دانشگاه صنعتی شریف

دانشکدهٔ مهندسی کامپیوتر

گزارش پروژه

سامانه‌های نهفته

دستیار صوتی هوشمند مبتنی بر مدل **Deepseek** با قابلیت آنلاین و آفلاین

گروه: ارشیا یوسف‌نیا

علی مجیدی

صادق سرگران

استاد درس: دکتر انصاری

تاریخ: تابستان ۱۴۰۴

فهرست مطالب

لیست تصاویر

لیست جداول

۱ چکیده

پیشرفت‌های چشمگیر اخیر در مدل‌های زبانی بزرگ دستیارهای هوشمند را متحول کرده است. در این پروژه یک دستیار صوتی هوشمند مبتنی بر یک مدل زبانی بزرگ طراحی و اجرا می‌شود. دستگاه لبه در یک معماری ابر و مه و لبه با کاربر تعامل صوتی دارد و بر حسب نیاز از ظرفیت‌های مه و ابر استفاده می‌کند تا پاسخ پرسش‌های کاربر را بدهد. مدل‌های زبانی بزرگ در مه و ابر با اندازه‌های مختلف مستقر هستند تا یک معماری منعطف و بهینه درست کنند. وجود بردهای رزبری‌پای و مدل‌های زبانی بزرگ متن‌باز این انعطاف را دوچندان می‌کند.

۲ مقدمه

مدل‌های زبانی بزرگ یک پیشرفت چشمگیر در پردازش زبان‌های طبیعی هستند. مدل‌های تجاری قوی در کنار مدل‌های متن‌باز گوناگون یک فضای بسیار مناسب و روبه‌رشد در این حوزه درست کرده است. وجود سرویس‌های API متنوع در کنار ابزارهای اجرای محلی مدل‌ها در اندازه‌های مختلف استفاده عملی از این مدل‌ها در کاربردهای مختلف را بسیار آسان کرده است. یک حوزه که ارتباط تنگاتنگی با مدل‌های زبانی بزرگ دارد دستیارهای صوتی هوشمند است. این حوزه با تعامل صوتی با کاربر سر و کار دارد و در قلب خود باید زبان طبیعی را درک کند. چالش‌هایی که پیش‌روی این استفاده وجود دارد منابع بسیار محدود در دستگاه لبه و در عین حال نیاز به سرعت پاسخ بالا است. مدل‌های زبانی بزرگ به منبع‌های پردازشی زیاد نیاز دارند و همین موضوع استفاده مستقیم از آنها در لبه را با چالش مواجه می‌کند. همچنین استفاده کامل از سرویس‌های ابری مشکلات هزینه، حریم خصوصی، و محرمانگی داده دارد. در ادامه معرفی‌ای از المان‌های اساسی این پروژه می‌آید:

۱.۲ رایانش ابر و مه و لبه

این معماری سلسله‌مراتبی، پردازش را در سه لایه لبه (دستگاه‌های نهایی)، مه (گره‌های میانی نزدیک به کاربر) و ابر (مراکز داده متمرکز) توزیع می‌کند. هر لایه نقش مکمل دیگری را ایفا می‌نماید: لبه برای پاسخ‌دهی فوری، مه برای پردازش نیمه‌متمرکز و ابر برای تحلیل‌های عمیق و ذخیره‌سازی کلان. این پارادایم یکپارچه، مزایای عملیاتی برجسته‌ای را فراهم می‌سازد که مهم‌ترین آن‌ها کاهش تأخیر، بهینه‌سازی پهنای باند و حفظ حریم خصوصی داده‌ها از طریق پردازش محلی است. در نهایت، این همزیستی هوشمندانه، راه‌حلی بهینه برای چالش‌های رایج در سیستم‌های اینترنت اشیاء و هوش مصنوعی ارائه می‌دهد [۱].

DeepSeek ۲.۲

در میان مدل‌های زبانی بزرگ، دیپ‌سیک (DeepSeek) در نقش یک مدل پیشرفته و مقرون‌به‌صرفه، خود را به‌عنوان گزینه‌ای جدی در عرصه هوش مصنوعی مطرح کرده است. این مدل با معماری نوآورانه و کارایی چشمگیر، نه‌تنها از لحاظ فنی با برترین مدل‌های تجاری رقابت می‌کند، بلکه با سیاست متن‌باز (open-source) خود، امکان دسترسی آزاد و شفاف را برای جامعه پژوهشی و توسعه‌دهندگان فراهم می‌سازد. علاوه بر این، بهینه‌سازی‌های انجام‌شده در دیپ‌سیک، استفاده از آن را حتی در محیط‌های با منابع محدود ممکن کرده که این امر هزینه‌های پیاده‌سازی و مقیاس‌پذیری را به‌طور قابل‌توجهی کاهش می‌دهد [۲].

OpenRouter ۳.۲

در میان سرویس‌های میانی دسترسی به مدل‌های زبانی بزرگ، اوپن‌روتر به عنوان پلتفرمی پیشرو، دسترسی یکپارچه و مقرون‌به‌صرفه به گسترده‌ترین مجموعه از مدل‌های هوش مصنوعی (اعم از تجاری و متن‌باز) را فراهم می‌سازد. این سرویس با ارائه یک API واحد و استاندارد، به توسعه‌دهندگان این امکان را می‌دهد تا بدون وابستگی به یک ارائه‌دهنده خاص، به راحتی بهترین مدل را برای نیاز خود انتخاب، آزمایش و به کار گیرند. این رویکرد نه‌تنها انعطاف‌پذیری و آزادی عمل را به حداکثر می‌رساند، بلکه با ایجاد رقابت در قیمت‌گذاری، هزینه استفاده از قدرتمندترین مدل‌های زبانی را به شکل چشمگیری کاهش می‌دهد [۳].

Ollama ۴.۲

در میان ابزارهای استقرار مدل‌های هوش مصنوعی، اولاما به عنوان یک پلتفرم کارآمد و کاربرپسند، اجرای روان مدل‌های زبانی بزرگ را روی سخت‌افزارهای محلی ممکن ساخته است. این ابزار با فراهم آوردن محیطی یکپارچه برای دانلود، مدیریت و اجرای مدل‌های بهینه‌شده، نیاز به وابستگی به سرویس‌های ابری پرهزینه را از بین می‌برد. اولاما با پشتیبانی از طیف وسیعی از مدل‌های متن‌باز و معماری‌های مختلف، کنترل کامل بر داده‌ها و حریم خصوصی را به کاربران بازمی‌گرداند و گزینه‌ای ایده‌آل برای توسعه‌دهندگان و سازمان‌هایی است که به دنبال حفظ امنیت و کاهش هزینه‌های عملیاتی هستند [۴].

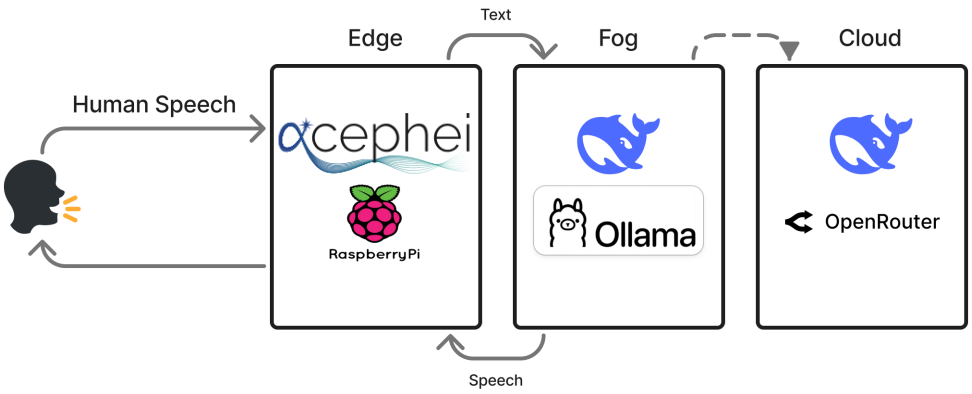
RaspberryPi ۵.۲

در میان پلتفرم‌های سخت‌افزاری، رزبری‌پای به عنوان یک رایانه تک‌برد مقرون‌به‌صرفه و بسیارانعطاف‌پذیر، نقش کلیدی در توسعه و استقرار سیستم‌های لبه ایفا می‌کند. این دستگاه با معماری کم‌مصرف و اندازه فشرده خود، نه‌تنها امکان پردازش محلی داده‌ها را در محیط‌های با منابع محدود فراهم می‌سازد، بلکه به لطف جامعه متن‌باز و اکوسیستم گسترده ماژول‌های افزودنی، امکان توسعه سریع و مقیاس‌پذیر راه‌حل‌های اینترنت اشیاء و هوش مصنوعی را فراهم می‌کند [۵].

۳ پیاده‌سازی

۱.۳ طراحی سیستم و چینش المان‌ها

بنا به مقدمه، یک معماری با ۳ المان اصلی درست می‌کنیم. دستگاه‌های لبه، مه، و ابر. دستگاه لبه باید با کاربر تعامل صوتی کند، صدا را به متن زبان طبیعی تبدیل کند و به دستگاه مه برای پاسخ بفرستد. دستگاه مه یک مدل‌زبانی بزرگ محلی را اجرا می‌کند و بسته به تنظیمات یا با آن مدل پاسخ می‌دهد یا با لایه ابر که یک سرویس خارجی است ارتباط می‌گیرد. صداهای گوناگون هم در همین دستگاه مه ساخته می‌شوند تا مشکل محدودیت منابع لبه را حل کند. در شکل ۱ این روند آمده است.



شکل ۱ : معماری سیستم دستیار صوتی و تقسیم نقش‌ها

۲.۳ جزییات ابزارها

جدول ۲.۳ مشخصه‌های فناوری و ابزارهای استفاده شده را نشان می‌دهد.

المان	مشخصات
سخت‌افزار لبه	برد رزبری‌پای ۳ با رم ۱ گیگابایت
گفتار به متن	مدل alphacephei [۶]
ورودی/خروجی صدا	میکروفون و بلندگو متصل به لبه یا گوشی موبایل به‌عنوان جایگزین
مدل‌های مه	دیپ‌سیک در ollama برای پردازش + edge-tts برای متن به گفتار
لایه ابری	اندپوینت openrouter برای ارتباط ابری

جدول ۱ : مشخصات فنی ابزارهای نرم‌افزاری و سخت‌افزاری استفاده شده

تنظیم استفاده از مدل مه یا ابری با فرمان صوتی تغییر می‌کند. مستندات فنی جزیی از اجزا و کارکردها به همراه

کد پیاده‌سازی جداگانه داده شده است. برای جزییات به آنجا رجوع کنید.

۴ نتیجه‌گیری و کارهای بیشتر

این پروژه را می‌توان از چند جهت بهبود داد:

- بهبود سخت‌افزاری دستگاه لبه
- ایجاد یک چارچوب برای اثرگذاری دستیار بر محیط و عاملیت
- بهبود رابط کاربری سخت‌افزاری با یک واسط گرافیکی

مراجع

[۱] URL: https://blog.isa.org/edge-fog-and-cloud-computing-whats-the-difference.

[۲] URL: https://www.deepseek.com/.

[۳] URL: https://openrouter.ai/.

[۴] URL: https://ollama.com/.

[۵] URL: https://www.raspberrypi.com/.

[۶] URL: https://alphacephei.com/vosk/models.