بسم الله الرحمن الرحیم

شرکت مهندسی نرم افزاری هلو

گزارش تحقیق در خصوص RAG ۳

کاری از امیرعلی نسیمی

فهرست

مقدمه‌ای بر تولید بازیابی-تقویتی (RAG)

تولید بازیابی-تقویتی (RAG) ترکیبی از دو فناوری قدرتمند است: مدل‌های زبانی بزرگ (LLM) و منابع دانش خارجی مثل پایگاه‌های داده‌ی برداری. LLM پاسخ‌ها را تولید می‌کند، در حالی که پایگاه داده خارجی زمینه‌های اضافی را برای بهبود دقت و مرتبط بودن پاسخ‌ها فراهم می‌کند. این روش محدودیت‌های استفاده از LLM به‌تنهایی را حل می‌کند، مانند:

- محدودیت در پنجره‌ی متنی: LLMها تنها می‌توانند مقدار محدودی از اطلاعات را همزمان پردازش کنند، بنابراین داشتن یک پایگاه داده خارجی اجازه می‌دهد اطلاعات بیشتری به آن‌ها ارجاع داده شود.
- اطلاعات قدیمی: LLMها ممکن است به جدیدترین اطلاعات دسترسی نداشته باشند، در حالی که یک پایگاه داده خارجی می‌تواند به‌طور مداوم به‌روز شود.

در این گزارش، یک رویکرد RAG پیشرفته نشان داده شده است که یک سیستم ساده RAG را بهبود می‌دهد و تکنیک‌های پیشرفته‌ای را به آن اضافه می‌کند. این تکنیک‌ها به سه دسته اصلی تقسیم می‌شوند:

۱. تکنیک‌های پیش از بازیابی: بر بهینه‌سازی پردازش پرسش قبل از شروع بازیابی تمرکز دارند.
۲. تکنیک‌های بازیابی: مرتبط با چگونگی دریافت اطلاعات از منبع خارجی هستند.
۳. تکنیک‌های پس از بازیابی: اطلاعات بازیابی‌شده را برای تطبیق بهتر با پرسش کاربر بهبود می‌بخشند.

مروری بر فناوری‌های اصلی

مدل Gemma

مدل Gemma یک مدل زبانی بزرگ است که از سوی گوگل و از طریق مدل‌های Kaggle ارائه شده است. این مدل به‌عنوان LLM در این سیستم RAG استفاده می‌شود. از آنجایی که LlamaIndex به‌طور پیش‌فرض از

Gemma پشتیبانی نمی‌کند، نیاز است که یک کلاس LLM سفارشی ساخته شود. این کار انعطاف‌پذیری را برای ادغام مدل‌های مختلف LLM فراهم می‌کند و همچنین امکان تنظیم دقیق عملکرد مدل برای وظایف خاص را می‌دهد.

Weaviate

Weaviate یک پایگاه داده‌ی برداری متن‌باز است که دانش را به‌صورت بردارهای برداری ذخیره می‌کند. در این دفترچه، Weaviate برای ذخیره‌ی قطعات متنی (Nodes) و متاداده‌های آن‌ها به کار می‌رود که این امر باعث بازیابی سریع‌تر زمینه‌های مرتبط با پرسش کاربر می‌شود. Weaviate از روش‌های جستجوی ترکیبی و فیلتر کردن متاداده نیز پشتیبانی می‌کند که در تنظیمات پیشرفته RAG مورد استفاده قرار می‌گیرد.

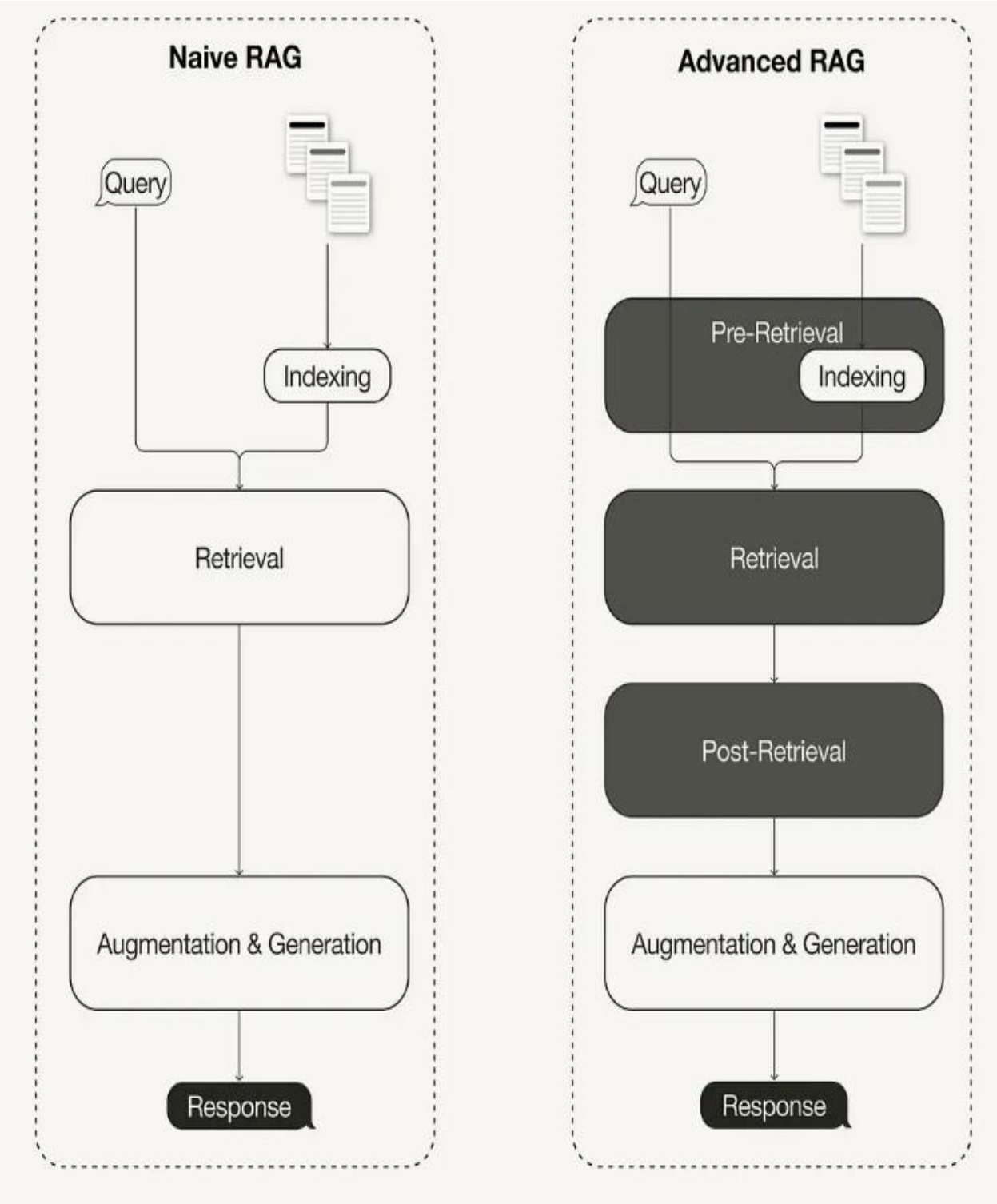
LlamaIndex

LlamaIndex (که قبلاً به عنوان GPT Index شناخته می‌شد) یک ابزار مبتنی بر پایتون است که برای مدیریت تعامل بین LLMها و پایگاه‌های داده خارجی طراحی شده است. این ابزار لایه‌ای برای ارکستراسیون فراهم می‌کند که نحوه‌ی ارسال پرسش به پایگاه داده‌ی برداری، دریافت نتایج و ادغام زمینه‌های بازیابی‌شده در فرآیند تولید پاسخ توسط LLM را مدیریت می‌کند. همچنین ویژگی‌های پیشرفته‌ای نظیر مرتب‌سازی مجدد و مهندسی درخواست few-shot را پشتیبانی می‌کند که به بهبود کیفیت پاسخ‌ها کمک می‌کند.

تجزیه و تحلیل کامل RAG پیشرفته

تکنیک‌های RAG پیشرفته با بهبود هر مرحله از پردازش پرسش تا تولید پاسخ نهایی، خروجی‌های دقیق‌تر و مرتبط‌تری را فراهم می‌کنند:

مراحل اجرای RAG پیشرفته



تعریف Gemma به‌عنوان LLM سفارشی

همانطور که پیش‌تر گفته شد، مدل Gemma به‌طور پیش‌فرض در LlamaIndex پشتیبانی نمی‌شود، بنابراین نیاز است که یک کلاس LLM سفارشی ساخته شود. این کار امکان یکپارچه‌سازی و تنظیم دقیق عملکرد مدل را فراهم می‌کند. اگر قصد دارید مدل Gemma را برای یک کار خاص تنظیم کنید، این روش اجازه‌ی کنترل بیشتری روی نحوه‌ی پردازش پرسش‌ها توسط مدل می‌دهد.

برای ادغام مدل

- مدل ایجاد بردار: این مدل بردارهای برداری از قطعات متن و پرسش کاربر ایجاد می‌کند. این بردارها به‌صورت ریاضی نمایش داده می‌شوند و می‌توان آن‌ها را در فضای برداری مقایسه کرد تا شباهت‌ها مشخص شود.
- LLM: مدل زبانی بزرگ (در این مورد Gemma) پاسخ‌ها را بر اساس پرسش کاربر و زمینه‌های بازیابی‌شده تولید می‌کند. LLM فقط به‌طور کورکورانه پاسخ نمی‌دهد، بلکه از دانش بازیابی‌شده برای تولید پاسخ‌های مرتبط استفاده می‌کند.

بارگذاری داده‌ها

در این مثال، داده‌ها از گزارش هوش مصنوعی Kaggle 2023 تهیه شده‌اند که شامل نوشته‌های مربوط به راه‌حل‌های ارائه‌شده در رقابت‌های مختلف Kaggle است. ساختارمند بودن این داده‌ها، آن را به گزینه‌ای مناسب برای RAG تبدیل می‌کند زیرا عناوین رقابت‌ها، روش‌ها و متاداده‌ها به‌صورت سازمان‌دهی‌شده موجود هستند.

انتخاب این مجموعه داده نشان می‌دهد که RAG می‌تواند داده‌های واقعی و تخصصی را پردازش کند و زمینه‌ی مناسبی را برای پاسخ‌گویی به پرسش‌های خاص فراهم کند (مانند "بهترین روش برای حل این رقابت چه بوده است؟").

ایجاد اسناد با متاداده

هر سند (مثل نوشته‌های راه‌حل‌های رقابت‌های Kaggle) با متاداده ذخیره می‌شود. متاداده در تکنیک‌های پیش از بازیابی مثل بازیابی خودکار و فیلتر کردن متاداده اهمیت زیادی دارد که به کاهش دامنه جستجو قبل از پردازش پرسش کمک می‌کند.

به‌عنوان مثال، اگر یک سند نشان‌دهنده‌ی گزارش یک رقابت باشد، متاداده‌هایی مانند عنوان رقابت، تاریخ ارسال و دسته‌بندی آن به همراه متن ذخیره می‌شود. این امر اجازه می‌دهد که سیستم RAG سریع‌تر اسناد مرتبط با رقابت‌های خاص را فیلتر کند.

تقسیم اسناد به قطعات (Nodes)

از آنجایی که LLMها تنها می‌توانند مقدار محدودی از متن را به‌طور همزمان پردازش کنند، اسناد بزرگ به قطعات کوچک‌تر به نام Nodes تقسیم می‌شوند. این فرآیند باعث می‌شود که همه‌ی بخش‌های مرتبط یک سند حتی اگر در یک پنجره متنی جا نشود، بتواند پردازش شود.

در این دفترچه، روش SentenceSplitter برای تقسیم متن به قطعات کوچک‌تر (ترجیحاً به‌صورت جملات) استفاده می‌شود. این روش ساده اما موثر است. با این حال، می‌توان از تکنیک‌های پیشرفته‌تری مانند HTMLNodeParser استفاده کرد که برای اسنادی که دارای ساختار سلسله‌مراتبی هستند (مثل صفحات وب یا گزارش‌های فنی) مناسب‌تر است.

ایجاد ایندکس

قطعات متنی (Nodes) به‌صورت برداری در پایگاه داده‌ی Weaviate ذخیره می‌شوند. این ذخیره‌سازی به بازیابی سریع‌تر و دقیق‌تر بر اساس بردارهای معنایی متن کمک می‌کند. استفاده از Weaviate باعث می‌شود که جستجو به‌صورت معنایی انجام شود، به این معنی که سیستم اسناد را بر اساس معنی آن‌ها جستجو می‌کند نه فقط کلمات کلیدی.

در این دفترچه، Weaviate به‌صورت Embedded mode استفاده می‌شود که نیاز به تنظیمات پیچیده یا استفاده از کلید API را از بین می‌برد. این حالت برای نمونه‌سازی و آزمایش محلی بسیار مفید است.

راه‌اندازی موتور پرسش RAG پیشرفته

در این مرحله چندین تکنیک پیشرفته به هم پیوند می‌خورند:

۱. بازیابی خودکار

بازیابی خودکار به‌طور خودکار فیلترهای متاداده را از پرسش کاربر استخراج می‌کند. به‌عنوان مثال، اگر کاربر بپرسد "بهترین روش برای حل رقابت Kaggle با عنوان 'Google - Isolated Sign Language Recognition' چه بوده است؟"، سیستم می‌تواند فیلتر متاداده‌ای مانند `competition_title == 'Google - Isolated Sign Language Recognition'` را بر اساس متن پرسش به‌طور خودکار ایجاد کند.

این تکنیک پیشرفته‌تر از سیستم‌های ساده RAG است که نیاز به تعیین دستی فیلترها توسط کاربر دارند.

۲. جستجوی ترکیبی

جستجوی ترکیبی روش‌های جستجوی معنایی (بر اساس شباهت برداری) و جستجوی کلمات کلیدی را ترکیب می‌کند تا نتایج دقیق‌تر و مرتبط‌تری فراهم شود.