به‌نام خدا

مقدمه

سلام، اول از همه بخوام بگم به‌طورکلی پروژه چی بود: **چت‌بات.** کاربران با ثبت‌نام می‌تونن وارد سیستم شوند و با چت‌بات‌های موجود چت کنند. چت‌بات‌ها توسط یک سری کاربر خاص ساخته می‌شن که بهشون کاربر چت‌بات‌ساز می‌گیم. کاربر چت‌بات‌ساز هم در همین سامانه و در قسمت مخصوص خودش می‌تونه چت‌بات خودش رو بسازه. کاربر چت‌بات‌ساز توسط ادمین مشخص می‌شه. مهم‌ترین قسمت پروژه چت‌بات‌ها هستند. چت‌بات‌ها knowledge base هستند، به این صورت که کاربر چت‌بات‌ساز، چت‌بات خودش رو بر اساس دیتاهای مدنظرش می‌سازه. حالا کاربران می‌تونن با اون چت‌بات چت کنند. چت‌بات هم با استفاده از اون دیتاهایی که توسط کاربر چت‌بات‌ساز در اختیارش قرار گرفته شده، به سوالات کاربر پاسخ می‌ده.

این سامانه رو با استفاده از جنگو، داکر، هم‌روش (با استفاده از پایپ‌لاین CI/CD ساخته‌شده) و پستگرس پیاده‌سازی کردم. چت‌بات‌ها هم با استفاده از API مربوط به openAI طراحی شده‌اند.

چت‌بات

همون‌طور که گفتم مهم‌ترین بخش کار، چت‌بات‌ها بودند. هر چت‌بات دیتاهای خاص خودش رو داره. که براساس اون دیتاها به سوالات کاربر پاسخ داده می‌شه. هر چت‌بات یک پرامپت سیستمی هم داره که توسط کاربر چت‌بات‌ساز مشخص می‌شه.

به عنوان نمونه:

```
system_prompt="Based on the specific data: {data}, answer to the following question."
```

از اون‌جایی که هر چت‌بات می‌تونه دیتاهای خیلی زیادی داشته‌باشه، یک از چالش‌های مطرح این بود که کدوم بخش از این دیتاها به پرامپت کاربر مربوط هست. توجه داشته‌باشید که با توجه به محدودیت توکن موجود در پرامپت‌های openAI نمی‌تونستم کل دیتا رو برای openAI بفرستم. حتی اگر محدودیت توکن وجود نداشت، فرستادن کل دیتا هزینه‌ی الکی‌ای هست. حتی شاید فرستادن کل دیتا روی دقت پاسخ‌ها هم تأثیر بذاره.

پیدا کردن دیتای مربوطه!

برای پیدا کردن دیتای مربوط به پرامپت کاربر از embedding خود openAI استفاده کردم. اول از همه باید دیتای چت‌بات رو امبد می‌کردم. پرامپت کاربر رو هم امبد می‌کردم. شبیه‌ترین دیتای چت‌بات به پرامپت کاربر رو که پیدا کردم دیگه میتونم یک درخواست برای openAI بفرستم. اینجا هم یک سری مسائل به وجود میاد. اگر بخوام با هر پرامپت کاربر کل دیتاهام رو امبد کنم؛ هم خیلی طول می‌کشه، هم هزینه زیادی داره. و اینکه چجوری شبیه‌ترین دیتا به پرامپت کاربر رو پیدا کنم. در ادامه این‌ها رو بررسی می‌کنیم.

ذخیره‌کردن و جست‌وجو در امبدینگ‌ها

زمانی که کاربر چت‌بات‌ساز یک دیتای جدید وارد می‌کنه یا یکی از دیتاهای قبلی رو ویرایش می‌کنه اون دیتا رو امبد می‌کنم. امبد دیتا رو داخل vector database ذخیره می‌کنم. حالا وقتی کاربر یک پرامپت می‌ده، اول پرامپت کاربر رو امبد می‌کنم تا با امبدهای داخل وکتور دیتابیسم مقایسه کنم و شبیه‌ترین دیتا رو پیدا کنم. برای جست‌وجو هم خود پستگرس قابلیتش رو داره.

توهم (hallucination)

یکی از مشکلات جدی مدل‌های زبانی hallucination هست یعنی وقتی یک چیزی رو بلد نیستن شروع میکنن به چرت و پرت جواب دادن. برای جلوگیری از این مشکل راه‌های زیادی هست. من به انتهای پیام کاربر (نه سیستم پرامپت!) متن زیر رو اضافه کردم.

```
usermessage += "\nSTRICTLY Do not give me any information about anything that is not mentioned in the PROVIDED CONTEXT."
```

به‌خاطر آوردن پیام‌های قبلی

کاربران از یک چت‌بات انتظار این رو دارند که واقعا حالت چت داشته‌باشه و چت‌بات پیام‌های گذشته رو به‌خاطر داشته باشه. برای این کار من پرامپت کاربر رو به دو بخش تقسیم کردم. قسمت اول همون پیام حال حاضر کاربر که در قسمت‌های قبل بهش پرداختیم. قسمت دوم تاریخچه‌ی چت هست. برای ذخیره‌کردن تاریخچه‌ی چت از خود openAI استفاده کردم. بعد از اینکه چت‌بات به پرامپت کاربر پاسخ داد. پرامپت کاربر و پاسخ چت‌بات رو کنار هم قرار می‌دم و از مدل‌زبانی می‌خوام که یک خلاصه‌ی یک جمله‌ای از اون مطلب به من بده. خلاصه‌ای که از مدل‌زبانی گرفتم رو به تاریخچه‌ی جست‌وجوم اضافه می‌کنم. در درخواست‌های بعدی تاریخچه‌ی چت رو در ابتدای پرامپت کاربر قرار می‌دم تا مدل‌زبانی پیام‌های قبلی رو به‌خاطر بیاره. سپس باز تاریخچه‌ی چت رو به‌روزرسانی می‌کنم. برای مطالعه‌ی بیشتر از این لینک استفاده کنید.

دیس‌لایک

کاربر اگر پاسخ خوبی دریافت نکرد. بر روی علامت دیس‌لایک مقابل پاسخ‌های چت‌بات کلیک می‌کند. یک جواب دیگر مانند گذشته برای کاربر تهیه می‌کنم. تا به الان کار خاصی برای این قسمت انجام ندادم و فقط از دوباره به openAI ریکوئست می‌زنم به امید این که جواب بهتری تولید کنه. البته تغییراتی برای آینده در نظر دارم. یکی ازراه‌ها این است که به چت‌بات بگیم پاسخت به این سوال رو دوست نداشتم و یک جواب دیگه تولید کن. یا می‌توانیم از کاربر بپرسیم چرا از این پاسخ ناراضی بوده و این دیتا رو هم در اختیار چت‌بات قرار بدیم تا بتونه جواب بهتری تولید کنه.

عنوان چت

برای اینکه چت‌ها عنوان داشته‌باشند، باز هم از خود openAI استفاده کردم. با دادن اولین پیام کاربر به چت‌بات ازش می‌خوام که برام یک عنوان تولید کنه.

فول تکست سرچ

در قسمت چت کاربران می‌توانند در بین پیام‌های موجود جست‌وجو انجام بدهند. برای جست‌وجو از full text search پستگرس استفاده کردم. پستگرس یک سرچ وکتور فیلد داره که با استفاده از اون سرچ وکتور هر پیام رو داخلش ذخیره می‌کنم که لازم نباشه با هر بار جست‌وجو همش رو از اول بسازم. هنگامی که کاربر کوئری خود را وارد کرد برای آن یک کوئری‌سرچ می‌سازم و آن را با وکتورهای پیام‌ها مقایسه می‌کنم. و نتایج به دست‌آمده را بر اساس از رنک بالا به پایین نمایش می‌دم.