به نام خدا

گزارش فاز اول پروژه: سامانه کنترل ربات مبتنی بر مدل‌های زبانی بزرگ

محمد مهدی اسدی

امیر حسین سلیمانی

استاد راهنما: دکتر مازیار پالهنگ

آذر ۱۴۰۴

فهرست

چکیده

مدل‌های زبانی بزرگ[1] به یکی از عناصر تحول‌آفرین 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[2] 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 به‌تدریج نقش هسته شناختی را بر عهده می‌گیرند؛ هسته‌ای که ادراک، تصمیم‌گیری، کنترل و تعامل انسان–ربات را یکپارچه می‌کند.

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 ماژولار برای ربات‌ها را تقویت می‌کنند. همچنین به روش‌هایی مانند برنامه‌ریزی مبتنی بر ثابت‌سازی تحلیل نحوی مبتنی بر هستی‌شناسی و چارچوب‌های نظام‌مند آموزش ربات از طریق زبان انسان می‌پردازیم.

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

با سازمان‌دهی این یافته‌ها در حوزه‌های برنامه‌ریزی، استدلال، دستکاری، ادراک، معماری‌های کنترلی، یک نگاه یکپارچه از وضعیت کنونی هوش تجسم‌یافته مبتنی بر مدل زبانی بزرگ ارائه می‌دهیم. در پایان نیز چالش‌هایی مانند ایمنی، 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.

[1] Large language model
[2] manipulation

مقدمه

پیشرفت‌های اخیر در حوزه‌ی هوش مصنوعی، به‌ویژه ظهور مدل‌های زبانی بزرگ، موجب تحول بنیادین در بسیاری از شاخه‌های مهندسی و علوم کامپیوتر شده است. یکی از حوزه‌هایی که به‌طور مستقیم از این پیشرفت‌ها تأثیر پذیرفته، علم رباتیک است. در رویکردهای کلاسیک رباتیک، سامانه‌های رباتیکی عمدتاً بر پایه‌ی برنامه‌نویسی صریح، قوانین از پیش‌تعریف‌شده و مدل‌های کنترلی سخت‌گیرانه طراحی می‌شدند. این رویکرد اگرچه در محیط‌های کنترل‌شده عملکرد قابل قبولی داشت، اما در مواجهه با محیط‌های پویا، دستورات مبهم انسانی و وظایف پیچیده و چندمرحله‌ای با محدودیت‌های جدی روبه‌رو بود.

در سال‌های اخیر، مدل‌های زبانی بزرگ به‌عنوان یک ابزار شناختی قدرتمند، امکان تعامل طبیعی‌تر میان انسان و ماشین را فراهم کرده‌اند. این مدل‌ها قادرند زبان طبیعی را نه‌تنها به‌عنوان یک ورودی متنی، بلکه به‌عنوان بستری برای استدلال، برنامه‌ریزی و تصمیم‌گیری به‌کار گیرند. استفاده از این قابلیت‌ها در رباتیک، مسیر جدیدی را تحت عنوان «رباتیک مبتنی بر مدل‌های زبانی بزرگ» گشوده است؛ مسیری که در آن زبان طبیعی به یک رابط اصلی میان انسان و ربات تبدیل می‌شود و ربات‌ها می‌توانند نیت انسانی را در سطحی فراتر از دستورات صریح درک کنند.

در این گزارش، این مبحث به‌صورت گام‌به‌گام و ساختاریافته در فصل‌های اول تا هفتم مورد بررسی قرار گرفته است. در فصل اول، تغییر پارادایم در رباتیک و گذار از رویکردهای کلاسیک به سامانه‌های مبتنی بر مدل‌های زبانی بزرگ تشریح شده و مفاهیمی نظیر هوش تجسم‌یافته و نقش زبان در ادراک و کنش ربات معرفی گردیده است .فصل دوم به درک زبان طبیعی و تعامل انسان–ربات به‌عنوان زیربنای شناختی این سامانه‌ها می‌پردازد و نشان می‌دهد که چگونه مدل‌های زبانی بزرگ می‌توانند ابهام، نیت ضمنی و گفت‌وگوی تعاملی را مدیریت کنند.

در ادامه، فصل سوم برنامه‌ریزی وظایف پیچیده و اجرای بلندمدت با هدایت زبان طبیعی را بررسی می‌کند و نقش مدل‌های زبانی را به‌عنوان برنامه‌ریزهای سطح‌بالا توضیح می‌دهد .فصل چهارم بر چالش‌های دستکاری فیزیکی و استدلال چندوجهی تمرکز دارد و نحوه‌ی پیوند معنا، ادراک بصری و کنترل فیزیکی را مورد بحث قرار می‌دهد. در فصل پنجم، معماری‌های ماژولار و شیوه‌های ادغام مدل‌های زبانی با پشته‌های رباتیکی موجود مانند ROS معرفی می‌شوند. سپس در فصل ششم، مسائل ایمنی، قابلیت اطمینان و نقش انسان در حلقه‌ی تصمیم‌گیری بررسی شده و اهمیت طراحی سازوکارهای کنترلی و نظارتی برجسته می‌گردد. در نهایت، فصل هفتم به محدودیت‌ها، چالش‌های باز و مسیرهای آینده‌ی این حوزه اختصاص دارد.

هدف از ارائه‌ی این گزارش، فراهم‌کردن دیدی جامع و تحلیلی نسبت به نقش مدل‌های زبانی بزرگ در سامانه‌های رباتیکی و بررسی مزایا، چالش‌ها و چشم‌اندازهای پیشِ رو است؛ به‌گونه‌ای که خواننده بتواند تصویری روشن از وضعیت فعلی و مسیر توسعه‌ی این فناوری به‌دست آورد.

فصل ۱: مقدمه و تغییر پارادایم در رباتیک مبتنی بر مدل‌های زبانی بزرگ

در دهه‌های گذشته، رباتیک عمدتاً بر پایه‌ی برنامه‌نویسی صریح، مدل‌های نمادین سخت‌گیرانه و زنجیره‌های کنترلی از پیش‌تعریف‌شده توسعه یافته است. در این رویکرد کلاسیک، هر رفتار ربات باید به‌صورت مستقیم و با جزئیات دقیق توسط مهندسان طراحی می‌شد؛ از تعریف سناریوهای ممکن گرفته تا مدیریت خطاها و تعامل با انسان. این شیوه اگرچه در محیط‌های صنعتی کنترل‌شده موفق بوده است، اما در مواجهه با محیط‌های پویا، دستورات مبهم انسانی و وظایف چندمرحله‌ای پیچیده به سرعت به بن‌بست می‌رسد. محدودیت اصلی این نمونه[۳]، ناتوانی آن در درک و استفاده از زبان طبیعی به‌عنوان یک رابط عمومی شناختی میان انسان و ربات است.

ظهور مدل‌های زبانی بزرگ (Large Language Models) این وضعیت را به‌طور بنیادین تغییر داده است. LLMها که بر حجم عظیمی از داده‌های متنی آموزش دیده‌اند، قادرند ساختارهای زبانی پیچیده، روابط معنایی پنهان و حتی نیت ضمنی انسان را استخراج کنند. زبان طبیعی از یک ابزار جانبی به یک مؤلفه‌ی مرکزی در ادراک، تصمیم‌گیری، برنامه‌ریزی و کنترل ربات تبدیل می‌شود. [1]در این چارچوب جدید، ربات دیگر صرفاً یک ماشین اجراکننده‌ی دستورات از پیش‌تعریف‌شده نیست، بلکه به سامانه‌ای با سطحی از استدلال شناختی و توانایی تفسیر هدف انسانی بدل می‌شود.

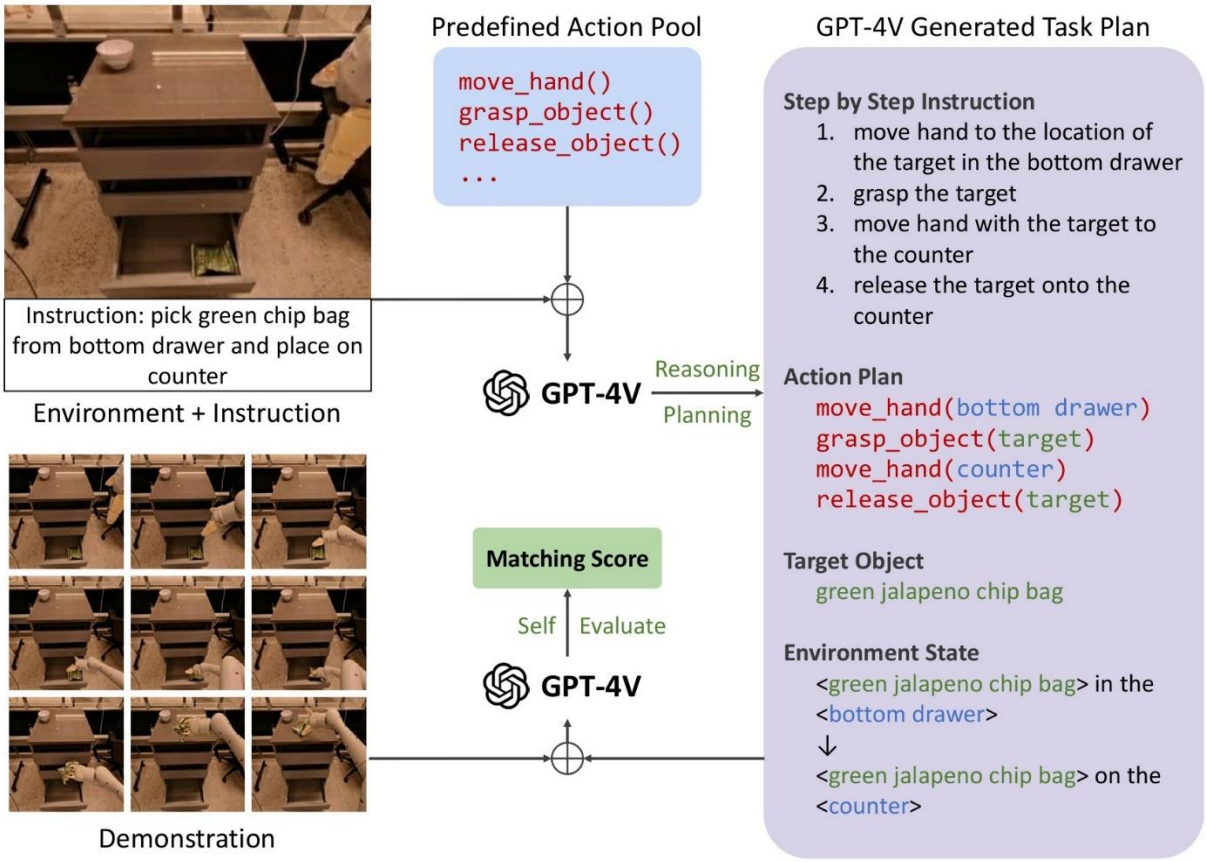
یکی از مفاهیم کلیدی که این گذار را توضیح می‌دهد، «هوش تجسم‌یافته» است. هوش تجسم‌یافته[۴] بر این ایده تأکید دارد که هوش واقعی تنها در سطح محاسبات نمادین شکل نمی‌گیرد، بلکه در تعامل مستمر میان ادراک، بدن فیزیکی و محیط معنا می‌یابد. مدل‌های زبانی بزرگ، با پیوند دادن زبان به ادراک چندوجهی[۵] و کنش فیزیکی، امکان شکل‌گیری چنین هوشی را فراهم می‌کنند. در سامانه‌های مدرن، LLMها به‌عنوان لایه‌ی شناختی عمل می‌کنند که اطلاعات زبانی، بصری و زمینه‌ای را یکپارچه کرده و آن را به تصمیم‌های قابل اجرا در جهان واقعی تبدیل می‌سازد. [2]

اهمیت این تغییر مدل زمانی روشن‌تر می‌شود که به نمونه‌های عملی نگاه کنیم. برای مثال، سامانه‌ی SayCan که توسط گوگل توسعه داده شده است، نشان می‌دهد چگونه می‌توان استدلال زبانی سطح‌بالا را با مهارت‌های کنترلی از پیش‌آموخته‌ی ربات ترکیب کرد. در این سامانه، ربات دستورات باز و غیرساخت‌یافته‌ای مانند «یک نوشیدنی برایم آماده کن» را دریافت کرده و آن‌ها را به توالی‌ای از اقدامات فیزیکی مانند حرکت در محیط، باز کردن کشوها، گرفتن اشیا و قرار دادن آن‌ها در مکان مناسب تبدیل می‌کند . [1] نکته‌ی مهم این است که کاربر نیازی به دانستن جزئیات کنترلی یا زبان برنامه‌نویسی ربات ندارد؛ زبان طبیعی به‌تنهایی به رابط اصلی تعامل تبدیل شده است.

نمونه‌ی برجسته‌ی دیگر، PaLM-E است که گامی فراتر از برنامه‌ریزی وظیفه برمی‌دارد و مدل زبانی را مستقیماً در حلقه‌ی ادراک و تصمیم‌گیری ربات قرار می‌دهد. در این سامانه، مدل زبانی نه‌تنها متن، بلکه تصاویر و بردارهای حالت ربات را نیز به‌عنوان ورودی دریافت می‌کند و می‌تواند به پرسش‌هایی درباره‌ی محیط پاسخ دهد یا تصمیم‌های کنترلی اتخاذ کند. [1] این رویکرد نشان می‌دهد که LLMها می‌توانند از یک «مولد برنامه[۶]» به یک «مغز تجسم‌یافته[۷]» برای ربات تبدیل شوند؛ مغزی که به‌طور پیوسته با محیط فیزیکی در تعامل است. تصویر 1)

[3] paradigm
[4] Embodied intelligence
[5] Multimodal perception
[6] Program generator
[7] The embodied brain

آنچه این مثال‌ها را به هم پیوند می‌دهد، حرکت از رباتیک مبتنی بر دستورالعمل‌های صریح به رباتیک مبتنی بر نیت و معنا است. در این نمونه جدید، مهندسی رباتیک بیش از آنکه به نوشتن کدهای کنترلی محدود شود، به طراحی معماری‌هایی می‌پردازد که بتوانند خروجی‌های احتمالی و انعطاف‌پذیرLLM ها را به رفتارهای ایمن و قابل اعتماد تبدیل کنند. [1]



تصویر 1. چارچوب روش پیشنهادی برای برنامه‌ریزی وظایف تجسدی 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 برمی‌گزیند. 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. [1]

فصل ۲: درک زبان طبیعی و تعامل انسان–ربات به‌عنوان زیربنای شناختی

درک زبان طبیعی سنگ‌بنای رباتیک مبتنی بر مدل‌های زبانی بزرگ است. اگر ربات نتواند زبان انسان را به‌درستی تفسیر کند، تمامی قابلیت‌های بعدی مانند برنامه‌ریزی، تصمیم‌گیری و کنترل فیزیکی با اختلال مواجه می‌شوند. در رباتیک کلاسیک، درک زبان عمدتاً به نگاشت‌های ساده‌ی

نحوی یا قواعد از پیش‌تعریف‌شده محدود بود. این سامانه‌ها معمولاً بر هستی‌شناسی‌های[8] دستی، وابستگی‌شناسی نحوی[9] و استخراج افعال و اشیای صریح تکیه داشتند و تنها در دامنه‌های بسیار محدود عملکرد قابل قبولی نشان می‌دادند. [3] نتیجه‌ی چنین رویکردی، سامانه‌هایی شکننده بود که در مواجهه با ابهام، بیان غیرمستقیم یا تغییرات جزئی در جمله به‌سرعت دچار خطا می‌شدند. [4]

مدل‌های زبانی بزرگ این وضعیت را به‌طور اساسی تغییر داده‌اند. LLMها به‌جای تکیه بر قواعد صریح، از بازنمایی‌های معنایی غنی استفاده می‌کنند که از داده‌های عظیم متنی آموخته شده‌اند. این بازنمایی‌ها به مدل اجازه می‌دهند روابط ضمنی، دانش عمومی و زمینه‌ی گفتار را در نظر بگیرد. به‌عنوان مثال، یک مدل زبانی بزرگ می‌تواند از جمله‌ای مانند «هوا سرد است» نیت ضمنی «پنجره را ببند» یا «بخاری را روشن کن» را استنتاج کند، بدون آنکه این عمل به‌طور مستقیم بیان شده باشد. [3] چنین قابلیتی برای تعامل طبیعی انسان–ربات حیاتی است، زیرا زبان انسانی ذاتاً سرشار از ابهام و ارجاعات ضمنی است. [3]

علاوه بر استخراج نیت، LLMها نقش مهمی در گفت‌وگوی تعاملی ایفا می‌کنند. در سامانه‌های مدرن، اگر دستور کاربر ناقص یا ناسازگار با وضعیت فیزیکی ربات باشد، مدل می‌تواند به‌جای شکست خاموش، پرسش تکمیلی مطرح کند یا گزینه‌های ممکن را پیشنهاد دهد. این رفتار گفت‌وگومحور، زبان را از یک کانال یک‌طرفه‌ی فرمان‌دهی به یک ابزار مذاکره و اصلاح مشترک میان انسان و ربات تبدیل می‌کند. [2] در نتیجه، تعامل انسان–ربات به‌جای آنکه شبیه برنامه‌نویسی باشد، به مکالمه‌ای هدف‌محور نزدیک می‌شود.

پیاده‌سازی عملی این مفاهیم را می‌توان در سامانه‌های کنترل زبانی ربات‌های واقعی مشاهده کرد. یکی از نمونه‌های کلاسیک، سیستم «Parsing Natural Language Sentences into Robot Actions» است که بر روی یک ربات انسان‌نمای واقعی مانند NAO یا Zora پیاده‌سازی شده است. در این سامانه، دستورات گفتاری کاربر مانند «دستت را بالا ببر» یا «سرت را به چپ بچرخان» به اعمال فیزیکی متناظر ترجمه می‌شوند. [4] نکته‌ی مهم این است که کاربر نیازی به دانستن نام موتورها، مفاصل یا توالی‌های کنترلی ندارد؛ زبان طبیعی به‌تنهایی برای هدایت ربات کافی است. این مثال نشان می‌دهد که چگونه حتی پیش از ظهورLLMهای مدرن، زبان به‌عنوان رابط کنترل مطرح بوده، اما با دامنه‌ای محدود و نیازمند مهندسی دستی گسترده.

توانایی رفع ابهام نیز در همین سامانه به‌صورت عملی نمایش داده شده است. زمانی که کاربر دستور ناقصی مانند «دستت را بالا ببر» صادر می‌کند، ربات به‌طور فعال می‌پرسد «دست چپ یا راست؟» و تنها پس از دریافت پاسخ، عمل را اجرا می‌کند. [4] این رفتار، نمونه‌ای ساده اما بسیار مهم از گفت‌وگوی تعاملی است که در ربات‌های واقعی آزمایش شده و نشان می‌دهد چرا درک زبان طبیعی نمی‌تواند به یک نگاشت ایستا محدود شود. LLMها این ایده را گسترش می‌دهند و امکان پرسش‌های پیچیده‌تر، پیشنهاد گزینه‌ها و حتی مذاکره درباره‌ی هدف را فراهم می‌کنند.

جنبه‌ی دیگر درک زبان طبیعی، آگاهی از وضعیت درونی و فیزیکی ربات است. در همان سامانه‌ی انسان‌نما، دو حالت اجرای «stateful» و «stateless» تعریف شده است؛ در حالت اول، ربات وضعیت بدنی خود را میان دستورات حفظ می‌کند و در حالت دوم، پس از هر عمل به وضعیت پیش‌فرض بازمی‌گردد. [4] این تمایز نشان می‌دهد که فهم زبان به‌تنهایی کافی نیست، بلکه تفسیر دستور باید در زمینه‌ی حالت فعلی ربات انجام شود. LLMها با توانایی نگهداری زمینه‌ی مکالمه و تاریخچه‌ی تعامل، این مسئله را در مقیاس بسیار وسیع‌تری حل می‌کنند.

[8] Based Ontologies
[9] Syntactic dependency theory

ایمنی نیز به‌طور مستقیم با درک زبان طبیعی گره خورده است. در مثال‌های واقعی، ربات انسان‌نما دستورات ناسازگار یا خطرناک، مانند بلند کردن یک پا در حالی که روی پای دیگر ایستاده است، را رد می‌کند و دلیل آن را به زبان طبیعی توضیح می‌دهد. [4] این رفتار نشان می‌دهد که نگاشت زبان به عمل باید همواره با دانش محدودیت‌های فیزیکی همراه باشد. در سامانه‌های مبتنی بر LLM، این دانش می‌تواند به‌صورت ضمنی از داده‌های آموزشی یا به‌صورت صریح از طریق لایه‌های ایمنی به مدل تزریق شود. [5]

فصل ۳: برنامه‌ریزی وظیفه و اجرای بلندمدت با هدایت زبان

پس از آنکه زبان طبیعی به‌عنوان یک رابط شناختی قابل اعتماد میان انسان و ربات تثبیت شد، گام منطقی بعدی استفاده از این توانمندی برای برنامه‌ریزی وظایف پیچیده و چندمرحله‌ای است. برنامه‌ریزی وظیفه[10] در رباتیک به معنای تبدیل یک هدف سطح‌بالا، که اغلب به‌صورت زبانی بیان می‌شود، به دنباله‌ای از اعمال اجرایی است که ربات بتواند آن‌ها را در دنیای فیزیکی انجام دهد. در رباتیک کلاسیک، این فرایند معمولاً به روش‌های نمادین مانند برنامه‌ریزی مبتنی بر حالت، درخت جست‌وجو یا زبان‌هایی نظیر PDDL محدود می‌شد. این روش‌ها اگرچه از نظر صوری دقیق هستند، اما نیازمند مدل‌سازی کامل محیط و تعریف صریح تمامی پیش‌شرط‌ها و اثرات اعمال‌اند؛ امری که در محیط‌های واقعی به‌ندرت امکان‌پذیر است. [1]

مدل‌های زبانی بزرگ رویکردی متفاوت به این مسئله ارائه می‌دهند. LLMها قادرند از دانش عمومی و ساختارهای روایی زبان برای تجزیه‌ی یک هدف کلی به زیروظایف معنادار استفاده کنند، حتی در شرایطی که مدل صریحی از محیط در اختیار ندارند. به بیان دیگر، آن‌ها می‌توانند نقش یک «برنامه‌ریز سطح‌بالا[11]» را ایفا کنند که خروجی آن نه دستورات کنترلی دقیق، بلکه یک توالی مفهومی از اقدامات است. [3]

با این حال، استفاده‌ی مستقیم از خروجی زبانی LLMها برای کنترل ربات با چالش‌های جدی مواجه است. خروجی‌های مدل‌های زبانی ذاتاً احتمالی‌اند و تضمین رسمی درباره‌ی صحت یا ایمنی آن‌ها وجود ندارد. به همین دلیل، معماری‌های عملی معمولاً از یک ساختار ترکیبی استفاده می‌کنند که در آن LLM وظیفه‌ی تولید یا ارزیابی برنامه‌ی سطح‌بالا را بر عهده دارد و اجرای واقعی توسط ماژول‌های کنترلی قابل اعتماد انجام می‌شود. این رویکرد ترکیبی، که گاه با عنوان «برنامه‌ریزی گراندشده[12]» شناخته می‌شود، یکی از محورهای اصلی پژوهش در رباتیک مبتنی بر LLM . [2]

پیاده‌سازی عملی این ایده به‌طور شاخص در سامانه‌ی SayCan دیده می‌شود. در SayCan، مدل زبانی بزرگ وظیفه دارد بر اساس دستور زبانی کاربر، مجموعه‌ای از اقدامات ممکن را ارزیابی کند و احتمال مناسب بودن هر اقدام برای رسیدن به هدف را تخمین بزند. این اقدامات از پیش به‌صورت مهارت‌های قابل اجرا روی ربات تعریف شده‌اند، مانند «حرکت به سمت میز»، «باز کردن کشو» یا «گرفتن فنجان» [1] سپس، این امتیازدهی زبانی با احتمال موفقیت هر مهارت که از داده‌های تجربی به‌دست آمده است ترکیب می‌شود. نتیجه، انتخاب عملی است که هم از نظر معنایی با هدف کاربر همخوانی دارد و هم از نظر فیزیکی قابل اجراست.

این معماری نشان می‌دهد که LLMها به‌جای جایگزینی کامل برنامه‌ریزی کلاسیک، می‌توانند آن را تکمیل کنند. در SayCan، زبان به‌عنوان راهنمای انتخاب در فضای اعمال عمل می‌کند، نه به‌عنوان کنترل‌کننده‌ی مستقیم موتورها. برای مثال، زمانی که کاربر می‌گوید «برایم یک نوشیدنی آماده کن»، مدل زبانی می‌داند که ابتدا باید به آشپزخانه رفت، سپس یک لیوان پیدا کرده و در نهایت آن را پر کرد، حتی اگر این

[10] Task programming
[11] High level programming
[12] Grounded planning

مراحل به‌طور صریح در دستور ذکر نشده باشند. [1] این توانایی استخراج ساختار وظیفه از زبان طبیعی دقیقاً همان چیزی است که در برنامه‌ریزی کلاسیک به‌سختی و با هزینه‌ی مهندسی بالا به‌دست می‌آید.

نمونه‌ی دیگر از این رویکرد، سامانه‌های مبتنی بر ROS-LLM هستند که در آن‌ها مدل زبانی بزرگ به‌عنوان یک لایه‌ی برنامه‌ریز روی پشته‌ی استاندارد رباتیکی ROS قرار می‌گیرد. در این سامانه‌ها، دستورات زبانی کاربر به توالی‌ای از فراخوانی سرویس‌ها و اکشن‌های ROS تبدیل می‌شود. [2] مزیت این روش در آن است که می‌توان بدون تغییر در زیرساخت کنترلی موجود، قابلیت برنامه‌ریزی زبانی را به ربات افزود. به‌علاوه، ROS-LLM نشان می‌دهد که چگونه LLMها می‌توانند با ابزارهای مهندسی جاافتاده ادغام شوند، نه اینکه به‌عنوان یک سامانه‌ی جداگانه عمل کنند.

اجرای بلندمدت وظایف یکی دیگر از جنبه‌های مهم برنامه‌ریزی زبانی است. بسیاری از وظایف دنیای واقعی، مانند تمیز کردن یک اتاق یا آماده‌سازی یک وعده‌ی غذایی، شامل ده‌ها مرحله هستند و ممکن است در طول اجرا با خطا یا تغییر شرایط مواجه شوند. LLMها به دلیل توانایی نگهداری زمینه و بازتفسیر هدف، برای مدیریت چنین وظایفی مناسب‌اند. [1] در عمل، این ویژگی به ربات اجازه می‌دهد پس از یک شکست جزئی، برنامه‌ی خود را اصلاح کند یا از کاربر راهنمایی بگیرد، به‌جای آنکه کل فرایند را متوقف سازد.

با این حال، مثال‌های عملی نشان می‌دهند که این توانمندی هنوز محدودیت‌هایی دارد. در سامانه‌های واقعی، اجرای بلندمدت اغلب نیازمند نظارت انسانی یا محدود کردن دامنه‌ی تصمیم‌گیری LLM است تا از انحراف از هدف اصلی جلوگیری شود. [2] این موضوع بار دیگر اهمیت معماری‌های ترکیبی را برجسته می‌کند؛ معماری‌هایی که در آن‌ها زبان به‌عنوان ابزار استدلال و هدایت به کار می‌رود، اما کنترل نهایی همچنان در چارچوب‌های ایمن و قابل پیش‌بینی انجام می‌شود.

فصل ۴: دستکاری فیزیکی، استدلال چندوجهی و پیوند معنا با کنترل

اگرچه در فصل پیش نشان داده شد که مدل‌های زبانی بزرگ می‌توانند اهداف سطح‌بالا را به برنامه‌های اجرایی معنادار تبدیل کنند، اما یکی از دشوارترین چالش‌ها در رباتیک همچنان باقی می‌ماند: دستکاری فیزیکی اشیا در دنیای واقعی. دستکاری نیازمند دقت عددی بالا، ادراک پیوسته و واکنش سریع به عدم‌قطعیت‌های محیطی است. به‌طور سنتی، این حوزه بر مدل‌های هندسی، برنامه‌ریزی مسیر و کنترل بازخوردی تکیه داشته و فاصله‌ی قابل توجهی میان استدلال نمادین و اجرای فیزیکی وجود داشته است. [6]

مدل‌های زبانی بزرگ به‌تنهایی قادر به حل مستقیم مسائل کنترل پیوسته نیستند، اما می‌توانند نقش مهمی در راهنمایی معنایی فرایند دستکاری ایفا کنند. ایده‌ی اصلی آن است که LLMها دانش سطح‌بالا درباره‌ی اشیا، کارکرد آن‌ها و راهبردهای متداول دستکاری را فراهم می‌کنند، در حالی که جزئیات عددی همچنان به ماژول‌های کنترلی تخصصی سپرده می‌شود. برای مثال، دانستن اینکه «فنجان را باید از دسته گرفت» یا «بطری را باید عمودی نگه داشت» نوعی دانش معنایی است که به‌راحتی در زبان بیان می‌شود اما به‌سختی در قالب قواعد هندسی صریح مدل‌سازی می‌شود. [7]

این رویکرد به‌ویژه در معماری‌های چندوجهی اهمیت می‌یابد. در چنین سامانه‌هایی، زبان، بینایی و حالت فیزیکی ربات به‌صورت هم‌زمان در تصمیم‌گیری دخیل هستند. LLM ها در این میان نقش ادغام‌کننده‌ی معنا را بازی می‌کنند؛ یعنی اطلاعات بصری و حسی را در چارچوب مفاهیم زبانی تفسیر می‌کنند و بر اساس آن، راهبرد مناسب دستکاری را پیشنهاد می‌دهند. [2]

پیاده‌سازی عملی این ایده به‌طور شاخص در سامانه‌ی RT-Grasp دیده می‌شود. RT-Grasp یک سیستم دستکاری است که از مدل‌های زبانی بزرگ برای هدایت فرایند گرفتن اشیا استفاده می‌کند. در این سامانه، پیش از آنکه شبکه‌ی عصبی عددی محل و زاویه‌ی گرفتن را پیش‌بینی کند، یک مرحله‌ی استدلال زبانی انجام می‌شود که هدف و محدودیت‌های دستکاری را مشخص می‌کند. [7] رای مثال، اگر هدف «برداشتن یک لیوان پر از آب» باشد، استدلال زبانی می‌تواند بر حفظ تعادل و جلوگیری از ریختن محتوا تأکید کند و بدین ترتیب، فضای جست‌وجوی کنترل عددی را محدود سازد.

اهمیت RT-Grasp در این است که نشان می‌دهد افزودن لایه‌ی زبانی می‌تواند هم ایمنی و هم قابلیت تعمیم سیستم را افزایش دهد. به‌جای آنکه مدل کنترل عددی مجبور باشد تمامی حالات ممکن را از داده بیاموزد، بخشی از دانش به‌صورت زبانی و قابل تفسیر در اختیار سیستم قرار می‌گیرد. این موضوع به‌ویژه در محیط‌های جدید یا اشیای نادیده‌شده اهمیت دارد، جایی که داده‌ی آموزشی محدود است. [6] [7]

نمونه‌ی دیگری از پیوند معنا و کنترل را می‌توان در سامانه‌ی VoxPoser مشاهده کرد. VoxPoser از زبان طبیعی برای تعریف قیود فضایی و هدف‌های دستکاری استفاده می‌کند؛ برای مثال، کاربر می‌تواند بگوید «این مکعب را کنار لیوان و دور از لبه‌ی میز قرار بده». سیستم این دستور را به قیود هندسی قابل اجرا برای برنامه‌ریز مسیر تبدیل می‌کند. [1] در اینجا، زبان نقش واسطی را ایفا می‌کند که مفاهیم کیفی مانند «کنار»، «دور از» یا «روی» را به محدودیت‌های کمی در فضای پیکربندی ربات ترجمه می‌کند.

این نوع ترجمه‌ی معنا به قیود کنترلی نشان می‌دهد که LLMها لزوماً نیازی به تولید مستقیم دستورات حرکتی ندارند. ارزش اصلی آن‌ها در ایجاد ساختار مفهومی برای مسئله است؛ ساختاری که ماژول‌های کلاسیک کنترل می‌توانند بر اساس آن عمل کنند. [8]

با وجود این پیشرفت‌ها، دستکاری فیزیکی مبتنی بر LLM همچنان با چالش‌های جدی مواجه است. عدم‌قطعیت در ادراک بصری، خطاهای مکانیکی و تفاوت‌های ظریف میان اشیای مشابه می‌توانند باعث شکست سیستم شوند. مثال‌های موجود در مقالات نشان می‌دهند که بسیاری از سامانه‌ها هنوز نیازمند محیط‌های نسبتاً کنترل‌شده یا نظارت انسانی هستند تا عملکرد قابل اعتمادی داشته باشند. [7] این واقعیت تأکید می‌کند که LLMها به‌تنهایی راه‌حل نهایی نیستند، بلکه بخشی از یک اکوسیستم پیچیده‌ی کنترلی‌اند.

فصل ۵: معماری‌های ماژولار، بازیابی دانش و ادغام با پشته‌های رباتیکی

با افزایش نقش مدل‌های زبانی بزرگ در برنامه‌ریزی و دستکاری، مسئله‌ی اساسی نحوه‌ی ادغام این مدل‌ها با سامانه‌های رباتیکی پیچیده و موجود مطرح می‌شود. استفاده‌ی مستقیم و یکپارچه از LLMها به‌عنوان یک جزء یکتا، اگرچه از نظر مفهومی جذاب است، اما در عمل با مشکلاتی مانند هزینه‌ی محاسباتی بالا، نبود تضمین ایمنی و دشواری اشکال‌زدایی همراه است. به همین دلیل معماری‌های ماژولار را به‌عنوان رویکرد غالب برای رباتیک مبتنی بر LLM معرفی میشود. [8] در این معماری‌ها، هر جزء مسئولیتی مشخص دارد و مدل زبانی تنها یکی از ماژول‌ها در یک زنجیره‌ی پردازشی بزرگ‌تر است.

ایده‌ی اصلی معماری ماژولار آن است که استدلال سطح‌بالا، که ذاتاً نمادین و احتمالی است، از کنترل سطح‌پایین، که نیازمند دقت و قابلیت پیش‌بینی بالاست، جدا شود. LLM ها در لایه‌های بالادستی قرار می‌گیرند و وظایفی مانند تفسیر زبان، تولید برنامه‌ی مفهومی و انتخاب مهارت مناسب را انجام می‌دهند، در حالی که ماژول‌های کنترلی کلاسیک مسئول اجرای دقیق این مهارت‌ها هستند. این تفکیک نه‌تنها ایمنی سیستم را افزایش می‌دهد، بلکه امکان جایگزینی یا به‌روزرسانی هر ماژول را بدون بازطراحی کل سامانه فراهم می‌کند. [2]

یکی از چالش‌های کلیدی در این معماری‌ها، محدودیت دانش درونی LLMهاست. هرچند این مدل‌ها بر داده‌های عظیمی آموزش دیده‌اند، اما دانش آن‌ها ثابت و وابسته به زمان آموزش است. برای غلبه بر این مشکل، رویکرد تولید مبتنی بر بازیابی[۱۳] مطرح شده است. در این روش، مدل زبانی به یک پایگاه دانش خارجی متصل می‌شود و پیش از تولید پاسخ یا برنامه، اطلاعات مرتبط را بازیابی میکند. [8] این پایگاه دانش می‌تواند شامل مستندات فنی ربات، نقشه‌ی محیط، محدودیت‌های ایمنی یا تجربه‌های پیشین باشد.

پیاده‌سازی عملی این ایده را می‌توان در سامانه‌ی ARRC مشاهده کرد .ARRC یک معماری رباتیکی ماژولار است که در آن زبان طبیعی به‌عنوان رابط اصلی میان کاربر و ربات عمل می‌کند، اما تصمیم‌گیری نهایی از طریق چندین لایه‌ی بررسی و اعتبارسنجی انجام می‌شود. [8] در این سامانه، دستور زبانی کاربر ابتدا توسط مدل زبانی تحلیل می‌شود، سپس اطلاعات مرتبط از پایگاه دانش بازیابی می‌گردد و در نهایت یک خروجی ساخت‌یافته تولید می‌شود که برای ماژول‌های کنترلی قابل فهم است. (*تصویر 2*)

نکته‌ی مهم در ARRC آن است که خروجی LLM به‌صورت متن آزاد مستقیماً اجرا نمی‌شود. به‌جای آن، مدل موظف است خروجی خود را در قالبی مشخص و محدود ارائه دهد؛ قالبی که تنها شامل اعمال مجاز و پارامترهای قابل کنترل است. این محدودسازی نقش مهمی در افزایش ایمنی و قابلیت اطمینان سیستم دارد و به‌طور مستقیم با مهار رفتارهای غیرقابل پیش‌بینی ها همسو است. [5]

ادغام با پشته‌های رباتیکی موجود، به‌ویژه ROS ، یکی دیگر از مزایای معماری‌های ماژولار است. ROS به‌عنوان یک چارچوب استاندارد، ابزارها و پروتکل‌های ارتباطی گسترده‌ای را برای رباتیک فراهم می‌کند. سامانه‌هایی مانند ROS-LLM نشان داده‌اند که می‌توان LLMها را به‌عنوان یک گره‌ی اضافی به این پشته افزود، بدون آنکه ساختار کلی سیستم تغییر کند. [2] در این رویکرد، مدل زبانی دستورات زبانی را به فراخوانی سرویس‌ها یا اکشن‌های موجود در ROS نگاشت می‌کند و بدین ترتیب، قابلیت‌های جدیدی به ربات افزوده می‌شود.

از منظر مهندسی، این نوع ادغام مزایای قابل توجهی دارد. نخست آنکه توسعه‌دهندگان می‌توانند از سرمایه‌گذاری‌های قبلی خود در طراحی کنترل‌کننده‌ها و برنامه‌ریزهای کلاسیک بهره ببرند. دوم آنکه اشکال‌زدایی و ارزیابی سیستم ساده‌تر می‌شود، زیرا هر ماژول رفتاری مشخص و قابل اندازه‌گیری دارد. این شفافیت معماری برای پذیرش صنعتی رباتیک مبتنی بر LLM حیاتی است. [8]

با وجود این مزایا، معماری‌های ماژولار بدون چالش نیستند. هماهنگی میان ماژول‌ها، تأخیر ناشی از فراخوانی مدل‌های بزرگ و مدیریت خطا در مرزهای ماژول‌ها از جمله مسائلی هستند که در مثال‌های عملی نیز دیده می‌شوند. [8] این مسائل نشان می‌دهند که طراحی معماری به‌اندازه‌ی انتخاب مدل زبانی اهمیت دارد و موفقیت یک سامانه‌ی رباتیکی بیش از آنکه به قدرت خام LLM وابسته باشد، به نحوه‌ی استفاده‌ی مهندسی‌شده از آن بستگی دارد.

[13] Retrieval-Augmented Generation

تصویر 2. معماری در سطح بالا: ماژول ادراک، مشاهداتِ شیءمحور تولید می‌کند. برنامه‌ریز RAG دانشِ مربوط به کار را بازیابی کرده و یک طرح JSON می‌سازد. اجراکننده نیز با بررسی‌های ایمنی، دستورها را از طریق XArm SDK اعتبارسنجی و اجرا می‌کند. [8]

فصل ۶: ایمنی، قابلیت اطمینان و تعامل انسان در حلقه

با ورود مدل‌های زبانی بزرگ به قلب سامانه‌های رباتیکی، مسئله‌ی ایمنی و قابلیت اطمینان به یکی از مهم‌ترین دغدغه‌های پژوهشی و مهندسی تبدیل شده است. برخلاف الگوریتم‌های کنترلی کلاسیک که رفتار آن‌ها در چارچوب مدل‌های ریاضی نسبتاً قابل پیش‌بینی است، LLMها ذاتاً سامانه‌هایی احتمالی هستند که خروجی آن‌ها می‌تواند بسته به زمینه، داده‌های آموزشی و حتی جزئیات ظریف ورودی تغییر کند. از یک سو، منبع انعطاف‌پذیری و تعمیم‌پذیری بالاست و از سوی دیگر، خطری بالقوه برای ایمنی ربات در محیط‌های واقعی.

یکی از چالش‌های اصلی، پدیده‌ی «توهم زبانی[14]» است؛ حالتی که در آن مدل زبانی پاسخی ظاهراً منسجم اما نادرست یا غیرقابل اجرا تولید می‌کند. در زمینه‌ی رباتیک، چنین خطایی می‌تواند به اعمال خطرناک یا آسیب‌زا منجر شود. به همین دلیل، پژوهش‌ها بر این نکته تأکید دارند که خروجی LLM نباید مستقیماً و بدون بررسی اجرا شود. در عوض، معماری‌های ایمن از لایه‌های اعتبارسنجی استفاده می‌کنند که خروجی مدل را از نظر سازگاری با محدودیت‌های فیزیکی، قوانین ایمنی و اهداف تعریف‌شده بررسی می‌کنند. [9]

نمونه‌های عملی متعددی از این رویکرد در مقالات دیده می‌شود. برای مثال، در سامانه‌های مبتنی برROS-LLM ، دستورات زبانی ابتدا به برنامه‌های ساخت‌یافته تبدیل می‌شوند و سپس توسط ماژول‌های کنترلی بررسی می‌گردند تا از مجاز بودن اعمال اطمینان حاصل شود. [2] اگر برنامه‌ی تولیدشده با محدودیت‌های ایمنی در تضاد باشد، یا اجرا نمی‌شود یا به کاربر بازخورد داده می‌شود. این فرایند نشان می‌دهد که ایمنی نه در یک نقطه‌ی خاص، بلکه در کل زنجیره‌ی تصمیم‌گیری توزیع شده است.

تعامل انسان در حلقه یکی دیگر از راهکارهای کلیدی برای افزایش قابلیت اطمینان است. برخلاف دیدگاه‌های اولیه که استقلال کامل ربات را هدف نهایی می‌دانستند، رویکردهای جدیدتر انسان را به‌عنوان بخشی فعال از سامانه در نظر می‌گیرند. در این چارچوب، انسان می‌تواند در مراحل حساس تصمیم‌گیری مداخله کند، خروجی مدل را تأیید یا اصلاح نماید و حتی با زبان طبیعی رفتار ربات را هدایت کند. [2]

سامانه‌ی VoxPoser نمونه‌ی روشنی از این رویکرد است. در VoxPoser، کاربر می‌تواند در طول اجرای وظیفه با زبان طبیعی قیود جدیدی اضافه کند یا مسیر اجرای ربات را اصلاح نماید. [1] برای مثال، اگر ربات در حال قرار دادن یک شیء در موقعیتی نامناسب باشد، کاربر می‌تواند بگوید «کمی دورتر از لبه‌ی میز بگذار» و سیستم این اصلاح را در برنامه‌ی کنترلی اعمال می‌کند. این تعامل بلادرنگ، هم ایمنی را افزایش می‌دهد و هم حس کنترل و اعتماد کاربر را تقویت می‌کند.

بازخورد زبانی انسان همچنین به‌عنوان ابزاری برای یادگیری و بهبود رفتار ربات عمل می‌کند. در برخی سامانه‌ها، اصلاحات کاربر ذخیره می‌شود و در تصمیم‌گیری‌های بعدی مورد استفاده قرار می‌گیرد. [2] این فرایند نشان می‌دهد که زبان می‌تواند نه‌تنها ابزار فرمان‌دهی، بلکه رسانه‌ای برای انتقال تجربه و دانش ایمنی باشد. [1]

با این حال، اتکا به انسان در حلقه محدودیت‌هایی نیز دارد. افزایش بار شناختی کاربر، تأخیر در تصمیم‌گیری و وابستگی بیش از حد به نظارت انسانی از جمله چالش‌هایی هستند که در مثال‌های عملی گزارش شده‌اند. [1] این مسائل نشان می‌دهند که طراحی تعامل انسان–ربات باید به‌دقت انجام شود تا توازن مناسبی میان استقلال و نظارت برقرار گردد.

فصل ۷: محدودیت‌ها، چالش‌های باز و مسیرهای آینده

با وجود پیشرفت‌های چشمگیر در رباتیک مبتنی بر مدل‌های زبانی بزرگ، بررسی انتقادی این رویکرد نشان می‌دهد که فاصله‌ی معناداری میان قابلیت‌های فعلی و چشم‌انداز «ربات‌های عمومی و مستقل» وجود دارد. LLMها را نباید به‌عنوان راه‌حلی نهایی، بلکه به‌عنوان یک ابزار قدرتمند اما ناقص در نظر گرفت. [1] این فصل به جمع‌بندی محدودیت‌های اساسی، چالش‌های پژوهشی باز و مسیرهای محتمل آینده می‌پردازد و مثال‌های عملی را برای عینیت‌بخشی این بحث به کار می‌گیرد.

[14] linguistic Hallucination

یکی از بنیادی‌ترین محدودیت‌ها، عدم تضمین صحت و سازگاری خروجی‌های زبانی است. همان‌طور که در فصل‌های پیش اشاره شد، LLMها ممکن است پاسخ‌هایی تولید کنند که از نظر زبانی قانع‌کننده اما از نظر فیزیکی یا منطقی نادرست باشند. این مسئله در سامانه‌های رباتیکی، که هر خطا می‌تواند پیامدهای فیزیکی داشته باشد، اهمیت دوچندان پیدا می‌کند. گزارش می‌کند که در برخی پیاده‌سازی‌های صنعتی، خروجی‌های مدل زبانی تنها پس از فیلترهای سخت‌گیرانه و آزمون‌های متعدد اجازه‌ی اجرا پیدا می‌کنند، و حتی در این شرایط نیز نظارت انسانی ضروری باقی می‌ماند. [6] این واقعیت نشان می‌دهد که اتکای کامل به LLMها در محیط‌های ایمنی‌حساس هنوز عملی نیست.

چالش مهم دیگر، تعمیم‌پذیری در دنیای واقعی است. بسیاری از مثال‌های موفق، مانند SayCan یا RT-Grasp، در محیط‌های نسبتاً کنترل‌شده یا با مجموعه‌ای محدود از اشیا آزمایش شده‌اند. [7] [1] در مقابل، محیط‌های واقعی سرشار از تغییرات پیش‌بینی‌نشده، نویز حسی و تعاملات پیچیده‌اند. اگرچه LLMها از نظر زبانی تعمیم‌پذیری بالایی دارند، اما این ویژگی لزوماً به تعمیم فیزیکی منجر نمی‌شود. [6] شکاف میان «دانستن» و «انجام دادن» همچنان یکی از چالش‌های باز در هوش تجسم‌یافته است.

مسئله‌ی هزینه‌های محاسباتی و زیرساختی نیز نباید نادیده گرفته شود. اجرای مدل‌های زبانی بزرگ، به‌ویژه در سناریوهای بلادرنگ رباتیکی، نیازمند منابع پردازشی قابل توجه است. این موضوع در مقالات به‌ویژه در کاربردهای صنعتی برجسته شده است، جایی که محدودیت‌های انرژی، تأخیر و قابلیت اطمینان سخت‌افزار نقش تعیین‌کننده دارند. [6] در نتیجه، بسیاری از سامانه‌های عملی به استفاده از مدل‌های کوچک‌تر، نسخه‌های فشرده یا اجرای ابری روی آورده‌اند که هر یک مصالحه‌هایی در زمینه‌ی تأخیر و حریم خصوصی به همراه دارد.

از منظر پژوهشی، یکی از چالش‌های باز مهم، هم‌ترازی اهداف انسانی با رفتار ربات است. اگرچه زبان ابزاری قدرتمند برای بیان نیت انسانی است، اما تضمین اینکه تفسیر مدل از این نیت با انتظارات کاربر همخوانی داشته باشد، همچنان دشوار است. بازخورد زبانی انسان، نظارت در حلقه و محدودسازی خروجی‌ها تنها راهکارهای موقتی‌اند. [9] مثال‌های عملی نیز نشان می‌دهند که بدون چنین سازوکارهایی، حتی سامانه‌های پیشرفته ممکن است رفتاری غیرمنتظره از خود نشان دهند. [2]

با این حال، مسیرهای آینده‌ی روشنی نیز قابل ترسیم است. یکی از این مسیرها، توسعه‌ی مدل‌های چندوجهی قوی‌تر است که زبان، بینایی، لمس و حالت فیزیکی را به‌صورت یکپارچه پردازش می‌کنند. نمونه‌هایی مانند PaLM-E نشان داده‌اند که این ادغام می‌تواند شکاف میان ادراک و استدلال را کاهش دهد. [1] انتظار می‌رود که با پیشرفت این مدل‌ها، ربات‌ها بتوانند درک غنی‌تری از محیط و پیامدهای اعمال خود داشته باشند.

مسیر دیگر، حرکت به‌سوی معماری‌های عامل‌محور و سلسله‌مراتبی است که در آن‌ها چندین عامل زبانی و کنترلی با سطوح مختلف انتزاع همکاری می‌کنند. چنین معماری‌هایی می‌توانند پیچیدگی تصمیم‌گیری را بهتر مدیریت کنند و در عین حال، ایمنی و قابلیت اطمینان را حفظ نمایند [8]، مثال‌های موجود در مقالات، هرچند در مقیاس محدود، نشان می‌دهند که این ایده از نظر عملی قابل پیاده‌سازی است.

نتیجه‌گیری

مطالب ارائه‌شده در این گزارش نشان می‌دهد که مدل‌های زبانی بزرگ، اگرچه به‌تنهایی راه‌حل نهایی برای تمامی مسائل رباتیک نیستند، اما نقشی کلیدی در تحول معماری‌ها و روش‌های تعامل انسان و ربات ایفا می‌کنند. بررسی فصل‌های مختلف نشان داد که قدرت اصلی این مدل‌ها در توانایی درک زبان طبیعی، استخراج نیت انسانی، برنامه‌ریزی سطح‌بالا و ایجاد پیوند میان معنا و عمل نهفته است. این قابلیت‌ها باعث شده‌اند که ربات‌ها از ماشین‌هایی صرفاً فرمان‌پذیر به سامانه‌هایی با سطحی از استدلال و انعطاف‌پذیری شناختی تبدیل شوند.

با این حال، نتایج فصول مختلف به‌روشنی نشان می‌دهد که استفاده‌ی مستقیم و بدون واسطه از خروجی‌های مدل‌های زبانی در کنترل ربات‌ها می‌تواند خطرناک و غیرقابل اعتماد باشد. ماهیت احتمالی این مدل‌ها، پدیده‌هایی مانند توهم زبانی و نبود تضمین رسمی در صحت خروجی‌ها، ضرورت استفاده از معماری‌های ترکیبی و ماژولار را برجسته می‌کند. در چنین معماری‌هایی، مدل زبانی نقش راهنما و تصمیم‌ساز سطح‌بالا را بر عهده دارد، در حالی که کنترل دقیق و ایمن به ماژول‌های کنترلی کلاسیک سپرده می‌شود.

از منظر ایمنی و قابلیت اطمینان، حضور انسان در حلقه‌ی تصمیم‌گیری همچنان اهمیت بالایی دارد. تعامل زبانی پیوسته، امکان اصلاح مسیر اجرا و ارائه‌ی بازخورد، نه‌تنها ایمنی سیستم را افزایش می‌دهد، بلکه اعتماد کاربر به ربات را نیز تقویت می‌کند. با این وجود، وابستگی بیش از حد به نظارت انسانی می‌تواند مانعی برای مقیاس‌پذیری و خودمختاری کامل ربات‌ها باشد؛ ازاین‌رو، ایجاد توازن میان استقلال ربات و نظارت انسان یکی از چالش‌های اصلی آینده محسوب می‌شود.

در جمع‌بندی می‌توان گفت که رباتیک مبتنی بر مدل‌های زبانی بزرگ، مسیری نویدبخش اما همراه با چالش‌های فنی، ایمنی و مهندسی است. آینده‌ی این حوزه به توسعه‌ی مدل‌های چندوجهی قوی‌تر، معماری‌های سلسله‌مراتبی و روش‌های اعتبارسنجی دقیق‌تر وابسته است. اگر این چالش‌ها به‌درستی مدیریت شوند، می‌توان انتظار داشت که ربات‌ها در آینده‌ای نه‌چندان دور، تعامل طبیعی‌تر، درک عمیق‌تر و عملکرد قابل‌اعتمادتری در محیط‌های واقعی داشته باشند و به ابزارهایی مؤثر در زندگی روزمره و صنعت تبدیل شوند.

منابع

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

[2] J. Wang, “ROS-LLM: A ROS framework for embodied AI with task feedback and structured reasoning,” و E. M. Christopher, W. Yuhui, Hongzhan, G. Antoine, J. Gonzalez-Billandon, M. Zimmer p. 26, 2024.

[3] Z. Xiaoli, “A Review of Natural-Language-Instructed Robot,” و L. Rui, G. Yibei, J. Runxiang p. 42, 2024.

[4] H. Andreas, “Domain-Specific Fine-Tuning of Large Language,” و A. Benjamin, K. Urs, T. Aleksandar, K. Darko p. 5, 2021.

[5] C. Robin, “ARRC: Advanced Reasoning Robot Control—Knowledge-Driven Autonomous Manipulation Using Retrieval-Augmented Generation,” و V. Eugene, M. Ammar Jaleel, . R. Salim p. 8, 2025.

[6] Francesco, “Parsing Natural Language Sentences into Robot,” و Danilo, M. Enrico B. Gianluca, . D. O. p. 4, 2019.

[7] Z. Liangjun, “RT-Grasp: Reasoning Tuning Robotic Grasping via,” و X. Jinxuan, J. Shiyu, L. Yutian, Z. Yuqian p. 8, 2024.

[8] P. S. Yu, “Large Language Models for Robotics: A Survey,” و F. Zeng, W. Gan, Y. Wang, N. Liu p. 19, 2023.

[9] Z. Yilun, “A Survey of LLM-Driven AI Agent Communication: Protocols,,” و Zhebo, L. Minghao, L. Yufeng K. Dezhang, L. Shi , . X. Zhenhua, W. p. 35, 2025.

[10] Z. Wenxiao, “Enhancing reliability in LLM-integrated robotic systems: A unified approach,” و D. Conan, . B. Thomas, B. H. Jin p. 14, 2025.

[11] D. z. Michał, “InCoRo: In-Context Learning for Robotics Control with Feedback Loops,” و Y. Z. Jiaqiang , G. C. Carla, V. David p. 20, 2024.

[12] T. Jie, “Large Language Models for Manufacturing,” و L. Yiwei, Z. Huaqin, J. Hanqi, P. Yi, L. Zhengliang , W. Zihao, S. Peng . p. 52, 2024.

[13] p. 10, 2023. T. Dzmitry, “LLM-BRAIn: AI-driven Fast Generation of Robot,” و L. Artem

[14] A. Rumaisa, “LLM-Driven Robots Risk Enacting Discrimination, Violence, and,” و H. Andrew p. 59, 2025.

[15] M. George, “Multi-Agent Systems for و Ziqi, . G. X. Haoyuan, Z. Dandan C. Junhong, Y. p. 11, 2024. Robotic Autonomy with LLMs,”

[16] Z. Shuai, “An LLM- و ZhenDong, N. ZhanShang, W. ShiXing, L. JunYi, C. YongTian C. p. 8, 2025. powered Natural-to-Robotic Language,”

[17] T. Stefanie, “Plug in the Safety Chip: Enforcing Constraints و Y. Ziyi, S. R. Shreyas, S. Ankit p. 15, 2023. for LLM-driven Robot,”

[18] B. Rogerio, “Reshaping Robot و B. Arthur, F. Luis, . H. Sami, K. Ashish, . M. Shuang p. 7, 2022. Trajectories Using Natural Language Commands:,”

[19] R. Nicholas, . و T. Stefanie, K. Thomas, D. Steven, R. W. Matthew, G. B. Ashis, T. Seth p. 8, 2011. “Understanding Natural Language Commands,”

[20] F. Yourong , “A Multimodal LLM- و D. Wenhao, Z. Xingting, . W. Pan, Z. Jingwei, S. Yutong . p. 22, 2025. Driven Robotic Control System,”