ROS-LLM: A ROS framework for embodied AI with task feedback and structured reasoning

مقاله سوم یک چارچوب جامع ارائه می‌کند که هدف آن استفاده از مدل‌های زبانی بزرگ در کنار سیستم ROS است تا ربات‌ها بتوانند با کمک زبان طبیعی، مهارت‌های پیچیده و رفتارهای چندمرحله‌ای را اجرا کنند .ROS یکی از اصلی‌ترین ابزارهای نرم‌افزاری در رباتیک است که برای ارتباط، کنترل و مدیریت بخش‌های مختلف ربات استفاده می‌شود. پیش از ظهور مدل‌های زبانی بزرگ، کاربر باید برای هر وظیفه، ساختارهای پیچیده‌ای در ROS تعریف می‌کرد، مانندAction Server، Node، Topicو .Service این مقاله یک سیستم ارائه می‌کند که این دشواری را کاهش می‌دهد و اجازه می‌دهد کاربر تنها با زبان انسانی، رفتارهای سطح بالا را برای ربات توصیف کند و مدل زبانی آن‌ها را به ساختارهای قابل اجرا در ROS تبدیل کند.

در ابتدای مقاله بیان می‌شود که چرا استفاده مستقیم از مدل‌های زبانی در رباتیک دشوار است. مدل زبانی فقط می‌تواند متن تولید کند و هیچ درکی از محدودیت‌های حرکتی ربات، قوانین فیزیکی یا ساختار ارتباطی ROS ندارد. همچنین، فرمان‌های زبانی معمولاً مبهم هستند و مدل باید بتواند آن‌ها را به بخش‌های کوچک‌تر و قابل کنترل تبدیل کند. به عنوان مثال اگر کاربر بگوید «چراغ میز را روشن کن»، مدل باید تشخیص دهد که ربات چه مهارتی دارد، کدام عملگرها باید فعال شوند، چهNode هایی باید فراخوانی شوند و در نهایت چه مسیر حرکتی تولید شود. این مقاله سیستمی طراحی کرده است که مدل زبانی را در چارچوبی محدود، ساختاریافته و مدیریت‌شده قرار می‌دهد تا رفتارهای رباتی قابل اطمینان تولید کند.

یکی از بخش‌های اصلی مقاله معرفی مفهوم Skill یا مهارت است. در ROS-LLM هر رفتار ربات به صورت یک مهارت تعریف می‌شود. این مهارت ممکن است ساده باشد، مانند بازکردن گیره بازوی ربات، یا ممکن است پیچیده باشد، مانند تنظیم موقعیت ربات نسبت به یک جسم مشخص. هر مهارت در قالب یک Action Server پیاده‌سازی می‌شود و ورودی، خروجی و مواردی که ممکن است در زمان اجرا رخ دهد به‌طور دقیق تعریف شده است. مدل زبانی به جای دستکاری مستقیم سخت‌افزار یا تولید کد کنترل خام، فقط مجاز است با این مهارت‌ها تعامل کند. این ایده باعث می‌شود که هر خروجی مدل زبانی در محدوده‌ای امن و قابل پیش‌بینی قرار بگیرد و احتمال رفتار خطرناک کاهش پیدا کند.

در ادامه مقاله معماری کامل ROS-LLM را توضیح می‌دهد. این معماری شامل چهار بخش اصلی است. بخش اول ماژول برنامه‌ریزی زبانی است که مدل زبانی در آن قرار دارد. این مدل فرمان کاربر را می‌گیرد، هدف را تشخیص می‌دهد، وظیفه را به گام‌های کوچک تقسیم می‌کند و فهرستی از مهارت‌هایی که ربات باید استفاده کند ارائه می‌دهد. بخش دوم ماژول اجرا است که ترتیب مهارت‌ها را مدیریت می‌کند و هر مهارت را به صورت یک وظیفه ROS فراخوانی می‌نماید. بخش سوم مهارت‌های فنی هستند که شامل action serverهای واقعی ربات می‌شوند و با سخت‌افزار، حسگر و الگوریتم‌های حرکتی ارتباط دارند. بخش چهارم سیستم بازخورد است که نتیجه اجرا را به مدل زبانی گزارش می‌دهد تا اگر خطایی رخ داد، مدل بتواند طرح جدیدی پیشنهاد دهد یا گام‌های قبلی را اصلاح کند.

یکی از مهم‌ترین قسمت‌های مقاله توضیح این است که چگونه LLM می‌تواند به شکل ایمن با ربات تعامل کند. نویسندگان تأکید می‌کنند که مدل زبانی نباید به صورت آزاد و بدون محدودیت به سخت‌افزار ربات دسترسی داشته باشد. برای این کار مجموعه‌ای از قواعد فیلترکننده طراحی شده است که بررسی می‌کند آیا فرمان تولیدشده توسط مدل زبانی با مهارت‌های موجود سازگار است یا خیر. اگر مدل فرمانی بدهد که خارج از محدوده امن باشد، سیستم اجرا آن را رد می‌کند و از مدل درخواست می‌کند که پاسخ جدیدی تولید کند. همچنین سیستم می‌تواند در صورت لزوم پرسش‌های شفاف‌سازی از مدل بخواهد، مثل اینکه «کدام جسم را باید بردارم؟» یا «آیا منظورت این است که ربات به نقطه A برود؟». این تعامل دوطرفه باعث می‌شود احتمال اشتباهات منطقی مدل زبانی کم شود.

در بخشی دیگر مقاله درباره نحوه برنامه‌ریزی وظایف چندمرحله‌ای صحبت می‌شود. مدل زبانی می‌تواند وظیفه را به مجموعه‌ای از گام‌ها تقسیم کند، اما این گام‌ها باید قابل اجرا باشند. برای مثال اگر کاربر بگوید «این میز را کاملاً مرتب کن»، مدل زبانی باید بفهمد منظور کاربر جمع‌کردن وسایل، جابه‌جایی اجسام غیرمرتبط، پاک‌کردن سطح و در نهایت تنظیم نهایی وسایل است. سپس باید برای هر یک از این بخش‌ها مهارت مناسب انتخاب کند. این فرآیند از آنجا اهمیت دارد که مدل زبانی بدون اطلاعات مناسب ممکن است گام‌هایی غیرواقع‌بینانه تولید کند. برای جلوگیری از این مشکل، سیستم از ربات بازخورد دریافت می‌کند. اگر در یک مرحله مهارت اجرا نشود یا جسم موردنظر شناسایی نگردد، مدل زبانی مجبور است طرح جدیدی ارائه دهد.

نویسندگان مقاله همچنین فرآیند یادگیری مهارت‌ها را توضیح می‌دهند. مهارت‌ها در ROS-LLM به صورت ماژولار طراحی شده‌اند. هر مهارت مانند یک واحد مستقل است که می‌تواند بارها توسط مدل زبانی در وظایف مختلف فراخوانی شود. این طراحی ماژولار باعث می‌شود توسعه‌دهندگان بتوانند مهارت جدید اضافه کنند بدون آنکه نیاز باشد کل سیستم را تغییر دهند. مدل زبانی نیز هنگام برنامه‌ریزی، فقط باید نام مهارت و پارامترهای آن را بنویسد، نه اینکه منطق داخلی مهارت را بفهمد. این موضوع باعث می‌شود سیستم هم انعطاف‌پذیر باشد و هم ایمن.

یکی از بخش‌های مهم مقاله توضیح آزمایش‌های عملی است. نویسندگان سامانه را روی ربات متحرک و بازوی رباتی آزمایش کرده‌اند و نشان داده‌اند که ربات می‌تواند وظایفی مانند پیدا کردن یک جسم، برداشتن آن، جابه‌جا کردن آن، تعامل با وسایل خانه و اجرای وظایف چندمرحله‌ای را تنها با یک فرمان زبانی انجام دهد. برای مثال ربات می‌تواند دستور «این جعبه را بردار و به اتاق کناری ببر» را با تقسیم وظیفه به مراحل کوچک و فراخوانی مهارت‌های مناسب اجرا کند. یا می‌تواند وظیفه‌ای مانند «محیط میز را مرتب کن» را با چندین مهارت پی‌درپی انجام دهد. این نتایج نشان می‌دهند که سیستم توانسته است تعامل زبان–ربات را به شکل قابل اتکا و واقعی اجرا کند.

مقاله سپس به تحلیل محدودیت‌ها می‌پردازد. یکی از نکات حساس این است که مدل‌های زبانی گاهی پاسخ‌های غلط تولید می‌کنند. اگر این پاسخ‌ها بدون فیلتر وارد سیستم ربات شوند ممکن است رفتار خطرناک اتفاق بیفتد. محدودیت دیگر مربوط به اطلاعات ناقص محیطی است. مدل زبانی نمی‌تواند بدون داشتن تصویر یا داده دقیق حسگر، وضعیت واقعی را بفهمد و تصمیم درست بگیرد. همچنین مدل‌های بزرگ نیاز به توان پردازشی زیاد دارند و ممکن است ربات‌های کوچک نتوانند آن‌ها را اجرا کنند. نویسندگان پیشنهاد می‌کنند که در آینده مدل‌های کوچک‌تر و تخصصی برای ربات‌ها طراحی شود که سرعت بیشتری داشته باشند و نیاز کمتری به پردازنده‌های قوی داشته باشند.

در پایان مقاله نتیجه‌گیری می‌کند که ROS-LLM یک قدم مهم در تلفیق هوش مصنوعی زبانی و رباتیک است. این سیستم نشان می‌دهد که ربات‌ها می‌توانند از طریق زبان طبیعی هدایت شوند، رفتارهای سطح بالا را با استدلال زبانی تولید کنند و در عین حال در چارچوب امن و قابل اعتماد ROS عمل کنند. نویسندگان باور دارند که این روش می‌تواند پایه‌ای برای نسل جدید ربات‌های هوشمند باشد؛ ربات‌هایی که می‌توانند نه تنها فرمان‌های ساده، بلکه وظایف پیچیده و چندمرحله‌ای را با کمک زبان انسان انجام دهند و این قابلیت می‌تواند رباتیک را وارد مرحله‌ای کاملاً جدید کند.