب‌گذاری شده (مانند هرزنامه یا غیرهرزنامه) که توسط متخصصان یا کاربران جمع‌آوری شده‌اند، ضروری است. در بخش‌های آتی، به بررسی مشکلاتی که مدل‌های زبانی بزرگ (LLM) قادر به حل آن‌ها هستند، چالش‌های پیش‌روی این مدل‌ها و معماری کلی آن‌ها خواهیم پرداخت.

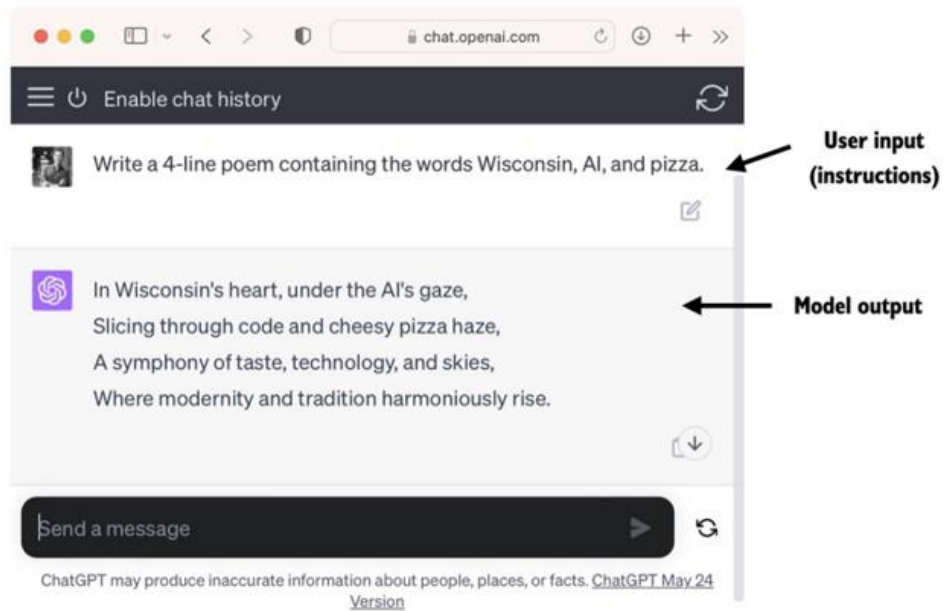
کاربردهای LLM

مدل‌های زبانی بزرگ (Large Language Models یا LLM) با توجه به قابلیت‌های پیشرفته‌شان در پردازش و درک داده‌های متنی بدون ساختار، کاربردهای گسترده‌ای در حوزه‌های مختلف دارند. امروزه، از این مدل‌ها برای وظایفی مانند ترجمه ماشینی، تولید متن، تحلیل احساسات، خلاصه‌سازی متن و بسیاری از کاربردهای دیگر استفاده می‌شود. اخیراً، LLMها حتی در تولید محتوای خلاقانه مانند داستان‌های تخیلی، مقالات و کدهای کامپیوتری نیز به کار گرفته شده‌اند.

LLMها قادرند چت‌بات‌های پیشرفته و دستیارهای مجازی مانند ChatGPT OpenAI یا Gemini گوگل (که پیش‌تر با نام Bard شناخته می‌شد) را تقویت کنند. این دستیارها می‌توانند به سؤالات کاربران پاسخ داده و حتی موتورهای جستجوی سنتی مانند گوگل یا بینگ را بهبود بخشند. علاوه بر این، LLMها می‌توانند در حوزه‌های تخصصی مانند پزشکی یا حقوق، به بازایی مؤثر دانش از حجم عظیمی از متون کمک کنند. این شامل بررسی اسناد، خلاصه‌سازی متن‌های طولانی و پاسخ به سؤالات فنی است.

به طور خلاصه، LLMها برای خودکارسازی تقریباً هر کاری که شامل تحلیل یا تولید متن باشد، ارزشمند هستند. کاربردهای آن‌ها تقریباً بی‌پایان است و با ادامه روند نوآوری، این مدل‌ها پتانسیل بازتعریف رابطه ما با فناوری را دارند و آن را تعاملی‌تر، شهودی‌تر و در دسترس‌تر می‌سازند.

در این کتاب، ما بر درک نحوه عملکرد LLMها از پایه تمرکز خواهیم کرد و یک مدل زبانی بزرگ را کدنویسی می‌کنیم که قادر به تولید متن باشد. همچنین با تکنیک‌هایی آشنا می‌شویم که به LLMها اجازه می‌دهند وظایفی مانند پاسخ به سؤالات، خلاصه‌سازی متن، ترجمه و غیره را انجام دهند. در واقع، در این کتاب، نحوه ساخت دستیارهای پیشرفته‌ای مانند ChatGPT را به صورت گام‌به‌گام فرا خواهیم گرفت.



شکل ۱-۳ رابط‌های کاربری LLM، ارتباط زبان طبیعی بین کاربران انسانی و سیستم‌های هوش مصنوعی را امکان‌پذیر می‌کند. شکل ۱-۴ اسکرین‌شات ChatGPT را نشان می‌دهد که شعری را مطلقاً با مشخصات کاربر می‌نویسد.