

بسم الله الرحمن الرحيم

شرکت مهندسی نرم افزاری هلو

# گزارش تحقیق در خصوص RAG

کاری از امیر علی نسیمی

## فهرست

۲.....	مقدمه‌ای بر هوش مصنوعی و بازیابی-تولید (RAG)
۲.....	مقدمه
۲.....	اهمیت RAG
۳.....	تفاوت RAG با دیگر سیستم‌های بازیابی و تولید
۳.....	نقش این تکنیک در پاسخ‌گویی به سوالات
۴.....	ساختار RAG (بازیابی-تولید)
۴.....	توضیح فرآیند بازیابی اطلاعات
۵.....	نحوه ترکیب بازیابی و تولید در RAG
۶.....	نقش مدل‌های زبان بزرگ (LLMs) در تولید پاسخ
۷.....	یکپارچگی و هماهنگی بازیابی و تولید در RAG
۸.....	تحقیقات انجام شده
۸.....	جدول مقایسه مدل‌های خودرگرسیو (GPT Large، GPT Medium، GPT Extra Large و LLaMA)
۱۰.....	جدول مقایسه مدل‌های Embedding
۱۳.....	خط لوله RAG با استفاده از Embedding و GPT
۱۴.....	اجزای خط لوله RAG
۱۵.....	مراحل خط لوله RAG
۱۶.....	موارد کلیدی برای ساخت خط لوله RAG
۱۷.....	پروژه انجام شده - تبدیل سوالات مربوط به بانک
۱۷.....	پردازش داده‌ها
۱۹.....	آماده‌سازی مستندات
۱۹.....	بارگذاری مدل‌ها
۱۹.....	ساخت پایپ‌لاین پرسش و پاسخ

## مقدمه‌ای بر هوش مصنوعی و بازیابی-تولید (RAG)

### مقدمه

هوش مصنوعی (AI) به سرعت در حال تغییر و تحول در بسیاری از صنایع است. یکی از تکنیک‌های جدید و موثر که توجه محققان و متخصصان را به خود جلب کرده است، تکنیک "بازیابی-تولید" یا به اختصار RAG می‌باشد. RAG به عنوان یک روش ترکیبی از بازیابی اطلاعات و تولید متن، در پاسخ به سوالات پیچیده و کاربردهای مرتبط با پردازش زبان طبیعی (NLP) به کار گرفته می‌شود.

در این تکنیک، سیستم ابتدا اطلاعات مرتبط با پرسش را از منابع بزرگ اطلاعاتی، مانند پایگاه‌های داده یا اینترنت، بازیابی کرده و سپس از یک مدل تولید زبان استفاده می‌کند تا پاسخی مناسب و دقیق ایجاد کند. این ترکیب از بازیابی داده‌های دقیق و تولید محتوا، به سیستم امکان می‌دهد تا نه تنها اطلاعات را از منابع موجود بازیابی کند بلکه بتواند آن را به شیوه‌ای طبیعی و انسانی بازنویسی و تولید کند.

هدف نهایی از انجام این پژوهش، توضیح در خصوص RAG، ساختار کلی آن، نقش مدل‌های LLM در آن می‌باشد. در بخش بعدی تحقیقاتی در خصوص روش‌های LLM انجام شده است و بخشی از آنها مورد بررسی قرار گرفته است. نهایتاً در آخرین بخش به بررسی پروژه مربوط به تبدیل سوالات بانک پرداخته شده است.

### اهمیت RAG

در بسیاری از سیستم‌های سنتی پاسخ‌گویی به سوالات، فرآیند به دو بخش مجزا تقسیم می‌شد: بخش اول به بازیابی اطلاعات مرتبط با سوال اختصاص داشت و بخش دوم وظیفه تولید پاسخ بر اساس اطلاعات موجود را بر عهده داشت. این روش‌ها اغلب با مشکلاتی از جمله دقت پایین در بازیابی اطلاعات یا ناتوانی در تولید پاسخ‌های جامع مواجه بودند.

RAG این مشکلات را با ادغام بازیابی و تولید برطرف می‌کند. این تکنیک به‌ویژه در حوزه‌هایی که نیاز به بازیابی دقیق اطلاعات و تولید پاسخ‌های منعطف و متناسب با سوال وجود دارد، بسیار مفید است. به عنوان مثال، در صنعت پزشکی، RAG می‌تواند اطلاعات حیاتی در مورد بیماری‌ها یا درمان‌ها را از پایگاه‌های داده پزشکی بازیابی کند و سپس به زبان ساده‌ای به کاربران ارائه دهد.

## تفاوت RAG با دیگر سیستم‌های بازیابی و تولید

در حالی که بسیاری از سیستم‌های سنتی تنها به یک بخش از فرآیند پاسخ‌گویی متمرکز بودند، RAG از یک مدل دوگانه بهره می‌برد که نه تنها اطلاعات را به‌طور موثر بازیابی می‌کند بلکه آن‌ها را به شیوه‌ای طبیعی و قابل فهم تولید می‌کند. در واقع، یکی از اصلی‌ترین تفاوت‌های RAG با دیگر سیستم‌ها این است که از دو رویکرد مکمل یعنی بازیابی اطلاعات و تولید زبان طبیعی به‌طور همزمان استفاده می‌کند.

مدل‌های قدیمی‌تر ممکن است تنها به بازیابی دقیق اطلاعات تکیه کنند، اما این اطلاعات بازیابی‌شده ممکن است به‌طور کامل نیاز کاربر را برآورده نکنند. RAG با استفاده از مدل‌های زبان بزرگ مانند BERT یا GPT، اطلاعات بازیابی‌شده را در قالبی مفهومی و معنادار ترکیب و تولید می‌کند.

## نقش این تکنیک در پاسخ‌گویی به سوالات

یکی از کاربردهای اصلی RAG در سیستم‌های پاسخ‌گویی به سوالات است. این سیستم‌ها در حوزه‌های مختلفی از جمله خدمات مشتریان، سیستم‌های جستجو، و حتی دستیارهای مجازی به کار گرفته می‌شوند. در یک سیستم سنتی، پاسخ‌دهی به سوالات اغلب به بازیابی مستقیم از یک منبع متنی وابسته بود که می‌توانست محدود به اطلاعات موجود باشد. اما RAG، با ترکیب داده‌های بازیابی‌شده با توانایی تولید محتوای جدید، می‌تواند پاسخ‌های جامع‌تری ارائه دهد.

به عنوان مثال، در یک سیستم خدمات مشتریان، RAG می تواند به سرعت اطلاعات مربوط به محصولات یا خدمات را از چندین منبع بازیابی کند و سپس آن را به زبانی ساده و قابل فهم برای کاربر نهایی تولید کند. این روش باعث بهبود کیفیت پاسخ ها و افزایش دقت و کارایی سیستم های پاسخ گویی می شود.

## ساختار RAG (بازیابی-تولید)

### توضیح فرآیند بازیابی اطلاعات

بازیابی اطلاعات در سیستم های RAG به عنوان نخستین مرحله از زنجیره تولید پاسخ، نقشی کلیدی ایفا می کند. هدف از بازیابی اطلاعات، یافتن اسنادی است که پاسخ صحیح به سوال مطرح شده را در خود دارند. برای این کار، سیستم باید به سرعت از میان حجم زیادی از داده ها یا اسناد، بهترین و مرتبط ترین اطلاعات را استخراج کند.

در سیستم های سنتی جستجو مانند مدل های TF-IDF یا BM25، واژه ها به طور ساده از متن استخراج و بر اساس فراوانی و اهمیتشان در یک سند رتبه بندی می شدند. با این حال، این روش ها به دلیل ناتوانی در درک معنایی عمیق تر از متن و هم بستگی معنایی میان کلمات، نتایج دقیقی ارائه نمی دادند. در مقابل، مدل های جدید بازیابی اطلاعات برداری (Vector-based Retrieval) مانند DPR (Dense Passage Retrieval)، با استفاده از شبکه های عصبی و نمایش برداری (Embeddings) توانایی تحلیل عمیق تری از معنای سوال و اسناد دارند.

در این روش ها، ابتدا سوال کاربر به بردارهای چندبعدی تبدیل می شود که این بردارها نمایانگر ویژگی های معنایی آن سوال هستند. سپس، سیستم اسناد موجود در پایگاه داده یا منابع مختلف را نیز به صورت بردارهای مشابهی تبدیل می کند. در مرحله بعد، سیستم با محاسبه شباهت های برداری (Vector Similarity) میان سوال و اسناد، مرتبط ترین اسناد را انتخاب و بازیابی می کند.

مدل‌های بازیابی برداری به دلیل توانایی درک و مدل‌سازی معنایی بهتر از متن، توانایی بالاتری در پیدا کردن اطلاعات دقیق دارند. به‌ویژه در سیستم‌های پیچیده‌ای که نیاز به درک عمیق از مفهوم سوالات و اسناد دارند، مانند سوالات مربوط به مباحث علمی یا حقوقی، این مدل‌ها عملکرد بهتری از خود نشان می‌دهند. این فرآیند، مرحله اساسی و اولیه‌ای است که زمینه را برای تولید محتوای نهایی فراهم می‌کند.

## نحوه ترکیب بازیابی و تولید در RAG

سیستم RAG توانایی منحصر به فردی در ترکیب بازیابی و تولید دارد که باعث می‌شود خروجی‌های آن از سیستم‌های سنتی پاسخ‌گویی بسیار متفاوت باشد. ترکیب این دو مرحله به‌طور هم‌زمان و بهینه، باعث ایجاد سیستم‌های پیشرفته‌ای می‌شود که قادرند نه تنها اطلاعات دقیق و مرتبط را بازیابی کنند، بلکه از آن‌ها برای تولید پاسخ‌هایی با زبان طبیعی استفاده کنند.

در مرحله تولید، مدل زبانی مانند GPT یا BERT از اطلاعات بازیابی شده به عنوان ورودی استفاده می‌کند. برخلاف سیستم‌های سنتی که بازیابی و تولید را به‌صورت جداگانه انجام می‌دادند، RAG این دو مرحله را در یک فرآیند هماهنگ و هم‌زمان ادغام می‌کند. این بدان معناست که مدل زبانی مستقیماً از اسناد بازیابی شده برای تولید پاسخ استفاده می‌کند، و نه از دانشی که از پیش در مدل ذخیره شده است. به این ترتیب، پاسخ‌ها دقیق‌تر و منطبق بر آخرین اطلاعات به‌روز موجود در منابع بازیابی شده هستند.

RAG دو روش کلیدی برای ترکیب بازیابی و تولید دارد:

### ۱. RAG-Sequence:

در این روش، ابتدا سیستم تمامی اسناد مرتبط با سوال را بازیابی می‌کند و سپس مدل تولید زبان از آن‌ها به‌عنوان یک بسته کامل ورودی برای تولید پاسخ استفاده می‌کند. این روش به مدل تولید زبان اجازه می‌دهد که تمامی

اطلاعات موجود را یک جا دریافت و پردازش کند و پاسخ نهایی را به طور کامل بر اساس آن تولید کند. این روش برای سوالات پیچیده که نیاز به تجمیع اطلاعات از چندین منبع دارند مناسب است.

## ۲. RAG-Token:

این روش پیچیده تر است و در آن مدل تولید زبان در حین تولید هر قسمت از پاسخ، به طور پویا اطلاعات بیشتری از منابع بازیابی شده دریافت می کند. این روش به مدل تولید زبان انعطاف بیشتری می دهد تا بر اساس هر نشانه (Token) یا بخش از پاسخ، از اطلاعات جدیدتر و دقیق تر استفاده کند. به عبارتی، سیستم به طور مداوم در حین تولید پاسخ، اسناد بازیابی شده را مرور می کند و از آن ها برای تولید هر بخش از پاسخ استفاده می کند. این رویکرد به تولید پاسخ های دقیق تر و منطقی تر کمک می کند، زیرا مدل زبانی می تواند در صورت نیاز به جزئیات بیشتری دست یابد.

## نقش مدل های زبان بزرگ (LLMs) در تولید پاسخ

مدل های زبان بزرگ (LLMs) مانند GPT (Generative Pre-trained Transformer)، BERT (Bidirectional Encoder Representations from Transformers) و T5 (Text-to-Text Transfer Transformer) از پیشرفته ترین مدل های پردازش زبان طبیعی هستند که توانایی تولید متون طبیعی و شبیه به انسان را دارند. این مدل ها در RAG برای مرحله تولید متن پس از بازیابی اطلاعات مورد استفاده قرار می گیرند.

GPT یکی از مشهورترین مدل های زبانی است که توسط شرکت OpenAI توسعه یافته است. این مدل از معماری Transformer بهره می برد و بر اساس یک پایگاه داده بزرگ از متون عمومی و تخصصی آموزش دیده است. مدل های مبتنی بر GPT می توانند بر اساس ورودی های داده شده (مانند اسناد بازیابی شده)، متنی جدید و طبیعی تولید کنند که شباهت زیادی به نوشتار انسانی دارد. در RAG، GPT اطلاعات بازیابی شده را تجزیه و تحلیل کرده و سپس بر اساس آن ها یک پاسخ تولید می کند که از نظر ساختار زبانی و معنایی بسیار شبیه به متون نوشته شده توسط انسان است.

BERT نیز یکی دیگر از مدل‌های مهم در حوزه پردازش زبان طبیعی است که برخلاف GPT از رویکرد دوسویه برای درک متن استفاده می‌کند. این ویژگی به BERT اجازه می‌دهد که متن را با توجه به بافت کل جمله و معنای کلی آن بهتر درک کند. در سیستم RAG، از BERT به عنوان یک مدل پایه برای بازیابی اطلاعات دقیق و متناسب استفاده می‌شود. این مدل می‌تواند سوالات پیچیده را بهتر درک کند و به پاسخ‌های دقیق‌تری دست یابد.

T5 نیز یکی دیگر از مدل‌های زبانی است که تمرکز اصلی آن بر تبدیل وظایف مختلف زبانی به فرمت‌های متنی مشابه است. این مدل قادر است تا وظایف مختلف مانند ترجمه، خلاصه‌سازی، و پاسخ‌گویی به سوالات را به‌طور یکپارچه انجام دهد. در سیستم RAG، T5 می‌تواند برای تولید پاسخ‌هایی که نیاز به تحلیل و تفسیر دقیق متن دارند به کار گرفته شود.

## یکپارچگی و هماهنگی بازیابی و تولید در RAG

یکی از ویژگی‌های کلیدی سیستم RAG این است که به جای اتکا به تنها یک بخش از پردازش، از یکپارچگی میان بازیابی اطلاعات و تولید پاسخ بهره می‌برد. این روش دو مزیت اصلی دارد:

۱. دقت در اطلاعات بازیابی‌شده: با استفاده از تکنیک‌های برداری و مدل‌های عصبی، سیستم می‌تواند اسنادی را بازیابی کند که با معنای سوال کاملاً همخوانی دارند.

۲. طبیعی بودن پاسخ‌ها: مدل‌های زبانی بزرگ مانند GPT و BERT باعث می‌شوند که پاسخ‌های تولیدشده از نظر ساختار زبانی و معنایی کاملاً قابل درک و طبیعی باشند.

این یکپارچگی به RAG اجازه می‌دهد که در صنایع مختلف و برای کاربردهای مختلف، از جمله پرسش و پاسخ در سیستم‌های خدمات مشتریان، تحقیق و جستجوی اطلاعات علمی، و تولید محتوای تعاملی به کار گرفته شود.



## تحقیقات انجام شده

### جدول مقایسه مدل های خودرگرسیو (GPT Medium, GPT Large, GPT Extra Large و LLaMA)

در این بخش، به مقایسه مدل های خودرگرسیو (Autoregressive) مختلف شامل GPT در اندازه های Medium، Large، Extra Large و مدل LLaMA (Large Language Model Meta AI) پرداخته خواهد شد. این مقایسه بر اساس معیارهای کلیدی مانند تعداد پارامترها، معماری، عملکرد، و کاربردهای اصلی ارائه می شود. معیار ارزیابی ارائه شده ROUGE می باشد که جهت بررسی کیفیت متن های تولید شده توسط مدل است. ارزیابی میزان همپوشانی بین متن تولید شده و متن مرجع در این خصوص اهمیت دارد.

مدل	تعداد پارامتر	معماری	حجم داده های آموزش	قدرت پردازش و عملکرد	کاربردها	ویژگی های کلیدی	ارزیابی جمله فارسی	ارزیابی گزارش شده رسمی
GPT Medium	345 میلیون پارامتر	Transformer-based	داده های متنی عمومی و علمی	عملکرد خوب در تولید متن	تولید محتوا، تکمیل متن، چت بات	سرعت پردازش خوب، مناسب برای کاربردهای ساده و عمومی	خوب	۷۰٪
GPT Large	762 میلیون پارامتر	Transformer-based	داده های متنوع تر و گسترده	قدرت پردازش بیشتر نسبت به مدل Medium	کاربردهای تخصصی تر مانند ترجمه، تحلیل زبان طبیعی	دقت بیشتر در تحلیل و تولید متن	به علت حجم زیاد ارزیابی نشد	۷۸٪

<b>GPT Extra Large</b>	1.5 میلیارد پارامتر	Transformer-based	داده‌های بسیار گسترده و متنوع	عملکرد بسیار قوی در تولید متن و تکمیل	پردازش زبان طبیعی پیشرفته، سوال و جواب‌های پیچیده	مناسب برای کاربردهای پیچیده و تخصصی	به علت حجم زیاد ارزیابی نشد	۸۴٪
<b>(7B) LLaMA</b>	7 میلیارد پارامتر	Transformer-based	داده‌های گسترده علمی، عمومی، و تخصصی	عملکرد بسیار قوی در تولید متن، سرعت بالا	تحقیق و توسعه، تولید محتوا در مقیاس بزرگ، سوال و جواب پیشرفته	مدل کوچک‌تر و کارآمدتر نسبت به دیگر مدل‌ها با تعداد پارامتر بالا	عالی	۸۰٪
<b>(13B) LLaMA</b>	13 میلیارد پارامتر	Transformer-based	داده‌های گسترده و تخصصی‌تر	قدرت پردازش بسیار بالا	کاربردهای پیچیده و تحقیقاتی، پردازش داده‌های حجیم	کارایی بالاتر نسبت به نسخه‌های قبلی، مناسب برای کاربردهای پیشرفته	به علت حجم زیاد ارزیابی نشد	۸۵٪

### ۱. GPT Medium:

این مدل با ۳۴۵ میلیون پارامتر، به عنوان یک مدل خودرگرسیو در اندازه متوسط، عملکرد خوبی در تولید متن و پردازش زبان طبیعی دارد. این مدل برای کاربردهای عمومی مانند تولید محتوا و پاسخ‌گویی به سوالات ساده مناسب است. از نظر سرعت پردازش و کارایی، مدل Medium گزینه‌ای اقتصادی برای کارهای روزمره است.

### ۲. GPT Large:

با ۷۶۲ میلیون پارامتر، مدل GPT Large دقت و توانایی پردازش بیشتری نسبت به مدل Medium دارد. این مدل برای کاربردهای تخصصی‌تر مانند ترجمه زبان، تحلیل متن‌های پیچیده، و چت‌بات‌های پیشرفته مناسب است. در مقایسه با مدل Medium، GPT Large در پردازش داده‌های متنوع و حجم بالاتر عملکرد بهتری دارد.

### ۳. GPT Extra Large:

این مدل با ۱.۵ میلیارد پارامتر، به طور قابل توجهی از نظر اندازه و توانایی پردازش قوی تر از مدل های قبلی است. این مدل برای سوالات پیچیده، تولید محتوا در مقیاس بزرگ، و تحلیل های پیشرفته در زمینه پردازش زبان طبیعی استفاده می شود. GPT Extra Large به دلیل قدرت پردازش بالا و دقت زیاد، برای کاربردهای پیچیده و تحقیقاتی بسیار مناسب است.

### ۴. LLaMA (7B):

مدل LLaMA با ۷ میلیارد پارامتر، یکی از مدل های پیشرفته ای است که توسط Meta AI توسعه یافته است. با وجود تعداد پارامترهای بسیار بالاتر از GPT Extra Large، این مدل بهینه سازی های بیشتری در کارایی و پردازش ارائه می دهد. این مدل برای تحقیقات پیشرفته، تولید محتوای پیچیده، و سوال و جواب های پیچیده بسیار مناسب است.

### ۵. LLaMA (13B):

با ۱۳ میلیارد پارامتر، نسخه پیشرفته تر مدل LLaMA است که عملکرد بسیار قوی تری در پردازش حجم های عظیم داده دارد. این مدل برای پردازش داده های پیچیده تر و تحقیقاتی مانند تحلیل متون تخصصی و علمی استفاده می شود و از کارایی بالایی برخوردار است. به عنوان یکی از مدل های بزرگ زبان، LLaMA 13B به ویژه در حوزه های تحقیق و توسعه پیشرفته مورد استفاده قرار می گیرد.

## جدول مقایسه مدل های Embedding

در این بخش، به مقایسه مدل های برجسته embedding مانند Zephyr, MPNet-all, LaBSE و دیگر مدل های معروف پرداخته می شود. این مدل ها ابزارهای قدرتمندی برای نمایش معنایی کلمات و جملات به صورت برداری در فضای عددی هستند و به ویژه در پردازش زبان طبیعی (NLP) استفاده می شوند. جهت ارزیابی از ضریب همبستگی

پیرسون استفاده می شود؛ یک معیار آماری است که قدرت و جهت رابطه خطی بین دو متغیر پیوسته را ارزیابی می کند. این ضریب مقداری بین -۱ تا ۱ می گیرد که:

- ۱ نشان دهنده یک رابطه خطی کاملاً مثبت است (وقتی یکی از متغیرها افزایش می یابد، متغیر دیگر نیز به طور متناسب افزایش می یابد).
- ۰ یعنی هیچ رابطه خطی وجود ندارد (تغییرات در یک متغیر پیش بینی کننده تغییرات در متغیر دیگر نیست).
- -۱ نشان دهنده یک رابطه خطی کاملاً منفی است (وقتی یکی از متغیرها افزایش می یابد، متغیر دیگر به همان نسبت کاهش می یابد).

در زمینه یادگیری ماشین ضریب همبستگی پیرسون برای سنجش میزان مطابقت پیش بینی های مدل با نتایج واقعی (مقادیر صحیح) استفاده می شود، به ویژه در وظایفی مانند:

- شباهت جملات: مقایسه نمرات شباهت پیش بینی شده بین جفت های جملات با نمرات شباهت ارزیابی شده توسط انسان ها.
- وظایف رگرسیون: اندازه گیری میزان تطابق مقادیر عددی پیش بینی شده با مقادیر واقعی.

برای بردار سازی جملات، ضریب همبستگی پیرسون اندازه گیری می شود که شباهت پیش بینی شده (مثلاً شباهت کسینوسی بین بردار ها) تا چه حد با قضاوت های انسانی درباره شباهت جملات همخوانی دارد. ضریب همبستگی پیرسون بالا (نزدیک به ۱) نشان می دهد که خروجی های مدل با قضاوت انسانی یا داده های واقعی به خوبی تطابق دارند.

مدل	تعداد پارامترها	معماری	حجم داده های آموزش	قدرت پردازش و عملکرد	معیار ارزیابی Pearson correlations
MPNet-all	Sentence Transformer	768	مشابه سازی جملات،	چند زبانه، پشتیبانی از	0.88
				دقت بسیار بالا در مشابه سازی جملات و مفاهیم، ترکیب	

			جستجوی معنایی	زبان‌های مختلف	دو روش ماسک و ترتیب‌گذاری
<b>LaBSE</b>	Sentence Transformer	768	جستجوی چند زبان، ترجمه	بیش از ۱۰۰ زبان	مدل چند زبانه با توانایی بسیار بالا در ترجمه و یافتن معانی در زبان‌های مختلف
<b>USE</b>	Transformer- based, Deep Averaging Network	512	پردازش زبان طبیعی، جستجوی متون	انگلیسی و چند زبان دیگر	مدل عمومی برای نمایش جمله‌ای، بهینه‌سازی شده برای متن‌های عمومی و محتوای جستجو
<b>SBERT (Sentence-BERT)</b>	Sentence Transformer	768	مقایسه جملات، رتبه‌بندی متن	چند زبانه، زبان‌های مختلف	مدل سفارشی‌سازی شده برای دقت بالا تر در مشابه‌سازی و جستجوی معنایی
<b>DistilBERT</b>	Transformer (distilled)	768	نمایش معنایی و طبقه‌بندی	انگلیسی و برخی زبان‌های دیگر	مدل سبک‌تر با کارایی بالا برای کاربردهای زمان واقعی و بهینه در مصرف حافظه

## 2- MPNet-all (all-mpnet-base-v2):

این مدل یکی از بهترین مدل‌های embedding برای مشابه‌سازی معنایی جملات است MPNet. از روش‌های ماسک و ترتیب‌دهی استفاده می‌کند و با ابعاد برداری ۷۶۸، دقت بسیار بالایی در مشابه‌سازی متون ارائه می‌دهد. این مدل برای جستجوی معنایی و مقایسه جملات به خوبی عمل می‌کند و پشتیبانی چندزبانه‌ای دارد.

## 3 - LaBSE (Language-agnostic BERT Sentence Embedding):

مدل LaBSE یک مدل BERT چندزبانه است که برای جستجوهای معنایی و ترجمه متون استفاده می‌شود. این مدل با ابعاد ۷۶۸، از بیش از ۱۰۰ زبان پشتیبانی می‌کند و به دلیل قابلیت‌های چندزبانه و قدرت پردازش در یافتن معانی در زبان‌های مختلف، در کاربردهای جهانی محبوب است.

#### 4- USE (Universal Sentence Encoder):

مدل Universal Sentence Encoder (USE) یکی از مدل‌های ساده‌تر و کاربرپسندتر برای ایجاد embeddingهای جملات است. با ابعاد ۵۱۲، این مدل برای کاربردهای عمومی پردازش زبان طبیعی و جستجو مناسب است. استفاده از USE ساده و سریع است، و برای بسیاری از کاربردهای پایه‌ای توصیه می‌شود.

#### 5- SBERT (Sentence-BERT):

مدل SBERT بر پایه BERT ساخته شده است و برای دقت بالاتر در مشابه‌سازی جملات به کار می‌رود. این مدل با ابعاد ۷۶۸، توانایی بالایی در مقایسه جملات و رتبه‌بندی متن دارد. SBERT، به خصوص برای کاربردهایی که نیاز به جستجوی معنایی دقیق دارند مناسب است.

#### 6- DistilBERT:

DistilBERT یک نسخه سبک‌تر و سریع‌تر از BERT است. این مدل با ابعاد ۷۶۸، برای کاربردهایی که نیاز به سرعت و مصرف بهینه حافظه دارند بسیار مناسب است. DistilBERT به دلیل کاهش تعداد پارامترها همچنان دقت خوبی دارد و برای کاربردهای زمان واقعی (real-time) استفاده می‌شود.

## خط لوله RAG با استفاده از Embedding و GPT

برای ایجاد یک خط لوله مبتنی بر RAG (تولید مبتنی بر بازیابی اطلاعات) با استفاده از یک مدل embedding و GPT، باید نحوه عملکرد RAG را درک کنیم و همچنین نحوه تعامل اجزای آن یعنی مدل embedding (برای

بازیابی اطلاعات) و GPT (برای تولید متن) را بفهمیم. هدف اصلی این خط لوله، بازیابی اطلاعات مرتبط از یک پایگاه داده یا دانش و سپس تولید پاسخ‌هایی طبیعی و دقیق بر اساس اطلاعات بازیابی شده است.

## اجزای خط لوله RAG

خط لوله RAG از دو بخش اصلی تشکیل شده است:

### ۱. بازیابی اطلاعات (Document Retriever) با استفاده از مدل Embedding

این بخش وظیفه پیدا کردن مرتبط‌ترین اسناد یا پاراگراف‌ها از یک پایگاه داده بزرگ را دارد. این کار با استفاده از یک مدل embedding که هم پرسش‌ها (ورودی‌های کاربر) و هم اسناد را به صورت بردارهایی در فضای معنایی نمایش می‌دهد انجام می‌شود. مراحل اصلی در این بخش شامل موارد زیر است:

- مرحله ۱: تولید بردار پرسش: زمانی که یک پرسش مطرح می‌شود، مدل embedding این پرسش را به یک بردار چگال (dense vector) تبدیل می‌کند. این کار معمولاً توسط مدل‌هایی مانند MPNet، LaBSE، USE، یا SBERT انجام می‌شود که جملات را به بردارهای چندبعدی تبدیل می‌کنند.

- مرحله ۲: تولید بردار اسناد: تمامی اسناد موجود در پایگاه دانش به صورت پیش‌پردازش شده و به صورت embedding ذخیره می‌شوند. این کار امکان مقایسه کارآمد آن‌ها با بردار پرسش را فراهم می‌کند.

- مرحله ۳: جستجوی شباهت (بازیابی): پس از تولید بردار پرسش، سیستم جستجوی شباهت (مانند شباهت کسینوسی یا ضرب داخلی) را برای یافتن مرتبط‌ترین بردارهای سند انجام می‌دهد. مرتبط‌ترین اسناد انتخاب و به مرحله بعد منتقل می‌شوند.

## ۲. تولید زبان (Language Generator) با استفاده از GPT

پس از بازیابی اسناد مرتبط توسط embedder، مرحله بعد شامل تولید پاسخ طبیعی زبان بر اساس آن اسناد است. در این مرحله از مدل GPT استفاده می‌شود. مراحل شامل:

– مرحله ۴: ورودی اسناد به GPT: اسناد بازیابی شده به همراه پرسش اصلی کاربر به عنوان زمینه (context) به مدل GPT داده می‌شوند. این کار به GPT اطلاعات لازم برای تولید پاسخ‌های معنادار را فراهم می‌کند.

– مرحله ۵: تولید پاسخ: GPT بر اساس پرسش کاربر و زمینه اسناد بازیابی شده، پاسخ طبیعی تولید می‌کند. چون GPT یک مدل خودرگرسیو است، کلمات بعدی را براساس کلمات قبلی پیش‌بینی می‌کند و در نتیجه پاسخ منسجم و روان تولید می‌شود.

## مراحل خط لوله RAG

۱. ورودی پرسش:

– کاربر یک سوال یا درخواست مطرح می‌کند

۲. تبدیل پرسش به بردار:

– پرسش از طریق یک مدل embedding (مثل MPNet یا LaBSE) به یک بردار چگال تبدیل می‌شود.

۳. بازیابی اسناد:



- سیستم جستجوی شباهت بین بردار پرسش و embedding های از پیش محاسبه شده اسناد را انجام داده و مرتبط ترین اسناد را بازیابی می کند.

۴. ایجاد زمینه اسناد:

- اسناد بازیابی شده با هم ترکیب شده و به عنوان زمینه به مدل GPT داده می شوند.

۵. تولید پاسخ:

- GPT براساس ورودی های ارائه شده، پاسخ طبیعی و منسجمی تولید می کند که به پرسش کاربر پاسخ می دهد.

۶. خروجی:

- پاسخ تولید شده توسط GPT به عنوان پاسخ نهایی به کاربر بازگردانده می شود.

## موارد کلیدی برای ساخت خط لوله RAG

۱. انتخاب مدل Embedder مناسب:

- انتخاب مدل مناسب برای embedding می تواند تأثیر زیادی بر عملکرد بخش بازیابی اطلاعات داشته باشد. برای کاربردهای چندزبانه، LaBSE ممکن است مناسب تر باشد، در حالی که برای مشابه سازی دقیق جملات، MPNet یا SBERT گزینه های بهتری هستند.

۲. بهینه سازی کارایی:

- محاسبه و ذخیره از پیش embedding های اسناد و استفاده از تکنیک های جستجوی سریع (مثل FAISS) می تواند سرعت بازیابی اطلاعات را حتی برای مجموعه داده های بزرگ افزایش دهد.

۳. تنظیم GPT برای دامنه های خاص:

- برای کاربردهای خاص، ممکن است نیاز به تنظیم دقیق مدل GPT روی داده‌های مرتبط با دامنه خاص باشد. این کار باعث می‌شود پاسخ‌ها دقیق‌تر و مناسب‌تر باشند.

۴. مدیریت طول زمینه در GPT:

- تعداد اسناد بازایی شده که به GPT داده می‌شوند باید در چارچوب محدودیت‌های طول توکن (مثلاً ۴۰۹۶ توکن برای GPT-3) باشند. انتخاب بیش از حد اسناد طولانی ممکن است منجر به حذف بخش‌هایی از زمینه شود، بنابراین تعادل بین مرتبط بودن اسناد و دقت پاسخ ضروری است.

## پروژه انجام شده - تبدیل سوالات مربوط به بانک

در این پروژه، ما به بررسی یک سیستم پرسش و پاسخ (Q&A) مبتنی بر مدل‌های یادگیری عمیق و استفاده از LangChain می‌پردازیم:

### پردازش داده‌ها

در مرحله اول، داده‌ها از یک فایل JSON بارگذاری می‌شوند. داده‌ها شامل سوالات و پاسخ‌های مختلف است که به فرمت مورد نیاز برای استفاده در مدل‌های یادگیری ماشین تبدیل می‌شود. بخشی از این سوال و جواب‌ها به شرح زیر است:

۱. ایجاد حساب کاربری:

پرسش: چگونه می‌توانم یک حساب کاربری ایجاد کنم؟

پاسخ: برای ایجاد حساب کاربری، روی دکمه 'Sign Up' در گوشه بالا سمت راست وبسایت کلیک کنید و دستورالعمل‌های ثبت‌نام را دنبال کنید.

۲. روش‌های پرداخت:

پرسش: چه روش‌های پرداختی را قبول می‌کنید؟

پاسخ: ما کارت‌های اعتباری اصلی، کارت‌های بدهی و PayPal را برای سفارش‌های آنلاین قبول می‌کنیم.

۳. پیگیری سفارش:

پرسش: چگونه می‌توانم سفارش خود را پیگیری کنم؟

پاسخ: با ورود به حساب کاربری و مراجعه به بخش 'Order History' می‌توانید سفارش خود را پیگیری کنید.

۴. سیاست بازگشت کالا:

پرسش: سیاست بازگشت کالاهای شما چیست؟

پاسخ: ما اجازه می‌دهیم تا محصولات را ظرف ۳۰ روز از تاریخ خرید بازگردانید و مبلغ کامل را بازپرداخت کنیم، مشروط بر اینکه محصولات در شرایط اصلی و بسته‌بندی خود باشند.

۵. لغو سفارش:

پرسش: آیا می‌توانم سفارش خود را لغو کنم؟

پاسخ: بله، اگر سفارش هنوز ارسال نشده باشد، می‌توانید آن را لغو کنید.

۶. مدت زمان ارسال:

پرسش: مدت زمان ارسال چقدر است؟

پاسخ: مدت زمان ارسال به مقصد و روش ارسال انتخاب شده بستگی دارد. ارسال استاندارد معمولاً ۳-۵ روز کاری طول می‌کشد، در حالی که ارسال فوری ممکن است ۱-۲ روز کاری زمان ببرد.

با توجه به تعداد زیاد سوال ها و جواب ها، از ادامه آن ها خودداری شده است.

## آماده سازی مستندات

LangChain برای پردازش و جستجو در این مستندات استفاده می شود لذا پس از پردازش داده ها، کلیه این موارد در قالب مشخص به این فریم ورک داده می شود.

## بارگذاری مدل ها

مدل های یادگیری عمیق، به ویژه مدل های بزرگ زبانی (مثل Zephyr-7b)، بارگذاری می شوند. این مدل ها برای پردازش و تولید پاسخ ها استفاده می شوند. تنظیمات مختلفی برای بهینه سازی عملکرد مدل ها و کاهش مصرف حافظه مورد استفاده قرار گرفته اند.

## ساخت پایپ لاین پرسش و پاسخ

یک پایپ لاین برای پرسش و پاسخ ایجاد می شود که شامل مدل زبانی و سیستم جستجوی مستندات است. این پایپ لاین به کاربران امکان می دهد تا سوالات خود را بپرسند و پاسخ های مناسب را دریافت کنند. جهت ساخت این پایپ لاین از langchain استفاده شده است.

## ارزیابی نتایج

جهت استفاده از بخش Embedder از مدل HuggingFaceEmbeddings استفاده شده است که در حقیقت از SBert استفاده می کند که قبلا در بخش ارزیابی مدل های Embedding در خصوص این مدل انتخابی توضیح

داده شده است. جهت ذخیره سازی ویژگی های خروجی از نمونه ها از Chroma استفاده شده است که نوعی پایگاه داده مناسب ذخیره سازی بردار های ویژگی می باشد.

مدل LLM استفاده شده در این حالت Zephyr با ۷ میلیارد پارامتر می باشد. مدل Zephyr یک مدل جدید برای ایجاد نمایش های معنایی از جملات است. ابعاد برداری آن ۵۱۲ است و برای کاربردهایی مانند تحلیل احساسات و ترجمه متون به کار می رود. دارای ۷ میلیارد پارامتر است و این مدل از قابلیت های چندزبانای برخوردار است و می تواند برای کاربردهای پیچیده پردازش زبان طبیعی استفاده شود. دقت این مدل ۸۴ درصد گزارش شده است.

مدل HuggingFacePipeline بوده که کاملاً بستگی به در نهایت از مدل RetrievalQA از Langchain استفاده شده است. این مدل توانایی نسبتاً خوبی در خصوص یافتن جملات دارد. دقت آن بر اساس معیار تطبیق دهی جملات انگلیسی، بین ۷۰ تا ۹۰ درصد اندازه گیری شده است. جهت انجام بررسی برای جملات فارسی، تعداد ۲۶ سوال و جواب به فارسی درست شده است. دقت نهایی این روش برای جملات فارسی مناسب نمی باشد. لذا استفاده از مدل هایی که برای زبان فارسی در نظر گرفته شده است توصیه می شود.