به نام خدا

آرین عبداللهی ثابت نژاد ۹۹۵۲۱۴۴۲

امیرمحمد کمیجانی ۹۹۵۲۲۰۳۲

محمدحسین میرزایی ۹۹۵۲۲۱۵۸

۱)

توضیح کلی پروژه : Retrieval-Augmented Generation (RAG)یک رویکرد پیشرفته در زمینه هوش مصنوعی است که با هدف افزایش کیفیت تولید متن از طریق دسترسی به منابع اطلاعاتی خارجی توسعه یافته است. این فناوری، مدل‌های پیشرفته تولید متن مانند مدل‌های زبانی بزرگ را با سیستم‌های جستجوی پیشرفته اطلاعات تلفیق می‌کند تا امکان دسترسی به اطلاعاتی دقیق و به‌روز فراهم آورد، و در نتیجه، به ارائه پاسخ‌هایی با کیفیت تر کمک کند. در این فرآیند، ابتدا سیستم بازیابی اطلاعات (IR)مجموعه‌ای از مستندات یا اطلاعات مرتبط با پرسش یا درخواست ورودی را شناسایی می‌کند. سپس، با استناد به این اطلاعات بازیابی شده، مدل تولید متن پاسخ یا متن مورد نظر را ایجاد می‌کند.این پروژه با هدف ارتقاء پاسخ‌دهی به سوالات مرتبط با گزارش‌های سازمانی توسط RAG طراحی شده است. استفاده از این تکنولوژی این امکان را فراهم می‌آورد که به سوالات کاربران بر اساس محتوای موجود در گزارش‌های سازمانی پاسخ‌هایی دقیق و کنونی ارائه دهیم. به صورت کلی دو فایل word داریم که همان کورپس یا منبع میباشد و باید از آن اطلاعات را بازیابی کنیم و سوالاتی در این فایل ها هست که دو فایل questions نیز داریم که سوالاتی را دارند و باید از آنها این سوالات را برداریم و مطالب مربوط به آنرا در کورپس را جست و جو کنیم و مطالب مربوط به آنرا پیدا کنیم و به مدل خود مثلا llama3 بدهیم و بگوییم این کورپس و این سوال را داریم و از مدل بخواهیم که پاسخ بدهد و سرچ کردن و ایندکس کردن را با استفاده از کتابخانه های مختلف انجام دادیم.

استفاده از تکنولوژی بازیابی-افزایش تولید (RAG) در پاسخ به سوالات متداول (FAQ) مزایای متعددی را به همراه دارد، به خصوص برای پروژه‌هایی که با حجم زیادی از داده‌ها و سوالات

پیچیده سروکار دارند.
با دسترسی به اطلاعات دنیای واقعی از طریق پایگاه‌های دانش خارجی، دقت و اعتبار پاسخ‌ها را به سوالات متداول افزایش می‌دهد .این امر به ویژه برای موضوعات تخصصی یا فنی که ممکن است در مجموعه داده‌های آموزش مدل‌های زبانی بزرگ (LLMs) به طور کامل پوشش داده نشوند، مفید است.

RAGبه LLM ها اجازه می دهد تا به اطلاعات به روز شده از منابع خارجی دسترسی داشته باشند و آنها را در پاسخ های خود ادغام کنند .این امر تضمین می کند که پاسخ ها همیشه دقیق و مرتبط با آخرین اطلاعات باشند.

RAGبه LLM ها کمک می کند تا با دامنه وسیع تری از سوالات، از جمله سوالات باز، چالش برانگیز یا عجیب، سازگار شوند .با دسترسی به اطلاعات خارجی، LLMها می توانند زمینه بیشتری را برای درک سوال و تولید پاسخ های جامع تر پیدا کنند.

RAGمی تواند به طور بالقوه بار محاسباتی پاسخگویی به سوالات متداول را کاهش دهد، به خصوص برای مجموعه داده های بزرگ سوالات .با بازیابی اطلاعات مرتبط از منابع خارجی، LLM ها می توانند کار کمتری را برای تولید پاسخ از ابتدا انجام دهند.

(۲

LLAMA3 مخفف **(Large Language Model Meta AI)**یک خانواده از مدل‌های زبانی پایه پیشرفته است که توسط **Meta AI** توسعه یافته‌اند .این مدل‌ها با استفاده از معماری Transformerو با حجم عظیمی از داده‌ها (بیش از ۱۵ تریلیون توکن) آموزش دیده‌اند و در دو اندازه مختلف ۸ میلیارد و ۷۰ میلیارد پارامتر ارائه می‌شوند.

LLAMA3 می‌تواند برای ایجاد سیستم‌های پاسخ به سوال قدرتمند استفاده شود که می‌توانند به سوالات مربوط به موضوعات مختلف پاسخ دهند.

LLAMA3 در مقایسه با مدل‌های زبانی قبلی، عملکرد قابل توجهی در variousوظایف

پردازش زبان طبیعی (NLP) مانند پاسخ به سوال، خلاصه‌سازی متن، ترجمه و تولید متن نشان می‌دهد.

LaBSE . (مخفف **Language Model for Benchmarks, Sentences, and English**) یک مدل زبانی بزرگ (LLM) است که توسط **Google AI** توسعه یافته است. این مدل با استفاده از معماری Transformer و با حجم عظیمی از داده‌ها (۶۰۰ میلیارد توکن) آموزش دیده است و برای انجام 各种 وظایف پردازش زبان طبیعی (NLP) تنظیم شده است.

LaBSEدر مقایسه با مدل‌های زبانی قبلی، عملکرد قابل توجهی در variousوظایف NLP مانند پاسخ به سوال، خلاصه‌سازی متن، استنتاج و طبقه‌بندی متنی نشان می‌دهد.

LaBSEبرای انجام 各种 وظایف NLP تنظیم شده است، به این معنی که می‌توان از آن برای انجام کارهای مختلف بدون نیاز به آموزش جدید استفاده کرد.

مدل‌های زبانی بزرگ (LLMs) مانند LLAMA3 و LaBSE پیشرفت‌های قابل توجهی در زمینه پردازش زبان طبیعی (NLP) ایجاد کرده‌اند .این مدل‌ها می‌توانند برای انجام various وظایف NLP مانند پاسخ به سوال، خلاصه‌سازی متن، ترجمه و تولید متن با دقت بالا استفاده شوند.
از سوی دیگر، سیستم‌های بازیابی اطلاعات (IR) برای جستجو و بازیابی اطلاعات مرتبط از مجموعه داده‌های بزرگ متن طراحی شده‌اند .این سیستم‌ها از تکنیک‌های مختلفی مانند مطابقت کلمات کلیدی، مدل‌سازی آماری و یادگیری ماشین برای رتبه‌بندی اسناد بر اساس مرتبط بودن آنها با یک پرس و جو استفاده می‌کنند.
ترکیب LLMs و سیستم‌های IR می‌تواند مزایای هر دو روش را به همراه داشته باشد و منجر به سیستم‌های بازیابی اطلاعات قدرتمندتر و کارآمدتر شود .در زیر برخی از راه‌های اصلی ترکیب LLMsو سیستم‌های IR آورده شده است:

- **بازیابی افزایشی :**در این رویکرد، از یک مدل زبانی برای تولید یک نمایش زبانی از یک پرس و جو کاربر استفاده می‌شود .سپس این نمایش زبانی برای بازیابی اسناد مرتبط از

یک سیستم IR استفاده می‌شود .این رویکرد می‌تواند به سیستم‌های IR کمک کند تا پرس و جوهای پیچیده و غیرقابل بیان را بهتر درک کنند

- **رتبه‌بندی مرتبط :**در این رویکرد، از یک مدل زبانی برای محاسبه امتیاز مرتبط بودن برای هر سند در مجموعه داده‌های IR استفاده می‌شود .سپس از این امتیازات برای رتبه‌بندی اسناد بر اساس مرتبط بودن آنها با پرس و جو کاربر استفاده می‌شود .این رویکرد می‌تواند به سیستم‌های IR کمک کند تا اسناد مرتبط‌تر را در بالاترین رتبه‌ها قرار دهند.
- **رتبه‌بندی مرتبط :**در این رویکرد، از یک مدل زبانی برای محاسبه امتیاز مرتبط بودن برای هر سند در مجموعه داده‌های IR استفاده می‌شود .سپس از این امتیازات برای رتبه‌بندی اسناد بر اساس مرتبط بودن آنها با پرس و جو کاربر استفاده می‌شود .این رویکرد می‌تواند به سیستم‌های IR کمک کند تا اسناد مرتبط‌تر را در بالاترین رتبه‌ها قرار دهند.

۳) از جمله چالش ها کرش کردن کولب بود به دلیل محدودیت رَم که از دستور os.kill استفاده کردیم برای ریست کردن رم و سِشِن میباشد.همچنین با استفاده از این دستور : !ls "/root/.cache/huggingface/hub/"
با استفاده از این دستور فایل های اضافی که در کَش میباشند را شناسایی و حذف میکنیم تا محدودیت رَم تا حدودی جبران شود.
سپس توکنایزر و مدل را لود کرده و مدل را quantize میکنیم تا رَم بسیار کمتری مصرف کند.
سپس داکیومنت های سوال ها و ورودی ها را preprocess میکنیم در نهایت آنرا به vectore store index میدهیم تا به عنوان یک کوئری انجین ۳ تا از نزدیک ترین ورودی ها را به سوال ما پیدا میکند و در هم آمیختن آنها با توجه به سوال فرد پاسخ مناسب میدهیم.

۴) کیفیت پاسخ نسبتا مناسب است البته با محدودیت هایی از جمله gpu,ram مواجه بودیم.
باتوجه به داشتن کانتکست پاسخ بر اساس وقایع و تصمیمات گذشته موجود در اسناد داده شده است نه اطلاعات عمومی.

۵) محدودیت gpu و محدودیت ram که باعث میشد در دانلود مدل بارها با مشکل اتمام ram مواجه شویم که به سراغ یافتن مدل های دیگر برویم که بتواند دقت و عملکرد خوبی برای

inference از context قبلی که بازیابی شده است؛که این پیدا کردن همچین مدلی از چالش ها بوده است.

۶) یادگیری روش هایی که به صورت بهینه بتوان از منابع استفاده کرد به خصوص در فرآیند کش کردن و در کل یادگیری فایل ها و دایرکتوری ها در محیط اجرای پروژه مثلا کولب
کسب تجربه در آموزش و کار کردن با مدل های بزرگ و تسک های سنگین
پیشنهاد : سرچ کردن برای مدل های سبک تر و البته مناسب تر برای تَسک که نیازمند زمان زیادی میباشد.

۷)

https://www.cohesity.com/glossary/retrieval-augmented-generation-rag/

https://zapier.com/blog/llama-meta/

https://towardsdatascience.com/labse-language-agnostic-bert-sentence-embedding-by-google-ai-531f677d775f

https://docs.llamaindex.ai/en/stable/api_reference/llms/huggingface/

کارگاه hazm

https://medium.com/@tejaswi_kashyap/rag-processing-using-llamaindex-43d9786f9d8e