

گزارش فاز یک (سامانه کنترل ربات مبتنی بر مدل‌های زبانی بزرگ)

محمدمهری اسدی، امیرحسین سلیمانی

استاد راهنما: دکتر مازیار پالهنج

نسخه ۱،۰

آذر ۱۴۰۴

با سازمان‌دهی این یافته‌ها در حوزه‌های برنامه‌ریزی، استدلال، دستکاری، ادراک، معماری‌های کنترلی، یک نگاه یکپارچه از وضعیت کنونی هوش تجسم‌یافته^۲ مبتنی بر مدل زبانی بزرگ ارائه می‌دهیم. در پایان نیز چالش‌هایی مانند اینمنی، پایداری گراندینگ، دقت عددی و تعمیم در دنیای واقعی را بیان کرده و مسیرهای امیدبخش آینده برای ربات‌های خودکار، همسو با انسان را بحث می‌کنیم.

چکیده

مدل‌های زبانی بزرگ^۱ به یکی از عناصر تحول آفرین در رباتیک مدرن تبدیل شده‌اند. این مدل‌ها به ربات‌ها امکان می‌دهند که زبان طبیعی را تفسیر کنند، وظایف چندمرحله‌ای را برنامه‌ریزی کنند، بر اساس ورودی‌های چندوجهی استدلال کنند و مهارت‌های دستکاری^۳ اشیا را در محیط‌های پویا اجرا کنند. با حرکت رباتیک به سمت سطوح بالاتری از خودمختاری و هوش تجسم‌یافته، سامانه‌های مبتنی بر مدل‌های زبانی بزرگ به تدریج نقش هسته شناختی را بر عهده می‌گیرند؛ هسته‌ای که ادراک، تصمیم‌گیری، کنترل و تعامل انسان–ربات را یکپارچه می‌کند.

واژگان کلیدی

مدل‌های زبانی بزرگ، رباتیک، هوش تجسم‌یافته، درک زبان طبیعی، برنامه‌ریزی وظایف، دستکاری اشیا، استدلال چندوجهی، تولید مبتنی بر بازیابی اطلاعات، تعامل انسان–ربات، سامانه‌های کنترل ربات

۱. مقدمه

۱.۱. پیش‌زمینه و انگیزه

یکپارچگی مدل‌های زبانی بزرگ با رباتیک، نقطه عطفی در نحوه فهم، استدلال و تعامل ربات‌ها با جهان فیزیکی محسوب می‌شود. در حالی که رباتیک سنتی به برنامه‌نویسی ساخت‌یافته^۸، سیاست‌های کنترلی تخصصی و مدل‌های دقیق مهندسی شده متکی بود، پیشرفت‌های اخیر در پردازش زبان

با انگیزه پیشرفت‌های سریع این حوزه، این مقاله یک جمع‌بندی جامع از دستاوردهای اخیر در رباتیک مبتنی بر مدل‌های زبانی بزرگ ارائه می‌دهد. در این مقاله بررسی می‌کنیم که مدل‌های زبانی بزرگ چگونه برنامه‌ریزی وظایف مبتنی بر زبان طبیعی، تصمیم‌گیری تعاملی و ساختارهای برنامه‌نویسی مأذولات برای ربات‌ها را تقویت می‌کنند. همچنین به روش‌هایی مانند برنامه‌ریزی مبتنی بر ثابت‌سازی^۴ [۱] تحلیل نحوی مبتنی بر هسته‌شناسی^۵ [۲] و چارچوب‌های نظاممند آموزش ربات از طریق زبان انسان [۳] می‌پردازیم.

ما علاوه بر این، راهبردهای تنظیم تخصصی مدل‌های زبانی برای برنامه‌نویسی صنعتی ربات‌ها^۶ [۴] و روش‌های دستکاری مبتنی بر استدلال چندوجهی^۷ [۵] را بررسی می‌کنیم. همچنین پیشرفت‌های اخیر در سامانه‌های کنترل ربات مبتنی بر بازیابی اطلاعات^۶ [۶] و چارچوب‌های ساخت‌یافته مبتنی بر ROS^۷ [۷] را مرور می‌کنیم.

⁵ ontology-based action parsing

⁶ retrieval-augmented robotic

⁷ embodied intelligence

⁸ structured

¹ Large language model

² Multimodal inputs

³ manipulation

⁴ Grounded planning

نسل جدید مدل‌های زبانی بزرگ این قابلیت را به صورت چشم‌گیری گسترش داده است. مدل‌های مدرن با درک معنایی عمیق، نیت ضمنی، وابستگی‌های زمانی و منطق موردنیاز برای انجام وظیفه را شناسایی می‌کنند. برای مثال پژوهش‌ها درباره اجرای وظایف بر اساس زبان طبیعی نشان داده‌اند که سامانه‌ها می‌توانند عبارت «من گرسنه‌ام» را با استناد به دانش عمومی خود به وظیفه «غذا تهیه کن» تبدیل کنند. [3] این توانایی، فاصله زیادی با سامانه‌های قدیمی مبتنی بر قواعد دارد و نشان‌دهنده جهت‌گیری به سمت تفسیر سازگار وابسته به زمینه است.

مدل‌های زبانی بزرگ همچنین در رفع ابهام‌های زبانی نقش مهمی دارند؛ زمانی که دستور ناقص است، مدل می‌تواند از کاربر سوال پرسد. این ویژگی با چارچوب‌های گفت‌وگو محور در تعامل انسان–ربات سازگار است، جایی که زبان نه تنها ابزار فرمان‌دادن، بلکه ابزاری برای مذاکره، اصلاح و روشن‌سازی است. [7] به این ترتیب، زبان طبیعی به ستری برای تعامل دوطرفه تبدیل می‌شود؛ بستری که هم برای صدور دستور و هم برای تصمیم‌گیری مشترک کاربرد دارد.

۳.۱. پیشرفت‌های برنامه‌ریزی و استدلال مبتنی بر مدل‌های زبانی بزرگ

مرور ادبیات نشان می‌دهد که برنامه‌ریزی وظایف، یکی از حوزه‌هایی است که مدل‌های زبانی بزرگ بیشترین تأثیر را در آن داشته‌اند. پژوهش‌ها ثابت کرده‌اند که مدل‌های زبانی بزرگ قادرند توالی‌های دقیق و منظم از اقدامات تولید کنند، از دانش عمومی بهره بگیرند و حتی با برنامه‌ریزهای کلاسیک همکاری کنند تا دقت و قابلیت اتکای برنامه‌ریزی افزایش یابد. [1] این هم‌افزایی به ربات اجازه می‌دهد که دستورهای بلندمدت را به گام‌های ساخت‌یافته تقسیم کند، پیش‌شرط‌ها¹³ و روابط فضایی¹⁴ را تحلیل کند، از مکانیزم‌های ثابت‌سازی احتمالاتی استفاده کند و از دانش پیشینی برای هدایت الگوریتم‌های جست‌وجو بهره ببرد.

علاوه بر این، چارچوب‌هایی مانند LLM+P و SayCan تلفیقی از راهنمایی معنایی مدل‌های زبانی بزرگ و سیاست‌های تقویتی را ارائه می‌دهند. در این رویکرد، مدل‌های زبانی بزرگ هدف‌های سطح بالا را مشخص می‌کند و یادگیری تقویتی عملی‌بودن آنها را تأیید می‌کند. [1] سامانه‌های دیگر از ساختارهایی مانند اشاره‌های خط¹⁵، بازخورد چندوجهی و مدل‌های

طبیعی، امکان چارچوب‌هایی را فراهم کرده است که در آنها ربات می‌تواند مقصود انسان را مستقیماً از طریق زبان طبیعی درک کند. این تغییر نشان‌دهنده گذار به‌سمت «هوش تجسم‌یافته» است. جایی که ربات‌ها با استفاده از بازنمایی‌های شهودی و غنی‌شده از نظر شناختی یاد می‌گیرند، عمل می‌کنند و همکاری می‌کنند. [8]

در گذشته، سامانه‌های رباتیکی نیازمند برنامه‌نویسی صریح و دانش دامنه‌ای گسترش داده بودند. اما مدل‌های زبانی بزرگ همچون موتورهای استدلالی عمومی⁹ عمل می‌کنند؛ که قادرند معنای زمینه‌ای¹⁰ را استخراج کنند، برنامه انجام وظایف را تولید کنند و این برنامه‌ها را به دستورهای قابل اجرا تبدیل کنند. این قابلیت امکان می‌دهد وظایف پیچیده از نوبری¹¹ تا دستکاری– به صورت مکالمه‌ای تعریف شوند، نه با کدنویسی تخصصی یا الگوهای نمادین پیچیده. برای مثال، سامانه‌های برنامه‌ریزی مبتنی بر مدل‌های زبانی بزرگ قادرند حتی بدون ورودی بصری، دستورهای زبانی پیچیده را به توالی اقدامات تبدیل کنند و هنگامی که ورودی چندوجهی در اختیارشان قرار گیرد، عملکرد بسیار بهتری نشان می‌دهند. [1]

مدل‌های چندوجهی نیز با ترکیب ورودی‌های بصری، زبانی و زمینه‌ای، بازنمایی‌های واقعیت محور تولید می‌کنند که پل میان ارتباط انسانی و کنترل ربات است. این مدل‌ها توانایی استدلال قابل اعتماد در محیط‌های آگاهی فضایی را فراهم می‌کنند، که برای عملکرد قابل اعتماد در محیط‌های واقعی ضروری است. چنین توانایی‌هایی نشان می‌دهد که استدلال مبتنی بر مدل‌های زبانی بزرگ عنصر مرکزی در حرکت به‌سوی خودمختاری، سازگاری و هم‌راستایی بیشتر با انسان است.

۲.۱. از زبان تا عمل: نقش درک زبان طبیعی

یکی از کاردهای اساسی مدل‌های زبانی بزرگ در رباتیک، تبدیل زبان انسان به بازنمایی¹²‌های نمادین، ساخت‌یافته یا قابل اجرا است. پژوهش‌های اولیه تمرکز خود را بر خط لوله‌های تحلیل زبان طبیعی قرار داده بودند. برای مثال استفاده از سامانه‌های مبتنی بر هسته‌شناسی که با تحلیل وابستگی‌های نحوی، افعال، اعضای بدن ربات و محدودیت‌ها را استخراج می‌کردد. [2] این سامانه‌ها نشان دادند که ربات می‌تواند به شکل هوشمندانه به دستورهای انسانی واکنش نشان دهد و در عین حال، سازگاری اقدامات پیشنهادی را با وضعیت فعلی خود بررسی کند. برای نمونه رد کردن دستوری که تعادل ربات را مختل می‌کند.

¹³ preconditions

¹⁴ spatial relationships

¹⁵ error cues

⁹ generalized reasoning engines

¹⁰ contextual meaning

¹¹ Navigation

¹² representations

۵.۱ معماری‌های ماژولار و نسل مبتنی بر بازیابی (RAG)

با وجود افزایش توانایی مدل‌های زبانی بزرگ، این مدل‌ها همچنان ممکن است دچار توهمندی و استدلال نادرست شوند. در حوزه رباتیک، چنین خطاهایی می‌توانند خطرناک باشند. از همین رو، تولید مبتنی بر بازیابی **Retrieval-augmented generation (RAG)** به یک رویکرد مهم برای ثابت سازی و اعتبارسنجی تبدیل شده است.

سامانه **ARRC** نمونه برجسته‌ای از این راهبرد است. در این سامانه، پایگاه دانشی شامل الگوهای حرکتی، راهبردهای اینمنی و قالب‌های وظیفه به شکل برداری ذخیره شده و در زمان برنامه‌ریزی، بخش مرطیط از آن بازیابی می‌شود. [6] سپس مدل‌های زبانی بزرگ با اتکا بر این داشت معتبر، برنامه کنشی ساختیافته را تولید می‌کند و در نهایت یک مرحله ارزیابی اینمنی، برنامه را کنترل می‌کند.

به همین ترتیب، سامانه‌های ماژولار مبتنی **ProgramPort** و **MetaMorph** از مدل‌های زبانی بزرگ برای تبدیل دستورات زبانی به ماژول‌های برنامه‌نویسی یا سیاست‌های سازگار با شکل‌بندن‌های مختلف ربات بهره می‌برند. [1] این رویکردها قابلیت مقایسه‌پذیری، تعمیم و ترکیب‌پذیری را افزایش می‌دهند و به ربات‌ها اجازه می‌دهند مهارت‌ها را ترکیب کرده، به ابزارهای جدید سازگار شوند یا در محیط‌های متنوع، وظایف پیچیده را انجام دهند.

در نهایت می‌توان گفت معماری‌های مبتنی بر مدل‌های زبانی بزرگ نه تنها به عنوان مفسر زبان، بلکه به عنوان چارچوب‌های سازمان‌دهنده رفتار ربات عمل می‌کنند.

۶. تعامل انسان–ربات و یادگیری تعاملی

مدل‌های زبانی بزرگ نقش اساسی در گسترش ظرفیت‌های تعامل انسان–ربات ایفا می‌کنند. برای نمونه، سامانه **ROS-LLM** نشان می‌دهد که ربات‌ها می‌توانند وارد گفتگوهای چندمرحله‌ای شوند، بازخورد انسانی را دریافت کنند، رفتارهای سلسله‌مراتبی اجرا کنند و در طول زمان بهبود یابند. [7] در این چارچوب‌ها، ورودی‌های چندوجهی مبتند متن، تصویر یا داده‌های حسگر به ربات کمک می‌کند تا ابهام را کاهش دهد، خط را اصلاح کند و به تغییرات محیط سازگار شود.

از سوی دیگر، مدل‌های مبتنی بر «عامل‌های مولد»^{۱۸} امکان ذخیره حافظه، یادگیری از تجربه و حتی الگوگیری از سبک‌های تعامل انسانی را فراهم

استدلال–عمل مبتنی **ReAct** بهره می‌برند. این روش‌ها موجب ثبات بیشتر ربات در محیط‌های پویا می‌شوند.

استدلال پیچیده تنها به برنامه‌ریزی محدود نمی‌شود و حوزه تصمیم‌گیری و سازگاری را نیز دربرمی‌گیرد. به عنوان نمونه **LM-Nav** ترکیبی از زبان، بینایی و اطلاعات ناوبری را برای دنبال کردن دستورهای سطح بالا به کار می‌گیرد. بدون آنکه نیاز به داده‌های پیمایش نشانه‌گذاری شده^{۱۹} باشد، [1] در کنار آن، مدل‌های آگاهی از عدم قطعیت مبتنی **KnowNo** به ربات کمک می‌کنند از تصمیمات اشتباه یا خطرناک اجتناب کند و قابلیت اطمینان خود را در وظایف طولانی افزایش دهد. [1] این مجموعه از روش‌ها بینانگ آن است که مدل‌های زبانی بزرگ از یک «پردازشگر زبان» فراتر رفته و به «عامل‌های استدلالی فعال» تبدیل شده‌اند که در هسته معماري کنترل ربات قرار می‌گيرند.

۴. دستکاری، ادراک و مهارت‌های تجسم‌یافته تقویت شده با مدل‌های زبانی بزرگ

در حالی که برنامه‌ریزی وظایف مبتنی بر زبان یکی از برجسته‌ترین پیشرفت‌های اخیر است، دستکاری فیزیکی و کنترل تجسم‌یافته همچنان از چالش‌برانگیزترین بخش‌های رباتیک محسوب می‌شود. روش‌های سنتی برای دستکاری اجسام به دقت عددی، مدل‌سازی هندسی و خط لوله‌های ادراکی بسیار تنظیم‌شده مตکی هستند. اما مدل‌های زبانی بزرگ بعد جدیدی را وارد این حوزه کرده‌اند.

سامانه‌هایی مبتنی **VIMA** و **LLM-GROP** نشان داده‌اند که سرنخ‌های زبانی قادرند استدلال فضایی ربات را غنی کنند. این سامانه‌ها می‌توانند تنها با تکیه بر نشانه‌های سطح بالا، وظایفی نظیر جای‌گذاری بدون برخورد یا تقلید چندوجهی را انجام دهند. [1] چارچوب‌های چندوجهی با یکپارچه‌سازی بینایی، زبان و حرکت، این امکان را می‌دهند که ربات‌ها بدون نیاز به داده‌های اختصاصی، میان دسته‌های مختلفی از وظایف تعمیم‌پذیری پیدا کنند.

از سوی دیگر، ورود روش‌هایی مبتنی **RT-Grasp** که بر «استدلال تنظیم‌شده»^{۲۰} تکیه دارند، یکی از موانع اصلی در دستکاری، یعنی دقت عددی را هدف قرار داده است. **RT-Grasp** با ترکیب خروجی‌های استدلالی مدل‌های زبانی بزرگ و قالب‌های پیش‌بینی عددی ساخت‌یافته، توضیحات معنایی مدل را برای اصلاح پارامترهای دقیق گرفتن اشیا به کار می‌گیرد. [5] این کار پلی میان فهم زبانی–مفهومی و کنترل هندسی ایجاد می‌کند، پلی که پیش از این بیشتر بر دوش روش‌های تحلیلی یا یادگیری مبتنی بر داده بود.

¹⁸ generative agents

¹⁶ trajectory annotations

¹⁷ reasoning-tuned

از سوی دیگر، مدل‌های زبانی بزرگ در گفت‌و‌گوی تعاملی برای اصلاح فرمان‌ها نیز مؤثرند. مدل قادر است هنگام نیاز به اطلاعات بیشتر، سؤال پرسید تا دستور نهایی دقیق و قابل اجرا شود. چنین تعاملی با رویکردهای مدرن تعامل انسان–ربات سازگار است. جایی که زبان علاوه بر نقش فرماندهی، نقش ابزار مذاکره و رفع ابهام را نیز دارد.^[7]

در مجموع، **NLU** مبتنی بر مدل‌های زبانی بزرگ زیربنای شناختی سامانه‌های رباتیکی را شکل می‌دهد و آنها را قادر می‌سازد تا دستورهای پیچیده انسانی را تفسیر کرده و مطابق آن عمل کنند.

۳. برنامه‌ریزی وظایف با استفاده از مدل‌های زبانی بزرگ

برنامه‌ریزی وظایف یکی از حوزه‌هایی است که مدل‌های زبانی بزرگ بیشترین تحول را در آن ایجاد کرده‌اند. روش‌های کلاسیک برنامه‌ریزی ربات بر برنامه‌ریزی‌های نمادین، الگوریتم‌های جست‌وجو یا یادگیری تقویتی مبتنی بر پاداش تکیه داشتند. این روش‌ها در محیط‌های پویا، مبهم یا دارای پیچیدگی معنایی، با محدودیت‌های جدی روبرو بودند. اما مدل‌های زبانی بزرگ این محدودیت را برطرف کرده و امکان برنامه‌ریزی سطح بالا با درک زمینه‌ای و استدلال معنایی را فراهم کرده‌اند.

پژوهش‌ها نشان می‌دهد که مدل‌های زبانی بزرگ قادرند از روی دستورهای زبانی، حتی بدون مشاهده کامل محیط، توالی اقدامات بلندمدت تولید کنند.^[1] یکی از نقاط قوت مدل‌های زبانی بزرگ، توانایی استخراج دانش عمومی و استفاده از آن در تدوین مراحل منطقی وظایف است. برای مثال، مدل به‌طور طبیعی می‌داند که برای تهیه یک نوشیدنی، ابتدا باید یک ظرف پیدا شده و سپس پر شود؛ حتی اگر این مراحل صریحاً در دستور ذکر نشده باشند.

در رویکردهای ترکیبی مانند **LLM+P** و **SayCan** مدل‌های زبانی با برنامه‌ریزهای رسمی مبتنی بر **PDDL** ادغام می‌شوند. در این مدل‌ها، مدل زبانی بزرگ وظایف سطح بالا را تفسیر کرده و برنامه‌ریزهای سنتی تضمین می‌کنند که این عملیات با محدودیت‌های جهان واقعی سازگار باشد.^[1] علاوه بر این، روش‌هایی مانند رمزگشایی‌های واقعیت محور^{۱۹} با هم‌تازی کردن خروجی مدل زبانی بزرگ با مدل‌های فیزیکی محیط، قابلیت اتکا و سازگاری برنامه‌ریزی را افزایش می‌دهند. این هم‌تازی احتمال تولید اقدامات غیرممکن **LM** یا ناسازگار را کاهش می‌دهد.^[1] در حوزه ناوبری، سیستم‌هایی مانند **Nav** با ترکیب مدل‌های پیش‌آموزش‌دیده بینایی، مدل‌های زبان و استدلال فضایی، به ربات امکان می‌دهند که بدون نیاز به داده‌های پیامایش برچسب‌خورده، مسیرهای پیچیده را دنبال کند. (تصویر ۱) مجموع این پیشرفت‌ها نشان می‌دهد که مدل‌های زبانی بزرگ پایه یک چارچوب برنامه‌ریزی انعطاف‌پذیر، قدرتمند و قابل تعمیم را تشکیل می‌دهند که می‌تواند

می‌کنند. نتیجه این پیشرفت‌ها، حرکت به سمت ربات‌هایی است که تعامل طبیعی‌تر، سازگارتر و مشارکتی تر دارند.

۲. درک زبان طبیعی برای سامانه‌های رباتیکی

درک زبان طبیعی (NLU) سنگ‌بنای توانایی ربات‌های مبتنی بر مدل‌های زبانی بزرگ برای تفسیر نیت انسان و تبدیل آن به رفتارهای قابل اجرا است. در ادبیات پژوهش، **NLU** مجموعه‌ای از فرایندها را شامل می‌شود. تحلیل ورودی زبانی، استخراج ساختار معنایی، شناسایی عناصر قابل اقدام، رفع ابهام و تبدیل زبان طبیعی به فرمان‌های اجرایی.

رویکردهای اولیه بهشت بر خط لوله‌های زبانی و بازنمایی‌های نمادین تکیه داشتند. برای مثال، چارچوب‌های مبتنی بر هستی‌شناسی با استفاده از وابستگی‌های نحوی، افعال، اعضای بدن و محدودیت‌های حرکتی را استخراج کرده و از آنها برای ایجاد فرمان‌های ربات استفاده می‌کردند.^[2]

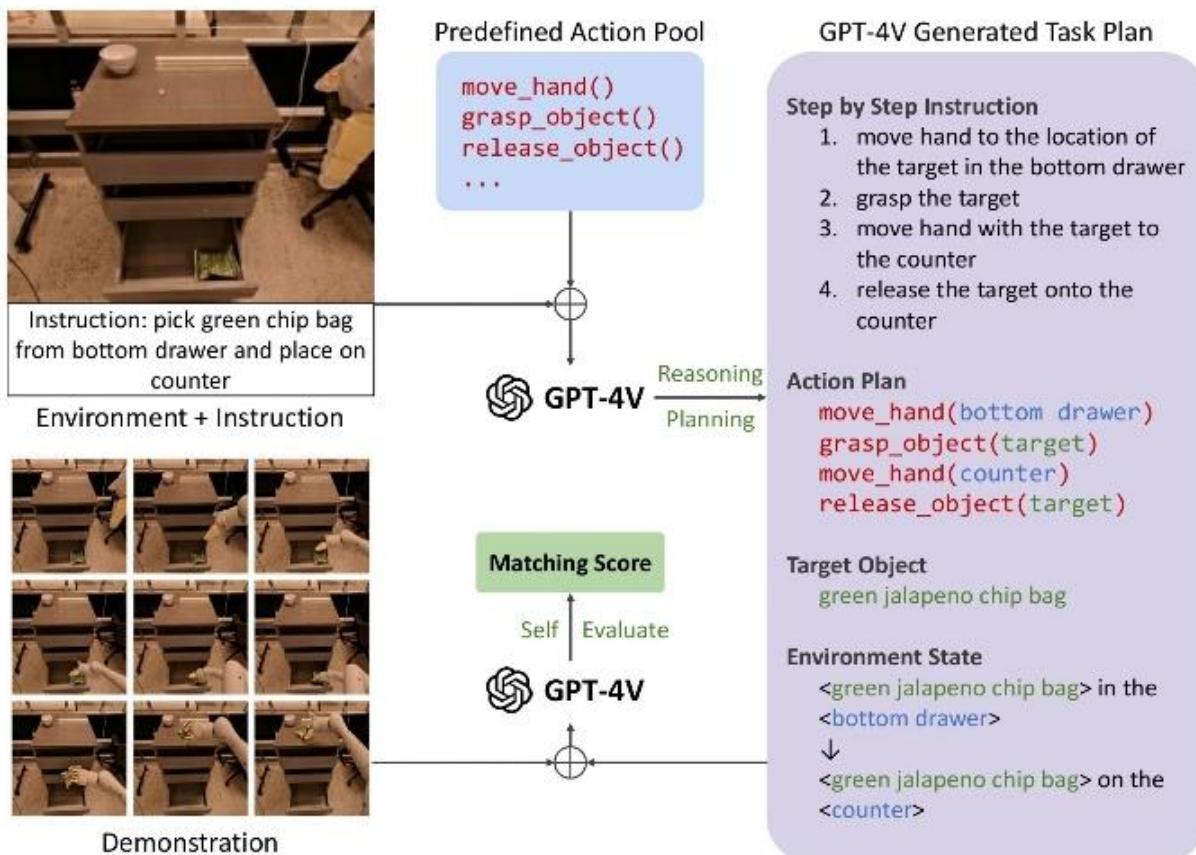
این سامانه‌ها قادر بودند در صورت ناقص بودن اطلاعات (برای نمونه، «دستت را بالا ببر» بدون ذکر راست یا چپ)، از کاربر سؤال پرسیده و مسیر صحیح را مشخص کنند. همچنین این چارچوب‌ها برای بررسی سازگاری فرمان با وضعیت فیزیکی فعلی ربات ضروری بودند.

با ظهور مدل‌های زبانی بزرگ مدرن، این قابلیتها گسترش چشمگیری یافته است. مدل‌های زبانی بزرگ با تکیه بر بازنمایی‌های معنایی قوی و آموزش گسترشده روی داده‌های متنی، قادرند روابط پنهان بین دستورها، محیط و نیت ضمنی را استخراج کنند. به عنوان نمونه پژوهش‌ها نشان می‌دهد که مدل‌های زبانی بزرگ می‌توانند بدون ورودی بصری نیز یک دستور را به توالی منسجم اقدامات تبدیل کنند و هنگامی که ورودی چندوجهی اضافه شود، دقت و ثابت‌سازی به‌طور قابل توجهی افزایش می‌یابد.^[1] پیشرفت در سامانه‌های اجرای وظایف از طریق زبان نیز نشان می‌دهد که ربات‌ها توانایی درک مقصود غیرمستقیم را پیدا کرده‌اند. برای مثال می‌توانند دستور ضمنی را از جملاتی مانند «هوا سرد است» استنتاج کنند.^[3] این رفتار نشان‌دهنده گذار از تطبیق الگوهای دستوری به استدلال مبتنی بر دانش عمومی و زمینه‌محور است.

در حوزه ایمنی، روش **KnowNo** با تخمین عدم قطعیت، مانع از انجام اقداماتی می‌شود که مدل نسبت به آنها اعتماد کافی ندارد. [1] این ویژگی در وظایف چندمرحله‌ای و حساس، اهمیت ویژه‌ای دارد.

در زمینه چندوجهی، مدل **LM-Nav** نشان می‌دهد که چگونه می‌توان زبان، ادراک بصری و مدل‌های ناوبری را برای تصمیم‌گیری هماهنگ ادغام کرد. برخی پژوهش‌ها نیز بر تصمیم‌گیری میان چند عامل تمرکز دارند؛ جایی که مدل‌های زبانی بزرگ قادرند تعامل میان چند ربات یا تعامل میان ربات و انسان را هماهنگ کنند. این توانایی برای محیط‌های مشارکتی ضروری است.

تصویر 1. چارچوب روش پیشنهادی برای برنامه‌ریزی وظایف تجسسی با GPT-4V فریم ابتدایی و متن دستورها به مدل داده می‌شود. GPT-4V دستور را به مجموعه‌ای از گام‌های عملیاتی تجزیه کرده و از میان یک مخزن از اعمال، نمایه‌های مناسب را بر می‌گزیند. هم‌زمان، شیء هدف و تغییرات محیطی قبل و بعد از انجام دستور در تصاویر تحلیل می‌شود. در پایان، GPT-4V طرح تولیدشده را با طرح مرجع مقایسه کرده و امتیاز می‌دهد.



۵. دستکاری و تعامل فیزیکی تقویت شده با LLMها

وظایف دستکاری -مانند گرفتن، جابه‌جایی و چیدمان اشیا- به دلیل نیاز به دقت هندسی، کنترل پیوسته و استدلال فضایی از چالش‌برانگیزترین حوزه‌های رباتیک محسوب می‌شوند. روش‌های سنتی برای این نوع وظایف معمولاً بر بهینه‌سازی عددی، برآورد موقعیت اشیا و سیاست‌های کنترلی

وظایف پیچیده را تحلیل کند، نیت انسان را استخراج کند و برنامه‌های قابل اعتماد برای محیط‌های پویا تولید کند.

۴. استدلال پیچیده و تصمیم‌گیری

پژوهش‌های مختلف نشان می‌دهد که مدل‌های زبانی بزرگ می‌توانند از حافظه معنایی پهله ببرند، میان وظایف مختلف تعمیم پیدا کنند و با استفاده از قالب‌های استدلالی ساخت‌یافته، تصمیم‌های منطقی اتخاذ کنند. برای مثال، استفاده از اشاره‌های خطابی ربات کمک می‌کند که برنامه تولیدشده را اصلاح و بهینه کند. [1] مدل‌هایی مانند **ReAct** استدلال و عمل را در یک چرخه

تصویر 1. چارچوب روش پیشنهادی برای برنامه‌ریزی وظایف تجسسی با GPT-4V.

فریم ابتدایی و متن دستورها به مدل داده می‌شود. GPT-4V دستور را به مجموعه‌ای از گام‌های عملیاتی تجزیه کرده و از میان یک مخزن از اعمال، نمایه‌های مناسب را بر می‌گزیند. هم‌زمان، شیء هدف و تغییرات محیطی قبل و بعد از انجام دستور در تصاویر تحلیل می‌شود. در پایان، GPT-4V طرح تولیدشده را با طرح مرجع مقایسه کرده و امتیاز می‌دهد.

ترکیب می‌کند؛ در نتیجه ربات می‌تواند حالت درونی خود را حفظ کند، پاسخ‌های نادرست را فیلتر کند و از تصمیم‌های متناقض جلوگیری کند.

چارچوب‌های تعامل محور مانند **LILAC** به کاربران اجازه می‌دهند تنها با زبان طبیعی و (در صورت نیاز) تصاویر، مسیر ربات یا پارامترهای کنترل را تنظیم کنند. [1] این روش‌ها باعث می‌شوند ربات‌ها سریع‌تر و طبیعی‌تر به خواسته‌های کاربر سازگار شوند.

سامانه **ROS-LLM** این مفهوم را گسترش می‌دهد و از مدل زبانی بزرگ برای به کارگیری بازخورد انسان در چرخه برنامه‌ریزی و اجرا استفاده می‌کند. در این سامانه، ربات می‌تواند:

- خطاهای اجرا را تشخیص دهد.
- پیشنهادهای اصلاحی کاربر را تفسیر کند.
- برنامه را بازسازی یا بهبود دهد. [7]

افزون بر این، عامل‌های مولد امکان ذخیره تجربه، شکل‌گیری حافظه درازمدت و سازگاری رفتاری را فراهم می‌کنند. این توانایی‌ها ربات‌ها را به موجوداتی یادگیرنده، سازگار و تعاملی‌تر تبدیل می‌کند. در مجموع، رویکردهای تعاملی مدل زبانی بزرگ محور باعث می‌شوند ربات‌ها بتوانند در زمان واقعی یاد بگیرند، از بازخورد انسان بهره ببرند، در محیط‌های ناپایدار یا ناشناخته عملکرد باثبات‌تری داشته باشند.

۷. رویکردهای مازوُلار برای هوش رباتیکی مقیاس‌پذیر

معماری مازوُلار یکی از اصول مهم طراحی در سامانه‌های رباتیکی است. مدل‌های زبانی بزرگ این رویکرد را به سطح جدیدی رسانده‌اند. در این معماری، یک وظیفه پیچیده به مجموعه‌ای از مازوُل‌های قابل استفاده مجدد تقسیم می‌شود. مانند مهارت‌ها، سیاست‌های کنترلی، برنامه‌ریزها، مازوُل‌های ادراک و درخت‌های رفتار.

مدل‌های زبانی بزرگ با تبدیل توصیف‌های زبانی به این مازوُل‌ها، امکان ساخت سازه‌های رفتاری پیچیده را بدون نیاز به برنامه‌نویسی دستی فراهم می‌کنند.

نمونه‌ای از این رویکرد، **PROGRAMPORT** است که زبان طبیعی را به مازوُل‌های برنامه‌ای برای دستکاری اشیا تبدیل می‌کند. [1] در این سامانه، به جای تولید مستقیم کد، ساختار معنایی دستورها به بلوک‌های عصبی قابل استفاده مجدد تبدیل می‌شود که قابلیت تعمیم در محیط‌ها و وظایف جدید را دارند.

رویکرد دیگر، استفاده از **MetaMorph** است؛ سیستمی که با مدل‌های ترنسفورمر، سیاست‌های کنترل سازگار با شکل‌بندی‌های مختلف ربات را یاد می‌گیرد. [1] این امر باعث می‌شود ربات بتواند بدون طراحی مجدد، به سخت‌افزارهای جدید سازگار شود.

داده‌محور تکیه دارند. اما مدل‌های زبانی بزرگ امکان یک رویکرد مفهومی-معنایی را برای هدایت این فرآیند فراهم می‌کنند.

برای مثال، پژوهش‌هایی مانند **LLM-GROP** نشان می‌دهند که مدل‌های زبانی قادرند سرنخ‌های معنایی را استخراج کرده و به ربات کمک کنند تا راهکارهای منطقی برای دستکاری انتخاب کند. مانند این که اشیای شکننده را زیر اجسام سنگین قرار ندهد. [1] چارچوب‌های چندوجهی مانند **VIMA** و **TIP** نیز با ترکیب متن و تصویر، به ربات‌ها امکان می‌دهند دستکاری را از طریق نشانه‌های زبانی و بصری یاد بگیرند و مهارت‌ها را به مواردی تعمیم دهنند که پیش‌تر ندیده‌اند.

پیشرفت مهم دیگری در این حوزه، روش **RT-Grasp** است. در این رویکرد، قبل از آنکه مدل خروجی عددی دقیق (مثل زاویه گرفتن یا موقعیت گیره) تولید کند، ابتدا از یک مرحله استدلال زبانی ساخت‌یافته استفاده می‌شود. این امر باعث می‌شود مدل ابتدا ماهیت شیء (مثال «فنجان»، «عینک»، «توب») و اصول صحیح گرفتن آن را توضیح دهد، سپس پارامترهای دقیق را محاسبه کند. [5] این ساختار پل مهمی میان استدلال سطح بالا و کنترل فیزیکی دقیق ایجاد می‌کند؛ شکافی که سال‌ها چالشی اساسی در رباتیک بود.

روش‌های دیگری مانند **R3M** و **LIV** نشان داده‌اند که آموزش مدل‌های بینایی-زبانی با داده‌های ویدئویی انسان، بازنمایی‌هایی غنی از قابلیت‌ها، روابط اشیا و ویژگی‌های محیط ایجاد می‌کند. این باعث می‌شود ربات‌ها با داده‌های بسیار کمتر، مهارت‌های دستکاری را یاد بگیرند. [1]

در سمت دیگر، سامانه‌هایی مانند **VoxPoser** به ربات امکان می‌دهند تا محدودیت‌های دستکاری را مستقیماً از متن استخراج کرده و رفتار خود را به‌طور پویا تنظیم کنند.

مجموع این پیشرفت‌ها نشان می‌دهد که دستکاری رباتیک در حال حرکت از یک مسئله صرفاً هندسی به یک مسئله ترکیبی از معنا، ادراک و استدلال چندوجهی است. رویکردی که به مهارت‌هایی نزدیک به انسان منجر می‌شود.

۶. راهبردهای تعاملی و ادغام بازخورد انسانی

ربات‌ها در محیط‌های واقعی باید عملکرد خود را بر اساس بازخورد انسان، خطاهای گذشته یا تغییرات محیط اصلاح کنند. مدل‌های زبانی بزرگ ابزارهای بسیار قدرتمندی برای هدایت این تعامل فراهم کرده‌اند.

سامانه **TEXT2REWARD** با تبدیل بازخورد زبانی انسان به کدهای پاداش، به عامل‌های یادگیری تقویتی امکان می‌دهد اصلاحات انسانی را در مسیر یادگیری خود اعمال کنند. [1] رویکرد **InstructRL** نیز از مدل زبانی بزرگ برای تولید سیاست‌های ابتدایی استفاده می‌کند و سپس به عامل اجازه می‌دهد با تکیه بر بازخورد انسان، رفتار خود را اصلاح کند.

مدل‌های زبانی بزرگ و مدل‌های بینایی-زبانی^{۲۰} در این زمینه پیشرفت‌های قابل توجهی ایجاد کرده‌اند.

سیستم **PaLM-E** نمونه‌ای مهم از ادغام ورویدی‌های چندوجهی است. در این مدل، تصاویر، بردارهای حالت و متن به صورت مشترک پردازش می‌شوند و مدل قادر است روی مجموعه گسترده‌ای از وظایف بینایی و دستکاری عمل کند. [8] این یکپارچگی به ربات اجازه می‌دهد به پرسش‌هایی مانند «داخل کشو چه چیزی است؟» یا «فنجان سبز کجاست؟» پاسخ مناسب بدهد.

سامانه **LM-Nav** نیز با ترکیب جایگذاری‌های CLIP و نقشه‌برداری مبتنی بر **ViNG**، به ربات امکان می‌دهد که صرفاً بر اساس زبان طبیعی و بدون نیاز به داده پیمایش برچسب‌خورده، به نقاط خاصی در محیط ناوبری کند. [6] این ترکیب، هم ادراک معنایی و هم استدلال فضایی را به شکلی یکپارچه فراهم می‌کند.

در حوزه دستکاری، **ARRC** با استفاده از **AprilTags+** عمق موقعیت سه‌بعدی دقیق اشیا را استخراج می‌کند و این اطلاعات برای اجرای صحیح وظایف **pick & place** ضروری است. [6]

به‌طور مشابه، مدل‌های چندوجهی مانند **VIMA** و **TIP** از زبان و تصویر برای تعريف محدودیت‌ها و اهداف دستکاری استفاده می‌کنند و به ربات امکان می‌دهند روابطی مثل «بلوک را سمت چپ مخروط قرار بده» را بفهمند.

این بازنمایی‌ها موجب افزایش کارایی نمونه^{۲۱} و تعمیم بدون آموزش اضافی در ربات‌های واقعی می‌شود.

در نهایت، می‌توان گفت ادراک مبتنی بر مدل‌های زبانی بزرگ تنها هندسه محیط را توصیف نمی‌کند، بلکه معنای صحنه را نیز استخراج می‌کند. توانایی‌ای که برای تعامل در محیط‌های واقعی، پویا و غیرساخت‌یافته^{۲۲} ضروری است.

۱۰. کنترل و اجرا در رباتیک مبتنی بر مدل زبانی بزرگ

کنترل - یعنی ترجمه هدف به حرکت فیزیکی - جایی است که استدلال سطح‌بالای مدل زبانی بزرگ با محدودیت‌های سخت‌افزار ربات تلاقی می‌کند. از آنجا که مدل‌های زبانی بزرگ خروجی‌هایی با سطح انتزاع بالا تولید می‌کنند، لازم است این خروجی‌ها با مأذول‌های کنترل پایین‌سطح و برنامه‌ریزی‌های حرکتی ادغام شوند.

در بسیاری از سامانه‌ها، خروجی مدل زبانی بزرگ به صورت ساخت‌یافته ارائه می‌شود، مانند توالی‌های **JSON** (در **ROS**-**Python**)، کدهای **ARRC** (در

سامانه‌های مبتنی بر مدل‌های بینایی-زبانی **NLMap** نیز برای ایجاد نقشه‌های معنایی و انعطاف‌پذیر مورد استفاده قرار می‌گیرند و ربات را قادر می‌کنند دستورهای پیچیده و محیط‌های باز را بهتر درک کند. [1]

۸. تولید مبتنی بر بازیابی (RAG) برای برنامه‌ریزی مطمئن در رباتیک

با افزایش توانایی‌های مدل‌های زبانی بزرگ، یک چالش مهم همچنان بر جا می‌ماند. در رباتیک، این خطاهای می‌توانند منجر به اقدامات خطرناک یا غیرممکن شوند. از این‌رو، رویکرد تولید مبتنی بر بازیابی به عنوان یک راهکار مهم برای ثابت‌سازی و کاهش خطای مطرح شده است.

سامانه **ARRC** نمونه‌ای برجسته از کاربرد **RAG** در رباتیک است. [6] در این سامانه:

- یک پایگاه دانش برداری شامل الگوهای حرکتی، قالب‌های وظیفه، راهبردهای ایمنی و رهنمودهای عملیاتی ذخیره می‌شود.
- مأذول بازیابی، بخش‌های مرتبط را بر اساس دستور کاربر و وضعیت محیط استخراج می‌کند.
- مدل زبانی بزرگ با تکیه بر این دانش معتبر، یک برنامه ساخت‌یافته **JSON** تولید می‌کند.
- در نهایت، یک لایه «دروازه ایمنی» برنامه را پیش از اجرا ارزیابی می‌کند.

این خط لوله باعث می‌شود احتمال تولید اقدامات اشتباه یا ناسازگار از سوی مدل زبانی بزرگ به شکل چشم‌گیری کاهش یابد.

علاوه بر این، **RAG** نیاز به آموزش مجدد مدل را کاهش می‌دهد؛ زیرا افزودن مهارت‌های جدید تنها به اضافه کردن اطلاعات به پایگاه دانش نیاز دارد.

مزیت دیگر این رویکرد، امکان استدلال چندوجهی مبتنی بر محیط است. برای نمونه، در **ARRC** با استفاده از **AprilTags** و داده عمق، موقعیت دقیق اشیا استخراج شده و به مدل زبانی بزرگ داده می‌شود تا برنامه‌ریزی با شناخت واقعی تری از محیط انجام شود. [6] در مجموع، **RAG** آمیزش قدرتمندی میان قطعیت و ایمنی در رباتیک کلاسیک و انعطاف و استدلال مدل زبانی بزرگ ایجاد می‌کند.

۹. درک صحنه و ادغام چندوجهی

ادرک، رکن اساسی هوش تجسم‌یافته است. ربات برای تعامل با جهان باید بتواند اطلاعات بصری، فضایی و معنایی را به صورت قابل اعتماد پردازش کند.

²² unstructured

²⁰ Vision Language Models

²¹ sample efficiency



نسبت به آن‌ها اعتماد کافی ندارد. [1] سامانه **ROS-LLM** نیز از «پرجم خطای» در ساختار **MDP** استفاده می‌کند تا ربات بتواند شکست اجرای وظایف را تشخیص داده و در خواست بازخورد یا اصلاح از کاربر داشته باشد. [7]

تصویر 2. معماری در سطح بالا: مازول ادراک، مشاهدات^{۲۳} محور تولید می‌کند. برنامه‌ریز **RAG** دانش مربوط به کار را بازیابی کرده و یک طرح JSON می‌سازد. اجراکننده نیز با بررسی‌های اینمنی، دستورها را از طریق **XArm SDK** اعتبارسنجی و اجرا می‌کند.

۱۲. تعامل انسان-ربات و هوش مشارکتی

تعامل انسان-ربات (**HRI**) یکی از حوزه‌هایی است که مدل‌های زبانی بزرگ بیشترین اثرگذاری را در آن داشته‌اند. مشکل قدیمی رباتیک این بود که کاربران غیرمتخصص نمی‌توانستند ربات‌ها را به آسانی کنترل یا برنامه‌ریزی کنند، اما مدل‌های زبانی این شکاف را پر کرده‌اند.

۱. تعریف وظایف از طریق مکالمه

(**LLM**)، درخت‌های رفتاری **XML** برنامه‌های شبکه کد (در **Cap**) یا هدف‌های میانی برای یادگیری تقویتی (در **SayCan**). (تصویر 2)

ROS-LLM. این خروجی‌ها مستقیماً به سرویس‌ها و اکشن‌های تبدیل می‌شوند و به این ترتیب مدل می‌تواند با موتورهای مسیریابی، حل‌کننده‌های سینماتیک و کتابخانه‌هایی مانند **Movelt** به‌طور کامل یکپارچه شود. [7]

۱۱. اجرای ایمن (Execution Safety-Constrained)

ایمنی یکی از مهم‌ترین دغدغه‌ها در رباتیک است، بهویژه زمانی که ربات در کنار انسان‌ها کار می‌کند یا با اشیای حساس سروکار دارد.

مدل‌های زبانی بزرگ، با وجود توانایی‌های چشمگیرشان، ماهیتی احتمالی دارند و ممکن است خروجی‌هایی تولید کنند که از نظر فیزیکی نادرست یا خطرناک باشند. به همین دلیل، بسیاری از سامانه‌ها لایه‌های اضافی برای تضمین ایمنی در نظر گرفته‌اند.

در سامانه **ARRC**، هر اقدام پیشنهادی از سوی مدل زبانی بزرگ باید از یک مجموعه چندلایه از «دروازه‌های ایمنی»^{۲۴} زیر عبور کند. شامل: [6]

اعتبارسنجی محدوده کاری (**Workspace Validation**): محل‌های هدف بررسی می‌شوند تا خارج از محدوده قابل دسترس ربات نباشند.

محدودیت‌های سرعت و شتاب: حرکات از سقف‌های تعیین شده فراتر نمی‌روند.

نظارت بر گشتاور و نیروی گریپر: در صورت افزایش نیروی نامعمول، عملیات قطع می‌شود.

زمان‌سنج برای هر مرحله: اگر یک حرکت بیش از حد طولانی شود، عملیات متوقف می‌شود.

حالاتی ایمنی اضطراری: در صورت بروز خطاها مکرر، ربات به حالت امن بازمی‌گردد.

تعداد تکرار محدود برای تلاش مجدد.

این لایه‌ها تضمین می‌کنند که حتی اگر مدل زبانی بزرگ برنامه‌ای غیرایمن یا ناسازگار تولید کند، اجرا هرگز بدون کنترل و اعتبارسنجی صورت نمی‌گیرد.

رویکرد **KnowNo** نیز یک لایه ایمنی نظری مکمل ارائه می‌دهد. این مدل با برآورد عدم‌قطعیت خروجی، از انجام اقداماتی جلوگیری می‌کند که مدل

²³ Safety gating

۳. نیاز به داده و مشکلات انتقال دامنه

تنظیم تخصصی مدل‌ها برای کارهای صنعتی نیازمند منابع محاسباتی و داده زیاد است، حتی با وجود روش‌های بهینه مانند **QLoRA**. [4]

۴. ایمنی و اعتبارسنجی

از آنجا که مدل زبانی بزرگ ذاتاً احتمالی هستند، نمی‌توان ایمنی کامل را تنها با اتکا به آن‌ها تضمین کرد. به همین دلیل، اجرای امن نیازمند محدودکننده‌های سخت‌افزاری و اعتبارسنجی نمادین است.

۵. تعمیم در دنیای واقعی

ربات‌ها اغلب در محیط‌های واقعی نسبت به شبیه‌سازی عملکرد ضعیف‌تری دارند. مدل‌های چندوجهی کمک می‌کنند، اما مسئله هنوز حل نشده است.

۶. تأخیر و نیازهای زمان واقعی

استنتاج مدل زبانی بزرگ ممکن است کند باشد و برای وظایف حساس به زمان مناسب نباشد. بهینه‌سازی مدل‌ها یا اجرای روی سخت‌افزار محلی می‌تواند این مشکل را کاهش دهد.

۷. وظایف بلندمدت

برنامه‌های چندمرحله‌ای طولانی، هنوز به دلیل خطاهای تجمعی و محدودیت پنجه زمینه، چالش برانگیز هستند.

۸. ملاحظات اخلاقی و اجتماعی

با افزایش خودمختاری ربات‌ها، نگرانی‌هایی مانند مسئولیت، حریم خصوصی، اثرات شغلی و امنیت انسان‌ها اهمیت بیشتری می‌یابد.

۱۴. نتیجه‌گیری

ظهور مدل‌های زبانی بزرگ بنیان‌های رباتیک مدرن را دگرگون کرده است. این مدل‌ها امکان سطحی از انتراع، استدلال و تعامل طبیعی را فراهم کرده‌اند که پیش از این در سامانه‌های رباتیکی وجود نداشت. مرور انجامشده - که حوزه‌هایی همچون برنامه‌ریزی چندوجهی، دستکاری مبتنی بر استدلال، تنظیم تخصصی صنعتی، کنترل مبتنی بر بازیابی و چارچوب‌های **ROS** محور را دربر می‌گیرد - نشان می‌دهد که مدل‌های زبانی بزرگ دیگر صرفاً نقش مکمل ندارند، بلکه به هسته شناختی سیستم‌های رباتیکی تبدیل شده‌اند.

روند غالب در تمام پژوهش‌های بررسی شده، گذار از پردازش زبان مبتنی بر قواعد به سوی درک معنایی غنی و انعطاف‌پذیر است. در گذشته، سامانه‌های

سامانه‌هایی که از مدل‌های زبانی بزرگ استفاده می‌کنند، می‌توانند دستورهای پیچیده را از جملات کاملاً طبیعی استخراج کنند. برای مثال، کاربر می‌تواند بگوید: «میز را تمیز کن و ظرف‌ها را داخل سینک بگذار.» مدل‌های زبانی بزرگ این جمله را به مجموعه‌ای از وظایف شامل حرکت، گرفتن، جابه‌جایی و رهاسازی ترجمه می‌کند. [1]

۲. بازخورد و اصلاح تعاملی

در **ROS-LLM**، ربات پس از هر مرحله می‌تواند بازخورد کاربر را دریافت کرده و رفتار خود را اصلاح کند. [7] این فرآیند مشابه روش‌های آموزشی انسانی است؛ مشاهده رفتار، ارائه اصلاح و یادگیری از تجربه.

۳. درک ضمنی

سامانه‌های **NLexe** قادرند از جملات غیرمستقیم مانند «تشنه هستم» معنای درخواست آب را استنباط کنند. [3] این توانایی برای تعامل طبیعی و انسانی کاملاً ضروری است.

۴. تعامل چندوجهی

برخی سامانه‌ها علاوه بر متن، از تصویر یا داده‌های حسگر نیز استفاده می‌کنند. برای مثال، کاربر می‌تواند تصویر را نشان دهد و بگوید «لیوان را کنار این قرار بده.» و ربات موقعیت اشاره‌شده را تفسیر می‌کند.

۵. عامل‌های مشارکتی و حافظه‌دار

پژوهش‌های جدید درباره «عامل‌های مولد» نشان می‌دهند که ربات می‌تواند حافظه بلندمدت شکل دهد، الگوهای رفتاری پایدار ایجاد کند و تجربه گذشته را برای بهبود تعامل استفاده کند. [1] در نتیجه، **HRI** مبتنی بر مدل‌های زبانی بزرگ به ربات‌هایی منجر می‌شود که قابل‌اعتمادتر، قابل تعاملی‌تر، قابل‌سازگاری‌تر و برای کاربران غیرمتخصص بسیار قابل استفاده‌تر هستند.

۱۳. محدودیت‌ها، چالش‌ها

۱. ثابت‌سازی و هم‌ترازی با جهان واقعی

مدل‌های زبانی بزرگ گاهی اقداماتی پیشنهاد می‌دهند که از نظر فیزیکی ممکن نیست. سامانه‌هایی مانند **ARRC** و **SayCan** این مشکل را کاهش می‌دهند، اما گران‌بینیگ کامل چندوجهی همچنان دشوار است.

۲. دقیقت عددی و محدودیت‌های فیزیکی

مدل‌های زبانی در استدلال و معنا بسیار توانمندند، اما در محاسبات دقیق، هندسه سه‌بعدی، پیش‌بینی پیوسته حرکت ضعیف‌تر عمل می‌کنند. از همین رو روش‌هایی مانند **RT-Grasp** ضروری‌اند. [5]

- [2] B. Gianluca, . D. Danilo, M. Enrico and O. Francesco, "Parsing Natural Language Sentences into Robot," p. 4, 2019.
- [3] L. Rui, G. Yibei, J. Runxiang and Z. Xiaoli, "A Review of Natural-Language-Instructed Robot," p. 42, 2024.
- [4] A. Benjamin, K. Urs, T. Aleksandar, K. Darko and H. Andreas, "Domain-Specific Fine-Tuning of Large Language," p. 5, 2021.
- [5] X. Jinxuan, J. Shiyu, L. Yutian, Z. Yuqian and Z. Liangjun, "RT-Grasp: Reasoning Tuning Robotic Grasping via," p. 8, 2024.
- [6] V. Eugene, M. Ammar Jaleel, . R. Salim and C. Robin, "ARRC: Advanced Reasoning Robot Control—Knowledge-Driven Autonomous Manipulation Using Retrieval-Augmented Generation," p. 8, 2025.
- [7] E. M. Christopher, W. Yuhui, Hongzhan, G. Antoine, J. Gonzalez-Billandon, M. Zimmer and J. Wang, "ROS-LLM: A ROS framework for embodied AI with task feedback and structured reasoning," p. 26, 2024.
- [8] F. Zeng, W. Gan, Y. Wang, N. Liu and P. S. Yu, "Large Language Models for Robotics: A Survey," p. 19, 2023.

مبتنی بر هستی‌شناسی [2] ربات‌ها را به مجموعه محدودی از الگوهای دستوری مقید می‌کردند. اما مدل‌های زبانی بزرگ امروزی با تکیه بر دانش جهان‌واقعی و توانایی استدلال مبتنی بر زمینه، قادرند دستورات مبهم، غیرمستقیم یا ناقص را تحلیل کنند، نیت انسان را تشخیص دهند و با تغییرات محیطی سازگار شوند. این توانایی گذار ربات‌ها را از «پیرو دستور» به «همکار هوشمند» ممکن کرده است.

در زمینه برنامه‌ریزی، مدل‌های زبانی اکنون قادرند برنامه‌های بلندمدت، زمینه‌محور و قابل اجرا تولید کنند و آن‌ها را با کمک مکانیزم‌های ثابت‌سازی و برنامه‌ریزی‌های کلاسیک تکمیل نمایند. [1] ادغام ورودی‌های زبانی چندوجهی و دانش پیشین نشان داده است که چشم‌انداز مدل‌های زبانی بزرگ به عنوان «برنامه‌ریزی‌های عمومی» روزبه‌روز قابل دسترس‌تر می‌شود.

در حوزه دستکاری، روش‌هایی مانند **RT-Grasp** و **LLM-GROP** و **VIMA** [5] نشان داده‌اند که زبان طبیعی می‌تواند راهنمایی برای درک ویژگی‌های اشیا، قیود ایمنی و راهبردهای دستکاری فراهم کند. ادغام استدلال زبانی با کنترل پیوسته، یکی از شکاف‌های دیرینه در رباتیک را پُر کرده است.

معماری‌های ماژولار مانند **MetaMorph** و **PROGRAMPORT** نشان می‌دهند که مدل‌های زبانی بزرگ قادرند ساختارهای رفتاری مقیاس‌پذیر ایجاد کنند. در کنار آن، رویکردهای **RAG** مانند [6] **ARRC** ثابت کرده‌اند که با تکیه بر پایگاه دانش واقعی، می‌توان خطأ و توهمندی مدل را کنترل کرده و برنامه‌ریزی قابل اعتماد ارائه داد.

در حوزه تعامل انسان–ربات، سامانه‌هایی نظیر **ROS-LLM** [7] نشان داده‌اند که زبان طبیعی می‌تواند واسطه اصلی تعامل باشد. این امکان به کاربران غیرمتخصص اجازه می‌دهد ربات‌ها را اصلاح، هدایت یا برنامه‌ریزی کنند – آن‌ها هم بدون نیاز به دانش فنی.

با وجود این پیشرفت‌ها، چالش‌هایی همچون دقت عددی، ایمنی، تعمیم، پردازش زمان واقعی و مسائل اخلاقی همچنان پابرجاست. مدل‌های زبانی بزرگ ممکن است در برخی شرایط رفتار غیرقابل اعتماد نشان دهند، بنابراین ترکیب آن‌ها با مکانیزم‌های محافظتی و مدل‌های مبتنی بر فیزیک ضروری است.

منابع :

- [1] W. Jiaqi, W. Zihao, L. Yiwei, J. Hanqi and S. Peng, "Large Language Models for Robotics: Opportunities, Challenges, and Perspectives," p. 18, 2021.