

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

۱۴۰۴ آذر

فهرست

3.....	چکیده.....
4.....	فصل 1
4.....	مقدمه.....
6.....	فصل 2
6.....	رویکردهای استفاده از مدل‌های زبانی در رباتیک
6.....	1-2 مقدمه.....
6.....	2-2 تغییر پارادایم در رباتیک مبتنی بر مدل‌های زبانی بزرگ
8.....	3-2 در ک ک زبان طبیعی و تعامل انسان-ربات به عنوان زیربنای شناختی
10.....	4-2 برنامه‌ریزی وظیفه و اجرای بلندمدت با هدایت زبان.....
11.....	5-2 دستکاری فیزیکی، استدلال چندوجهی و پیوند معنا با کنترل.....
13.....	6-2 معماری‌های مازولار، بازیابی دانش و ادغام با پشته‌های رباتیکی
15.....	7-2 اینمی، قابلیت اطمینان و تعامل انسان در حلقه
17.....	8-2 محدودیت‌ها، چالش‌های باز و مسیرهای آینده
18.....	فصل 3
18.....	نتیجه‌گیری
19.....	منابع

چکیده

مدل‌های زبانی بزرگ^۱ 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 جندوچهی استدلال کنند و مهارت‌های دستکاری^۲ اشیا را در محیط‌های پویا اجرا کنند. با حرکت رباتیک به سمت سطوح بالاتری از خودمحختاری 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 انسان‌ربات را یکپارچه می‌کند.

این گزارش 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 چندوچهی 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 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.

¹ Large language model

² manipulation

فصل ۱

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 عمده‌ای بر پایه‌ی برنامه‌نویسی صریح، قوانین از پیش تعريف شده و مدل‌های کنترلی سخت‌گیرانه طراحی می‌شوند. این رویکرد اگرچه در محیط‌های کنترل شده عملکرد قابل قبولی داشت، اما در مواجهه با محیط‌های پویا، دستورات مبهم انسانی و وظایف پیچیده و چندمرحله‌ای با محدودیت‌های جدی روبرو بود.

در سال‌های اخیر، مدل‌های زبانی بزرگ به‌عنوان یک ابزار شناختی قدرتمند، امکان تعامل طبیعی تر میان انسان و ماشین را فراهم کرده‌اند. این مدل‌ها قادرند زبان طبیعی را نه تنها به‌عنوان یک ورودی متنی، بلکه به‌عنوان بستری برای استدلال، برنامه‌ریزی و تصمیم‌گیری به کار 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 به صورت گام به گام 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 نظری هوش تجسم یافته 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 سطح بالا 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 مژولار و شیوه‌های ادغام مدل‌های زبانی با پشته‌های رباتیکی موجود مانند ROS معرفی می‌شوند. سپس در فصل ششم، مسائل اینمنی، قابلیت اطمینان و نقش انسان در حلقه‌ی تصمیم‌گیری بررسی شده و اهمیت طراحی سازوکارهای کنترلی و نظارتی بر جسته می‌گردد. در نهایت، فصل هفتم به محدودیت‌ها، چالش‌های باز و مسیرهای آینده‌ی این حوزه اختصاص دارد.

هدف از ارائه‌ی این گزارش، فراهم کردن دیدی جامع و تحلیلی نسبت به نقش مدل‌های زبانی بزرگ در سامانه‌های رباتیکی و بررسی مزایا، چالش‌ها و چشم‌اندازهای پیش‌رو است؛ به گونه‌ای که خواننده بتواند تصویری روشن از وضعیت فعلی و مسیر توسعه‌ی این فناوری به دست آورد.

فصل 2

رویکردهای استفاده از مدل‌های زبانی در رباتیک

1-2 مقدمه

2-2 تغییر پارادایم در رباتیک مبتنی بر مدل‌های زبانی بزرگ

در دهه‌های گذشته، رباتیک عمدهاً بر پایه‌ی برنامه‌نویسی صریح، مدل‌های نمادین سخت‌گیرانه و زنجیره‌های کنترلی از پیش تعریف شده توسعه یافته است. در این رویکرد کلاسیک، هر رفتار ربات باید به صورت مستقیم و با جزئیات دقیق توسط مهندسان طراحی می‌شد؛ از تعریف سناریوهای ممکن گرفته تا مدیریت خطاهای و تعامل با انسان. این شیوه اگرچه در محیط‌های صنعتی کنترل شده موفق بوده است، اما در مواجهه با محیط‌های پویا، دستورات مبهم انسانی و ظایف چند مرحله‌ای پیچیده به سرعت به بن‌بست می‌رسد. محدودیت اصلی این نمونه ناتوانی آن در درک و استفاده از زبان طبیعی به عنوان یک رابط عمومی شناختی میان انسان و ربات است.

ظهور مدل‌های زبانی بزرگ³ این وضعیت را به طور بنیادین تغییر داده است. مدل‌های زبانی بزرگ که بر حجم عظیمی از داده‌های متنه آموزش دیده‌اند، قادرند ساختارهای زبانی پیچیده، روابط معنایی پنهان و حتی نیت ضمنی انسان را استخراج کنند. زبان طبیعی از یک ابزار جانی به یک مؤلفه‌ی مرکزی در ادراک، تصمیم‌گیری، برنامه‌ریزی و کنترل ربات تبدیل می‌شود^[1]. در این چارچوب جدید، ربات دیگر صرفاً یک ماشین اجرا کننده‌ی دستورات از پیش تعریف شده 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 تجسم یافته»⁴ است. هوش تجسم یافته⁴ 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 چندوجهی⁵ و کنش فیزیکی، امکان شکل‌گیری چنین هوشی را فراهم می‌کنند. در سامانه‌های مدرن، مدل‌های زبانی بزرگ به عنوان 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].

³ Large Language Models

⁴ Embodied intelligence

⁵ Multimodal perception

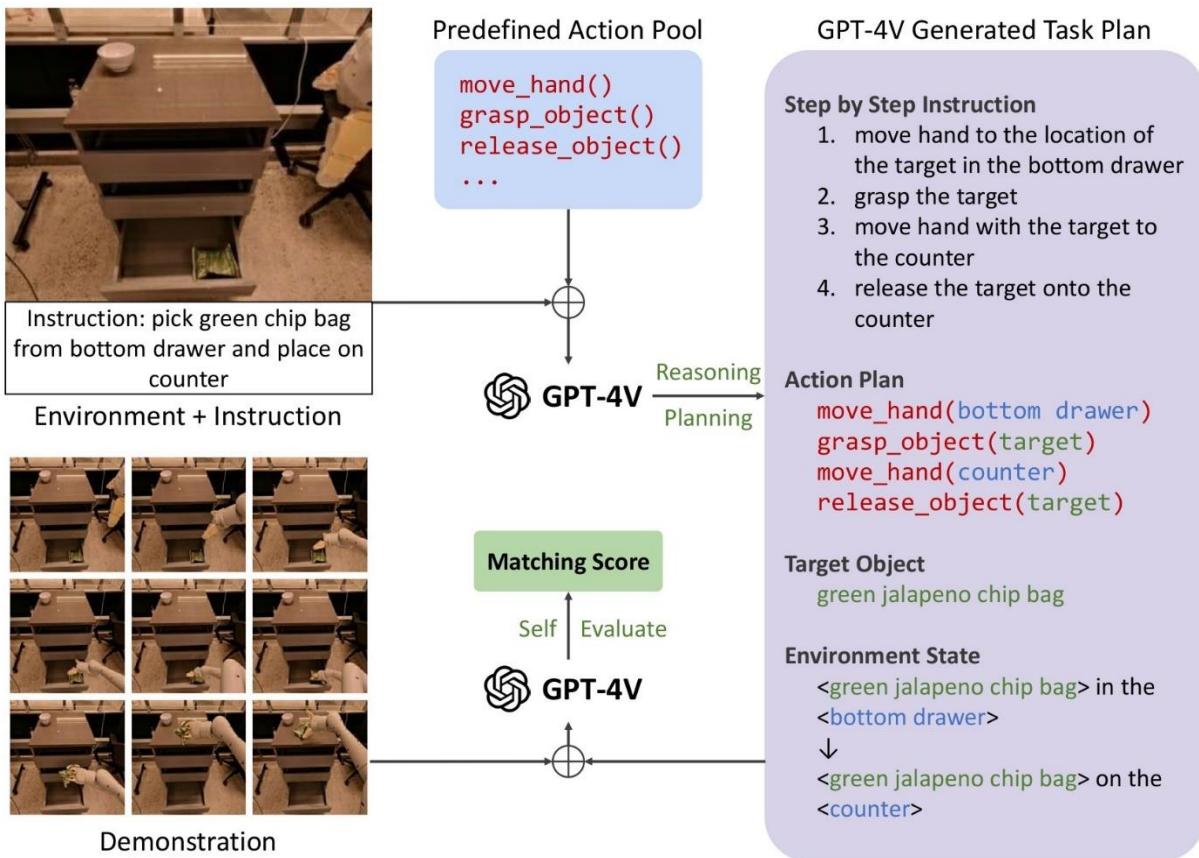
اهمیت این تغییر مدل زمانی روش‌نتر می‌شود که به نمونه‌های عملی نگاه کنیم. برای مثال، سامانه‌ی SayCan که توسط گوگل توسعه داده شده است، نشان می‌دهد چگونه می‌توان استدلال زبانی سطح بالا را با مهارت‌های کنترلی از پیش‌آموخته‌ی ربات ترکیب کرد. در این سامانه، ربات دستورات باز و غیرساخت یافته‌ای مانند «یک نوشیدنی 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 به تنها‌یی به رابط اصلی تعامل تبدیل شده است.

نمونه‌ی برجسته‌ی دیگر، PaLM-E است که گامی فراتر از برنامه‌ریزی وظیفه بر می‌دارد و مدل زبانی را مستقیماً در حلقه‌ی ادراک و تصمیم‌گیری ربات قرار می‌دهد. در این سامانه، مدل زبانی نه تنها متن، بلکه تصاویر و بردارهای حالت ربات را نیز به عنوان ورودی دریافت می‌کند و می‌تواند به پرسش‌هایی درباره‌ی محیط پاسخ دهد یا تصمیم‌های کنترلی اتخاذ کند. [1] این رویکرد نشان می‌دهد که مدل‌های زبانی بزرگ می‌توانند از یک «مولد برنامه»⁶ به یک «مغز تجسم یافته»⁷ برای ربات تبدیل شوند؛ مغزی که به‌طور پیوسته با محیط فیزیکی در تعامل است. تصویر 1)

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 نوشتمن کدهای کنترلی محدود شود، به طراحی معماری‌هایی می‌پردازد که بتوانند خروجی‌های احتمالی و انعطاف‌پذیر مدل‌های زبانی بزرگ را به رفتارهای ایمن و قابل اعتماد تبدیل کنند. [1]

⁶ Program generator

⁷ The embodied brain



تصویر ۱. 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 برمی‌گیرند. 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 مرجح مقایسه کرده و امتیاز می‌دهد. [1]

2-3 درک زبان طبیعی و تعامل انسان-ربات به عنوان زیربنای شناختی

درک زبان طبیعی سنگ بنای رباتیک مبتنی بر مدل‌های زبانی بزرگ است. اگر ربات نتواند زبان انسان را به درستی 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 پیش تعریف شده محدود بود. این سامانه‌ها معمولاً بر هستی‌شناسی‌های⁸ دستی، وابستگی‌شناسی نحوی⁹ و استخراج افعال و اشیای صریح تکیه داشتند و تنها در دامنه‌های بسیار محدود عملکرد قابل قبولی نشان می‌دادند. [3] نتیجه‌ی

⁸ Based Ontologies

⁹ Syntactic dependency theory

چنین رویکردی، سامانه‌هایی شکننده بود که در مواجهه با ابهام، بیان غیرمستقیم یا تغییرات جزئی در جمله به سرعت دچار خطا می‌شدند.

[4]

مدل‌های زبانی بزرگ این وضعیت را به طور اساسی تغییر داده‌اند. LLM‌ها به جای تکیه بر قواعد صریح، از بازنمایی‌های معنایی غنی استفاده می‌کنند که از داده‌های عظیم متی آموخته شده‌اند. این بازنمایی‌ها به مدل اجازه می‌دهند روابط ضمنی، دانش عمومی و زمینه‌ی گفتار را در نظر بگیرد. به عنوان مثال، یک مدل زبانی بزرگ می‌تواند از جمله‌ای مانند «هوا سرد است» نیت ضمنی «پنجره را بیند» یا «بخاری را روشن کن» را استنتاج کند، بدون آنکه این عمل به طور 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، مدل‌های زبانی بزرگ نقش مهمی در گفت‌و‌گوی تعاملی ایفا می‌کنند. در سامانه‌های مدرن، اگر دستور کاربر ناقص یا ناسازگار با وضعیت فیزیکی ربات باشد، مدل می‌تواند به جای شکست خاموش، پرسش تکمیلی مطرح کند یا گزینه‌های ممکن را پیشنهاد دهد. این رفتار گفت‌و‌گو محور، زبان را از یک کانال یک‌طرفه فرماندهی به یک ابزار مذاکره و اصلاح مشترک میان انسان و ربات تبدیل می‌کند. [2] در نتیجه، تعامل انسان–ربات به جای آنکه شبیه برنامه‌نویسی باشد، به مکالمه‌ای هدف محور 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» مانند Zora یا NAO پیاده‌سازی شده است. در این سامانه، دستورات گفتاری کاربر مانند «دست 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 به‌نهایی برای هدایت ربات کافی است. این مثال نشان می‌دهد که چگونه حتی پیش از ظهور مدل‌های زبانی بزرگی مدرن، زبان به عنوان رابط کنترل مطرح بوده، اما با دامنه‌ای محدود و نیازمند مهندسی دستی گسترده.

توانایی رفع ابهام نیز در همین سامانه به صورت عملی نمایش داده شده است. زمانی که کاربر دستور ناقصی مانند «دست را بالا ببر» صادر می‌کند، ربات به طور فعلی 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 در ک 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 در ک زبان طبیعی، آگاهی از وضعیت درونی و فیزیکی ربات است. در همان سامانه‌ی انسان‌نمای، دو حالت اجرای «stateless» و «stateful» 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 به‌نهایی 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.

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 به صورت ضمنی از داده‌های آموزشی یا به صورت صریح از طریق لایه‌های ایمنی به مدل تزریق شود. [5]

4-2 برنامه‌ریزی وظیفه و اجرای بلندمدت با هدایت زبان

پس از آنکه زبان طبیعی به عنوان یک رابط شناختی قابل اعتماد میان انسان و ربات ثبت شد، گام منطقی بعدی استفاده از این توانمندی برای برنامه‌ریزی وظایف پیچیده و چندمرحله‌ای است. برنامه‌ریزی وظیفه¹⁰ در رباتیک به معنای تبدیل یک هدف سطح بالا، که اغلب به صورت 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 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. مدل‌های زبانی بزرگ 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 سطح بالا»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 سطح بالا را بر عهده دارد و اجرای واقعی توسط مأذول‌های کنترلی قابل اعتماد انجام می‌شود. این رویکرد ترکیبی، که گاه با عنوان «برنامه‌ریزی گراندشه»¹² شناخته می‌شود، یکی از محورهای اصلی پژوهش در رباتیک مبتنی بر LLM. [2]

¹⁰ Task programming

¹¹ High level programming

¹² Grounded planning

پیاده‌سازی عملی این ایده به طور 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 بزنند. این اقدامات از پیش به صورت مهارت‌های قابل اجرا روی ریات 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 بدست 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 مدل‌های زبانی بزرگ به جای جایگزینی کامل برنامه‌ریزی کلاسیک، می‌توانند آن را تکمیل کنند. در SayCan، زبان به عنوان راهنمای انتخاب در فضای اعمال عمل می‌کند، نه به عنوان کنترل کننده‌ی 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 مراحل به طور صریح در دستور ذکر نشده باشند. [1] این توانایی استخراج ساختار وظیفه از زبان طبیعی دقیقاً همان چیزی است که در برنامه‌ریزی کلاسیک به سختی و با هزینه‌ی مهندسی بالا بدست می‌آید.

نمونه‌ی دیگر از این رویکرد، سامانه‌های مبتنی بر ROS-LLM هستند که در آن‌ها مدل زبانی بزرگ به عنوان یک لایه‌ی برنامه‌ریز روی پشتی 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. به علاوه، ROS-LLM نشان می‌دهد که چگونه مدل‌های زبانی بزرگ می‌توانند با ابزارهای مهندسی جافتاده ادغام شوند، نه اینکه به عنوان 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. مدل‌های زبانی بزرگ به دلیل توانایی نگهداری زمینه و باز تفسیر 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، به جای 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 به عنوان ابزار استدلال و هدایت به کار می‌رود، اما کنترل نهایی همچنان در چارچوب‌های ایمن و قابل پیش‌بینی انجام می‌شود.

5-2 دستکاری فیزیکی، استدلال چندوجهی و پیوند معنا با کنترل

اگرچه در فصل پیش نشان داده شد که مدل‌های زبانی بزرگ می‌توانند اهداف سطح بالا 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 عدم قطعیت‌های محیطی است. به طور 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، به تهایی قادر به حل مستقیم مسائل کنترل پیوسته نیستند، اما می‌توانند نقش مهمی در راهنمایی معنایی فرایند دستکاری ایفا کنند. ایده‌ی اصلی آن است که مدل‌های زبانی بزرگ دانش سطح بالا 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 به راحتی در زبان بیان می‌شود اما به سختی در قالب قواعد هندسی صریح مدل‌سازی می‌شود. [7]

این رویکرد به ویژه در معماری‌های چندوجهی اهمیت می‌یابد. در چنین سامانه‌هایی، زبان، بینایی و حالت فیزیکی ربات به صورت هم‌زمان در تصمیم‌گیری دخیل هستند. مدل‌های زبانی بزرگ در این میان نقش ادغام کننده‌ی 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 به طور 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 اینمی و هم قابلیت تعیین سیستم را افزایش دهد. به جای آنکه مدل کنترل عددی مجبور باشد تمامی حالات ممکن را از داده بیاموزد، بخشی از دانش به صورت زبانی و قابل تفسیر در اختیار سیستم قرار می‌گیرد. این موضوع به ویژه در محیط‌های جدید یا اشیای نادیده‌شده اهمیت دارد، جایی که داده‌ی آموزشی محدود است. [7] [6]

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 لبه 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 مدل‌های زبانی بزرگ لزوماً نیازی به تولید مستقیم دستورات حرکتی ندارند. ارزش اصلی آن‌ها در ایجاد ساختار مفهومی برای مسئله است؛ ساختاری که مژوول‌های کلاسیک کنترل می‌توانند بر اساس آن عمل کنند. [8]

با وجود این پیشرفت‌ها، دستکاری فیزیکی مبتنی بر LLM همچنان با چالش‌های جدی مواجه است. عدم قطعیت 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 کنترل شده یا نظارت انسانی هستند تا عملکرد قابل اعتمادی داشته باشند. [7] این واقعیت تأکید می‌کند که مدل‌های زبانی بزرگ به تنها‌ی راه حل نهایی نیستند، بلکه بخشی از یک اکوسیستم پیچیده کنترلی‌اند.

2-6 معماری‌های مژوولار، بازیابی دانش و ادغام با پشتهمای رباتیکی

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

ایده‌ی اصلی معماری مژوولار آن است که استدلال سطح‌بالا، که ذاتاً نمادین و احتمالی است، از کنترل سطح‌پایین، که نیازمند دقت و قابلیت پیش‌بینی بالاست، جدا شود. مدل‌های زبانی بزرگ در لایه‌های بالادستی قرار می‌گیرند و وظایفی مانند تفسیر زبان، تولید برنامه‌ی مفهومی و انتخاب مهارت مناسب را انجام می‌دهند، در حالی که مژوول‌های کنترلی کلاسیک مسئول اجرای دقیق این مهارت‌ها هستند. این تفکیک نه تنها اینمی سیستم را افزایش می‌دهد، بلکه امکان جایگزینی یا به روزرسانی هر مژوول را بدون بازطراحی کل سامانه فراهم می‌کند. [2]

یکی از چالش‌های کلیدی در این معماری‌ها، محدودیت دانش درونی مدل‌های زبانی بزرگ است. هرچند این مدل‌ها بر داده‌های عظیمی آموخته دیده‌اند، اما دانش آن‌ها ثابت و وابسته به زمان آموختش است. برای غلبه بر این مشکل، رویکرد تولید مبتنی بر بازیابی¹³ 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 اینمی یا تجربه‌های پیشین باشد.

پیاده‌سازی عملی این ایده را می‌توان در سامانه‌ی ARRC مشاهده کرد. یک معماری رباتیکی مژوولار است که در آن زبان طبیعی به عنوان رابط اصلی میان کاربر و ربات عمل می‌کند، اما تصمیم‌گیری نهایی از طریق چندین لایه‌ی بررسی و اعتبارسنجی انجام

¹³ Retrieval-Augmented Generation

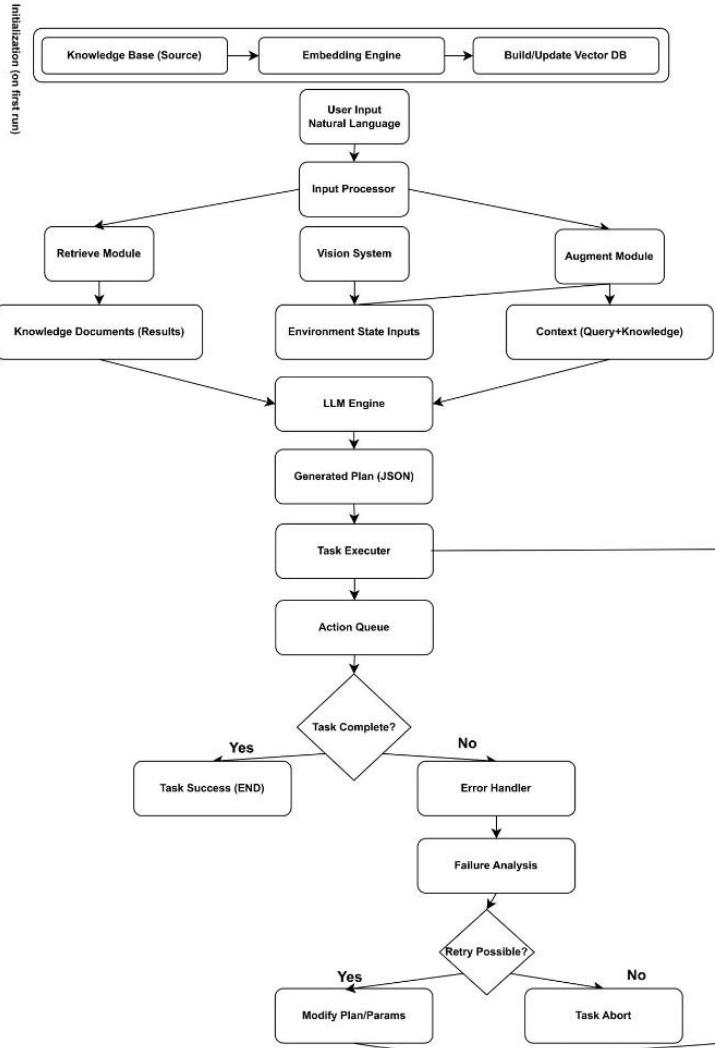
می‌شود. [8] در این سامانه، دستور زبانی کاربر ابتدا توسط مدل زبانی تحلیل می‌شود، سپس اطلاعات مرتبط از پایگاه دانش بازیابی می‌گردد و در نهایت یک خروجی ساخته‌یافته تولید می‌شود که برای مأذول‌های کنترلی قابل فهم است. (تصویر 2)

نکته‌ی مهم در ARRC آن است که خروجی LLM به صورت متن آزاد مستقیماً اجرا نمی‌شود. به جای 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 به طور مستقیم با مهار رفتارهای غیرقابل پیش‌بینی ها همسو است. [5]

ادغام با پشتنهای رباتیکی موجود، به ویژه ROS ، یکی دیگر از مزایای معماری‌های مأذول‌وار است. ROS به عنوان یک چارچوب استاندارد، ابزارها و پروتکل‌های ارتباطی گسترده‌ای را برای رباتیک فراهم می‌کند. سامانه‌هایی مانند ROS-LLM نشان داده‌اند که می‌توان مدل‌های زبانی بزرگ را به عنوان یک گرهی اضافی به این پشتنه 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 مأذول رفتاری مشخص و قابل اندازه‌گیری دارد. این شفافیت معماری برای پذیرش صنعتی رباتیک مبتنی بر LLM حیاتی است. [8]

با وجود این مزایا، معماری‌های مأذول‌وار بدون چالش نیستند. همانگی میان مأذول‌ها، تأخیر ناشی از فراخوانی مدل‌های بزرگ و مدیریت خطای مرزهای مأذول‌ها از جمله مسائلی هستند که در مثال‌های عملی نیز دیده می‌شوند. [8] این مسائل نشان می‌دهند که طراحی معماری به اندازه‌ی انتخاب مدل زبانی اهمیت دارد و موفقیت یک سامانه‌ی رباتیکی بیش از آنکه به قدرت خام LLM وابسته باشد، به نحوه استفاده‌ی مهندسی شده از آن بستگی دارد.



تصویر 2. معماری در سطح بالا: مژول ادراک، مشاهدات شیء محور تولید می‌کند. برنامه‌ریزی RAG دانش مربوط به کار را بازیابی کرده و یک طرح JSON می‌سازد. اجراکننده نیز با بررسی‌های اینمی، دستورها را از طریق XArm SDK اعتبارسنجی و اجرا می‌کند. [8]

2-ایمنی، قابلیت اطمینان و تعامل انسان در حلقه

با ورود مدل‌های زبانی بزرگ به قلب سامانه‌های ریاضیکی، مسئله‌ی اینمی 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، مدل‌های زبانی بزرگ 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 تعیین‌پذیری بالاست و از سوی دیگر، خطری بالقوه برای ایمنی ربات در محیط‌های واقعی.

یکی از چالش‌های اصلی، پدیده‌ی «توهم زبانی¹⁴» است؛ حالتی که در آن مدل زبانی پاسخی ظاهرآ منسجم اما نادرست یا غیرقابل اجرا تولید می‌کند. در زمینه‌ی رباتیک، چنین خطای 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 بر ROS-LLM، دستورات زبانی ابتدا به برنامه‌های ساخت‌یافته تبدیل می‌شوند و سپس توسط مأذول‌های 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 به عنوان 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 لبه میز بگذار» و سیستم این اصلاح را در برنامه‌ی کنترلی اعمال می‌کند. این تعامل بلاذرنگ، هم ایمنی را افزایش می‌دهد و هم حس کنترل و اعتماد کاربر را تقویت می‌کند.

بازخورد زبانی انسان همچنین به عنوان 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 نه تنها ابزار فرماندهی، بلکه رسانه‌ای برای انتقال تجربه و دانش ایمنی باشد. [1]

¹⁴ linguistic Hallucination

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 بدقت انجام شود تا توازن مناسبی میان استقلال و نظارت برقرار گردد.

2-8 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. مدل‌های زبانی بزرگ را باید به عنوان راه حلی نهایی، بلکه به عنوان 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.

یکی از بنیادی‌ترین محدودیت‌ها، عدم تضمین صحت و سازگاری خروجی‌های زبانی است. همان‌طور که در فصل‌های پیش اشاره شد، مدل‌های زبانی بزرگ ممکن است پاسخ‌هایی تولید کنند که از نظر زبانی قانع کننده اما از نظر فیزیکی یا منطقی نادرست باشند. این مسئله در سامانه‌های رباتیکی، که هر خطای 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 اجازه اجرا پیدا می‌کنند، و حتی در این شرایط نیز نظارت انسانی ضروری باقی می‌ماند. [6] این واقعیت نشان می‌دهد که اتکای کامل به مدل‌های زبانی بزرگ در محیط‌های ایمنی حساس هنوز عملی نیست.

چالش مهم دیگر، تعمیم‌پذیری در دنیای واقعی است. بسیاری از مثال‌های موفق، مانند SayCan یا RT-Grasp، در محیط‌های نسبتاً کنترل شده یا با مجموعه‌ای محدود از اشیا آزمایش شده‌اند. [7] [1] در مقابل، محیط‌های واقعی سرشار از تغییرات پیش‌بینی نشده، نویز حسی و تعاملات پیچیده‌اند. اگرچه مدل‌های زبانی بزرگ 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 تجسم یافته است.

مسئله‌ی هزینه‌های محاسباتی و زیرساختی نیز نباید نادیده گرفته شود. اجرای مدل‌های زبانی بزرگ، به ویژه در سناریوهای بلاذرنگ رباتیکی، نیازمند منابع پردازشی قابل توجه است. این موضوع در مقالات به ویژه در کاربردهای صنعتی بر جسته 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 مصالحه‌ایی 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 به صورت 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 در ک غنی‌تری از محیط و پیامدهای اعمال خود داشته باشند.

مسیر دیگر، حرکت به سوی معماری‌های عامل محور 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، هر چند در مقیاس محدود، نشان می‌دهند که این ایده از نظر عملی قابل پیاده‌سازی است.

فصل 3

نتیجه‌گیری

مطلوب ارائه شده در این گزارش نشان می‌دهد که مدل‌های زبانی بزرگ، اگرچه به تنها‌ی راه حل 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 در ک زبان طبیعی، استخراج نیت انسانی، برنامه‌ریزی سطح بالا 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 به روشنی 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 توهمندی زبانی و نبود تضمین رسمی در صحبت خروجی‌ها، ضرورت استفاده از معماری‌های ترکیبی و مازولار را برجسته می‌کند. در چنین معماری‌هایی، مدل زبانی نقش راهنمای و تصمیم‌ساز سطح بالا را بر عهده دارد، در حالی که کنترل دقیق و ایمن به مازولهای 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، نه تنها 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 مانع برای مقیاس‌پذیری و خودمختاری کامل ربات‌ها باشد؛ از این‌رو، 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 به درستی مدیریت شوند، می‌توان انتظار داشت که ربات‌ها در آینده‌ای نه چندان دور، تعامل طبیعی‌تر، در کم عمق‌تر و عملکرد قابل اعتمادتری در محیط‌های واقعی داشته باشند و به ابزارهایی مؤثر در زندگی روزمره و صنعت تبدیل شوند.

- [1] S. Peng, “Large Language Models for , W. Jiaqi, W. Zihao, L. Yiwei, J. Hanqi p. 18, 2021. Robotics: Opportunities, Challenges, and Perspectives,”
- [2] E. M. Christopher, W. Yuhui, Hongzhan, G. Antoine, J. Gonzalez-Billandon, M. J. Wang, “ROS-LLM: A ROS framework for embodied AI with task feedback , Zimmer p. 26, 2024. and structured reasoning,”
- [3] Z. Xiaoli, “A Review of Natural-Language-Instructed , L. Rui, G. Yibei, J. Runxiang p. 42, 2024. Robot,”
- [4] Francesco, “Parsing Natural Language O. , Danilo, M. Enrico B. Gianluca, . D. p. 4, 2019. Sentences into Robot,”
- [5] Z. Wenxiao, “Enhancing reliability in LLM- , D. Conan, . B. Thomas, B. H. Jin p. 14, 2025. integrated robotic systems: A unified approach,”
- [6] H. Andreas, “Domain-Specific Fine- , A. Benjamin, K. Urs, T. Aleksandar, K. Darko p. 5, 2021. Tuning of Large Language,”
- [7] Z. Liangjun, “RT-Grasp: Reasoning Tuning , X. Jinxuan, J. Shiyu, L. Yutian, Z. Yuqian p. 8, 2024. Robotic Grasping via,”
- [8] C. Robin, “ARRC: Advanced Reasoning , V. Eugene, M. Ammar Jaleel, . R. Salim Robot Control—Knowledge-Driven Autonomous Manipulation Using Retrieval- p. 8, 2025. Augmented Generation,”
- [9] D. z. Michał, “InCoRo: In-Context Learning for , Y. Z. Jiaqiang , G. C. Carla, V. David p. 20, 2024. Robotics Control with Feedback Loops,”
- [10] P. S. Yu, “Large Language Models for Robotics: , F. Zeng, W. Gan, Y. Wang, N. Liu p. 19, 2023. A Survey,”
- [11] Z. Yilun, “A , Zhebo, L. Minghao, L. Yufeng K. Dezhong, L. Shi , . X. Zhenhua, W. p. 35, 2025. Survey of LLM-Driven AI Agent Communication: Protocols,,”
- [12] T. Jie, “Large , L. Yiwei, Z. Huaqin, J. Hanqi, P. Yi, L. Zhengliang , W. Zihao, S. Peng . p. 52, 2024. Language Models for Manufacturing,”

- [13] p. 10, 2023. T. Dzmitry, “LLM-BRAIn: AI-driven Fast Generation of Robot,” , L. Artem
- [14] A. Rumaisa, “LLM-Driven Robots Risk Enacting Discrimination, , H. Andrew
p. 59, 2025. Violence, and,”
- [15] M. George, “Multi-Agent Systems ,Ziqi, . G. X. Haoyuan, Z. Dandan C. Junhong, Y.
p. 11, 2024. for Robotic Autonomy with LLMs,”
- [16] Z. Shuai, “An LLM- ,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 ,Y. Ziyi, S. R. Shreyas, S. Ankit
p. 15, 2023. Constraints for LLM-driven Robot,”
- [18] B. Rogerio, “Reshaping Robot ,B. Arthur, F. Luis, . H. Sami, K. Ashish, . M. Shuang
p. 7, 2022. Trajectories Using Natural Language Commands:,”
- [19] R. Nicholas, .,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 ,D. Wenhao, Z. Xingting, . W. Pan, Z. Jingwei, S. Yutong .
p. 22, 2025. LLM-Driven Robotic Control System,”