
# ریز تنظیم یک مدل زبانی بزرگ به منظور طراحی مدل پرسش و پاسخ فارسی

**علی باقرزاده**
دانشجوی کارشناسی دانشگاه علم و صنعت ایران
دانشکده مهندسی کامپیوتر

استاد راهنما:
دکتر بهروز مینایی

با تشکر از زحمات فراوان:
جناب آقای دکتر حسینی
سرکار خانم صراف

## چکیده:

مدل‌های زبانی بزرگ چند زبانه، نه‌تنها یکی از مهم‌ترین اختراعات در زمینه هوش مصنوعی و پردازش زبان طبیعی هستند، بلکه این ابتکار نیز توانسته است تا حد زیادی مسائل چالش‌برانگیز مرتبط با زبان‌های مختلف را حل کند. در این چکیده، ما به معرفی مدل‌های زبانی بزرگ چند زبانه و کاربردهای آن‌ها در تسک Question Answering به زبان فارسی می‌پردازیم.

این مدل‌ها بر پایه معماری BERT (Bidirectional Encoder Representations from Transformers) ساخته شده و توانایی درک و تولید متن در زبان‌های مختلف را دارند. به عبارت دیگر، آن‌ها قادرند اطلاعات عمیقی از زبان‌های مختلف را یاد بگیرند و در ترجمه، خلاصه‌سازی متون و حتی تسک‌های سوال و جواب به زبان‌های متعدد عملیاتی باشند.

در مورد زبان فارسی، مدل‌های BERT مثل ParsBERT توانسته‌اند بهبود قابل ملاحظه‌ای در تسک Question Answering داشته باشند. این مدل‌ها می‌توانند از متون فارسی استفاده کنند و پاسخ‌های دقیق به سوالاتی که از آن‌ها پرسیده می‌شود ارائه دهند.

به طور خلاصه، مدل‌های زبانی بزرگ چند زبانه مانند BERT با امکانات برتر در زمینه پردازش زبان طبیعی، تسک Question Answering به زبان فارسی را به مرحله جدیدی از دقت و اطمینان ارتقاء داده‌اند. این ابتکارات امکان پژوهش‌های آینده در این زمینه را باز کرده و در کاربردهای گسترده‌تری مفید خواهند بود

در این مقاله من از روند تولید یک دیتاست تا train کردن آن بر روی یک مدل pretrained و قدرت بخشیدن به آن به منظور استفاده از آن در زمینه پرسش و پاسخ در زبان فارسی صحبت خواهم کرد.

## مقدمه:

سوال و پاسخ یکی از وظایف مهم در پردازش زبان طبیعی است که در آن سیستم‌هایی توسعه داده می‌شوند که به صورت خودکار پاسخ یا اطلاعات مورد نیاز برای سوالات داده شده از متن‌های موجود تولید یا بازیابی می‌کنند. بر خلاف سیستم‌های جستجویی که مجموعه‌ای از مستندات مرتبط را تولید می‌کنند، مدل‌های سوال و پاسخ پاسخ را از پایگاه دانش تولید می‌کنند. این سیستم‌ها در مختلف حوزه‌ها از استخراج اطلاعات استفاده می‌کنند و به دلیل رشد در حوزه فنی، اهمیت بزرگی پیدا کرده‌اند.

این سیستم‌ها در روش‌های تولید پاسخ متفاوتی تفاوت دارند، مانند: سوال و پاسخ برداشتی و سوال و پاسخ تولیدی. مدل‌های سوال و پاسخ برداشتی پاسخ‌ها را مستقیماً از پایگاه دانش داده شده تولید می‌کنند و معمولاً از مدل‌های مبتنی بر ترانسفورمر مانند BERT استفاده می‌کنند. در این رویکرد، پاسخ از متن اصلی انتخاب یا کپی می‌شود و به عنوان نشانه کلمه برای پاسخ برداشته‌شده عمل می‌کند. از سوی دیگر، مدل‌های سوال و پاسخ تولیدی پاسخ‌های متن آزادی را بر اساس اطلاعات متنی تولید می‌کنند و از مدل‌های تولید متن استفاده می‌کنند. در این حالت، پاسخ از ابتدا تولید می‌شود و به هیچ قسمت خاصی از متن اصلی محدود نمی‌شود. به عبارت دیگر، این پاسخ به صورت خودکار ایجاد می‌شود و از رویکرد برداشتی متمایز می‌شود.

علاوه بر تفاوت میان مدل‌های سوال و پاسخ برداشتی و تولیدی، سیستم‌های سوال و پاسخ می‌توانند بر اساس حوزه عملکردشان دسته‌بندی شوند. این تقسیم‌بندی شامل تفکیک بین سیستم‌های سوال و پاسخ باز و بسته است. سیستم‌های سوال و پاسخ باز پاسخ را از متن یا پایگاه دانش گسترده‌تری بازیابی می‌کنند. به عبارت دیگر، آن‌ها محدود به پاسخ‌های مشخص پیش‌تعریف نشده نیستند و می‌توانند پاسخ‌هایی ارائه دهند که به طور صریح در متن اصلی وجود ندارند. این سیستم‌ها انعطاف‌پذیرتر هستند و قادر به پردازش یک طیف گسترده از سوالات هستند، حتی آن‌هایی که در دوره آموزش با آن‌ها

روبرو نشده‌اند. از سوی دیگر، سیستم‌های سوال و پاسخ بسته برای تولید پاسخ‌ها تنها بر اساس دانش موجود در داده‌های آموزشی آن‌ها طراحی شده‌اند. آن‌ها محدود به ارائه پاسخ‌ها در یک مجموعه پیش‌تعریف شده از پاسخ‌های ممکن هستند. بنابراین، سیستم‌های سوال و پاسخ بسته ممکن است برای موارد کاربردی خاص با سوالات و پاسخ‌های خوب تعریف‌شده و ساختاری مناسب‌تر مناسب باشند. به طور خلاصه، سیستم‌های سوال و پاسخ باز دامنه گسترده‌تری دارند و می‌توانند پاسخ‌ها را خارج از داده‌های آموزشی خود تولید کنند، در حالی که سیستم‌های سوال و پاسخ بسته محدودتر هستند و به دامنه‌های خاص با مجموعه‌های پیش‌تعریف شده از پاسخ‌ها مناسب هستند.

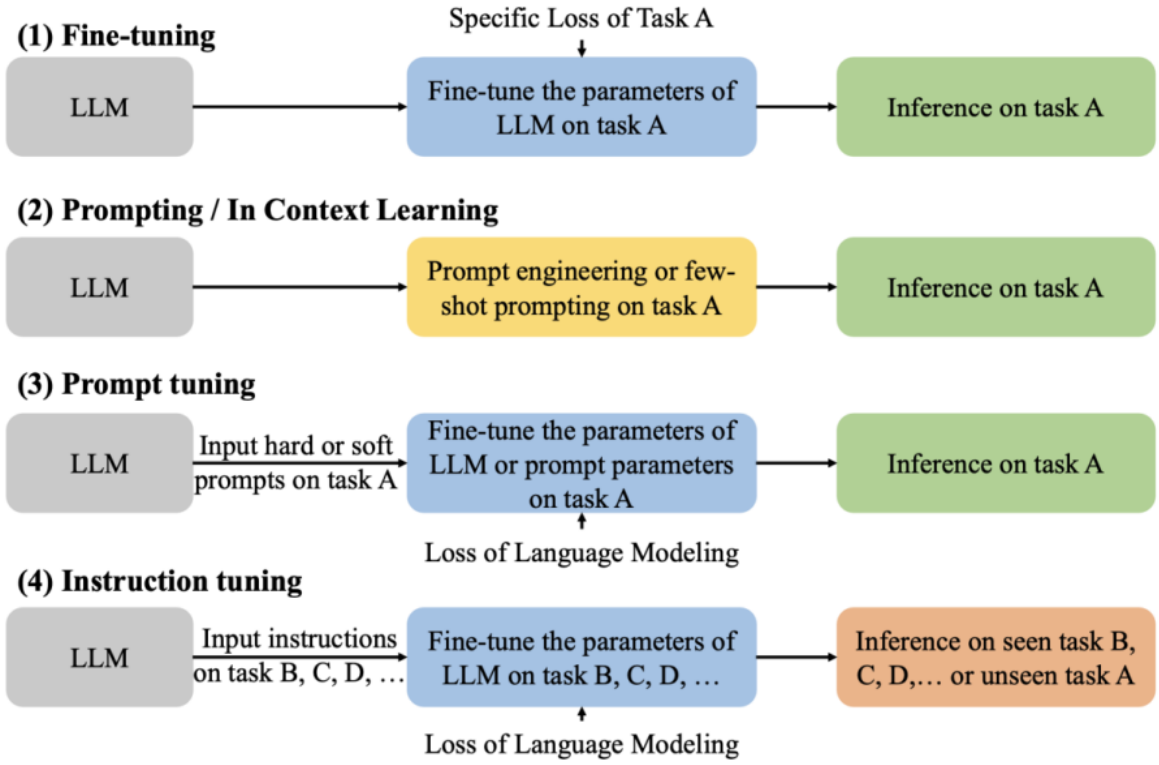
مدل‌های سوال و پاسخ اغلب برای زبان انگلیسی ساخته و آموزش داده می‌شوند و با یک مجموعه داده انگلیسی بزرگ بهینه‌سازی می‌شوند (مانند SQuAD (Rajpurkar و همکاران 2016)). چنین مجموعه‌داده‌هایی برای زبان فارسی در دسترس نیستند و به همین دلیل به عنوان زبان‌های منابع کم شناخته می‌شوند. کمبود کافی از کرپوراها به پیشرفت مدل‌های پردازش زبان طبیعی برای این زبان‌های منابع کم مانعی می‌شود.

ایران جمعیتی بالغ بر 85 میلیون نفر دارد. زبان فارسی توسط نزدیک به 150 میلیون نفر در دنیا صحبت می شود. با وجود محبوبیت این زبان‌ها، سیستم‌های پردازش زبان طبیعی در این زبان‌ها بی‌بررسی مانده‌اند.
بهترین دانش نویسندگان، این مجموعه داده را بزرگترین مجموعه داده سوال و پاسخ برای زبان فارسی معرفی می‌کنند که با ترجمه SQuAD 2.0 ایجاد شده است. رویکرد ما به حل مشکل در تعیین دقیق شماره پاسخ در سیاق می‌پردازد. مشارکت‌های ما می‌تواند به شرح زیر خلاصه شود:
• ما مجموعه بزرگی از داده سوال و پاسخ را برای زبان فارسی منتشر می‌کنیم.

• ما مدل مناسب و عملکرد عالی برای سوال و پاسخ را بر اساس ارزیابی و آزمایش جامع منتشر می‌کنیم.

## روش های مختلف تنظیم مجدد یک مدل زبانی بزرگ:

برای تنظیم مجدد یک مدل زبانی بزرگ ما روش های متفاوتی داریم که برخی از این روش ها در تصویر زیر اشاره شده است.




تصویر 1: توضیح با جزئیات پنج روش آموزشی مختلف (انطباق دامنه)

تنظیم نهایی یک تکنیک آموزشی استفاده شده در هوش مصنوعی (AI) به منظور بهینه‌سازی عملکرد مدل‌های AI می‌باشد. این تکنیک شامل سازگارسازی یک مدل پیش‌آموزش‌داده شده برای عملکرد بهتر در وظایف خاص یا در حوزه‌های خاص است، با استفاده از مجموعه‌های داده جدید، که اغلب سفارشی‌سازی شده‌اند، برای تنظیم وزن‌های آن مدل.

اما مهندسی دستور، مهارت یا نقش ویژه‌ای در AI است که بر روی هدایت و تنظیم پاسخ‌های مدل‌های یادگیری ماشین تمرکز دارد. مهندسان دستور پرسش‌ها یا دستورات دقیق و خاص را ایجاد می‌کنند تا پاسخ‌های موردنظر را از مدل AI بیرون کشانده و ایجاد کنند. این فرآیند نیازمند درک عمیق از معماری مدل و محدودیت‌های مجموعه‌های داده موجود می‌باشد.

تفاوت‌های کلیدی بین مهندسی پرومپ و تنظیم نهایی به شرح زیر می‌باشد:

1. **تمرکز:** مهندسی دستور به بهبود خروجی یا پاسخ‌های یک سامانه AI برای کاربران تمرکز دارد، در حالی که تنظیم نهایی بر روی بهبود عملکرد کلی مدل در وظایف خاص تمرکز دارد.

2. **رویکرد:** مهندسی دستور با ایجاد ورودی‌ها یا پرسش‌های موثرتر برای مدل AI بهبود خروجی‌ها را فراهم می‌کند. تنظیم نهایی با آموزش مدل بر روی داده‌های جدید به منظور افزایش دانش آن در حوزه‌های خاص به بهبود عملکرد می‌پردازد.

3. **کنترل:** مهندسی دستور کنترل دقیقی را بر روی عملکرد و پاسخ‌های سامانه AI ارائه می‌دهد تا با ایجاد پرومپ‌های سفارشی، پاسخ‌های مطلوب را استخراج کند. تنظیم نهایی عمق و جزئیات بیشتری به دانش مدل در حوزه‌های مشخص اضافه می‌کند.

4. **نیاز به منابع:** مهندسی دستور بر دستورهای ساخته‌شده توسط انسانها تکیه می‌کند و تقریباً به هیچ منبع محاسباتی نیاز ندارد. در مقابل، تنظیم نهایی اغلب نیاز به منابع محاسباتی قابل توجهی برای آموزش و داده‌های جدید دارد.

استفاده از هر دو تکنیک مهندسی دستور و تنظیم نهایی می‌تواند بهبود عملکرد مدل و بهبود خروجی‌ها را داشته باشد. انتخاب بین این دو تکنیک بستگی به اهداف و الزامات خاص یک پروژه و همچنین مهارت مهندسین انسانی دارد. ترکیب هر دو تکنیک می‌تواند به نتایج بهتری در سیستم‌های AI منجر شود.

## مدل های زبانی پرسش و پاسخ به کمک مدل های زبانی بزرگ:

پرسش و پاسخ با مدل‌های زبانی بزرگ (LLMs) یک وظیفه مهم در حوزه پردازش زبان طبیعی (NLP) و بازیابی اطلاعات (IR) است. در این وظیفه، سعی می‌شود به سوالاتی که به زبان طبیعی مطرح می‌شوند، پاسخ داده شود بر اساس اسناد بدون ساختار در مقیاس بزرگ [1]. این فرآیند به دو مرحله انجام می‌شود:

1. **بازیابی پاراگراف‌های مرتبط از اسناد مرتبط:** پاراگراف‌های مرتبط از اسناد مرتبط بازیابی می‌شوند.
2. **شناسایی دامنه پاسخ:** این مرحله که به عنوان درک خوانش ماشینی (MRC) شناخته می‌شود، شامل شناسایی دامنه پاسخ در پاراگراف‌های مرتبط بازیابی‌شده می‌شود.

وظیفه ماشینی درک مطالب به گونه‌ای است که ماشین توانایی تفسیر زبان طبیعی را داشته باشد و با خواندن یک متن، به سوالات پاسخ دهد. سیستم‌های OpenQA سنتی با استفاده از روش‌های جدید MRC عصری تکامل یافته‌اند تا پاسخ‌ها را از مستندات استخراج کنند.

آموزش مدل‌های عصبی برای وظایف پرسش و پاسخ نیازمند مجموعه‌داده‌های بزرگی است که اطلاعات ضرور برای این وظیفه را فراهم کنند. به همین دلیل، با افزایش تحقیقات OpenQA به ویژه در تکنیک‌های یکپارچه‌سازی با درک خوانش ماشینی عصبی، نیاز به تولید مجموعه‌داده‌ها افزایش یافته است.

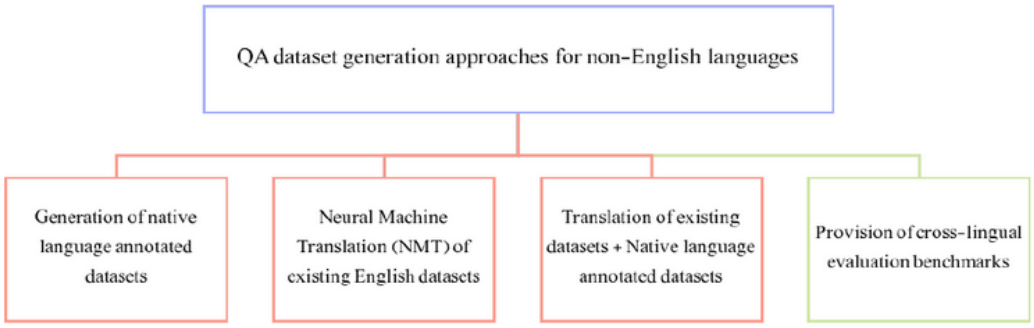
تا به امروز، مجموعه‌داده‌های مقیاس بزرگی برای پرسش و پاسخ در زبان‌های دیگر مانند فارسی تولید نشده است. اغلب مجموعه‌داده‌های ارائه‌شده به تازگی برای پرسش و پاسخ به زبان انگلیسی عرضه شده‌اند، از جمله مجموعه‌داده‌هایی مانند CNN/Daily Mail، MS MARCO، RACE و SQuAD.

میان این مجموعه‌داده‌ها، مجموعه‌داده Stanford Question Answering Dataset (SQuAD) به‌طور گسترده‌ای استفاده می‌شود. این مجموعه‌داده حاوی (متن متناظر، سوال، پاسخ) است و به مجموعه‌های آموزش و توسعه تقسیم شده است. SQuAD 2.0 سوالات غیرقابل پاسخ را اضافه کرد تا مدل‌ها را به چالش بکشد و آن‌ها را آموزش دهد که سوالاتی را که پاسخ قابل قبولی ندارند را به‌درستی اشاره کنند.

برای ایجاد یک مجموعه‌داده فارسی مبتنی بر SQuAD 2.0، نخستین گام ترجمه مجموعه‌داده‌های آموزش و توسعه SQuAD 2.0 به زبان فارسی با استفاده از رابط ماشینی ترجمه گوگل انجام شد، سپس سوالاتی که پاسخ ترجمه‌شده آن‌ها با بخشی از متن متناظر تطابق داشت، انتخاب و موارد دیگر حذف شدند. سپس با اصلاح و تنظیم بیشتر، مجموعه‌داده‌ای فارسی برای وظایف پرسش و پاسخ ایجاد شد که به آن ParSQuAD گفته می‌شود.

## نحوه ساخت دیتاست:

برای ساخت دیتاست به زبان فارسی دو روش در پیش است. یکی این است که از همان ابتدا پیش برویم و صفحات را crawl کنیم و از آن ها اطلاعات استخراج کنیم یا اینکه دیتاست آماده ای  به زبان انگلیسی مانند SQuAD را به زبان فارسی تبدیل کنیم. (تصویر 2)



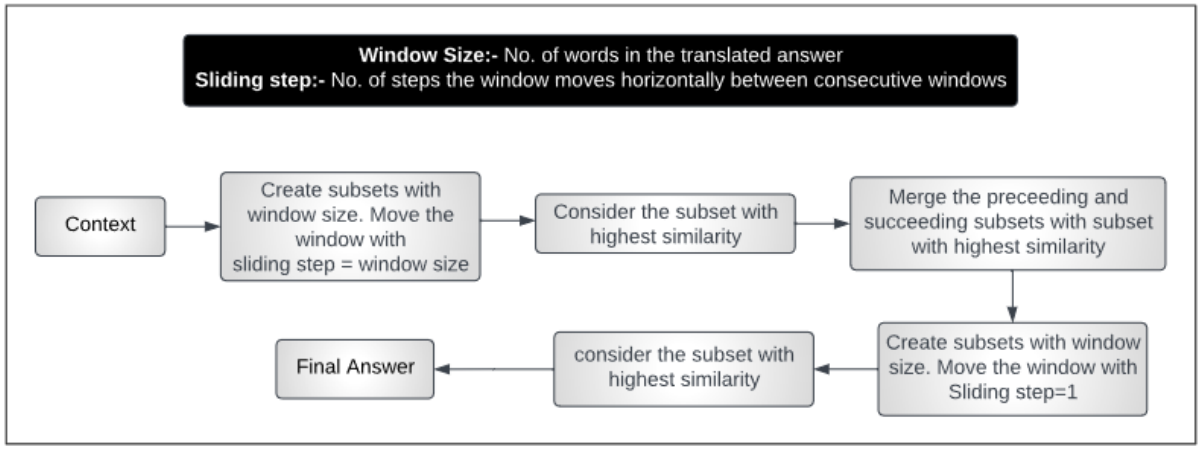

تصویر 2: نحوه ساخت دیتاست برای زبان های غیر انگلیسی

برای توسعه مدل پرسش و پاسخ، ضروری است که مکان شروع کاراکتری پاسخ در متن مشخص شود. اما به دلیل تغییرات در ساختار جملات در زبان‌های مختلف، نمی‌توان از اندیس‌های اصلی مجموعه‌داده به طور مستقیم استفاده کرد. برای حل این مشکل، یک روش نوآورانه برای تعیین اندیس‌های بهینه با دقت پایین طراحی کردیم. اطمینان از اینکه متن پاسخ دقیقاً با ظاهر آن در متن مطابقت داشته باشد چالشی است به دلیل ماهیت جعبه سیاه مدل‌های یادگیری ماشین که ممکن است به ترجمه‌های متنوعی از متن منجر شود.

در طول تولید مجموعه‌داده، ما با دو چالش اصلی روبرو شدیم:
1. تعیین اندیس شروع پاسخ مناسب در متن.
2. جایگزین کردن پاسخ ترجمه‌شده با پاسخ متناظر دقیق که در متن ظاهر می‌شود، به منظور دقت در معنی اصلی و همچنین تطابق دقیق.

شکل 3 فرآیند استخراج پاسخ نهایی از متن را توضیح می‌دهد. برای شناسایی اندیس‌های پاسخ در متن، از یک تکنیک پنجره لغزان استفاده کردیم. این رویکرد شامل استفاده از یک پنجره با طول برابر با اندازه متن پاسخ بود. با حرکت در سراسر متن با این اندازه پنجره، هدف ما یافتن زیرمجموعه‌ای از متن متناظر با بیشترین شباهت به متن پاسخ بود. این کار ما را قادر می‌سازد تا اندیس‌های دقیق متناظر با پاسخ در متن را مشخص کنیم. برای پرداختن به احتمال پاسخ نهایی که ممکن است در چندین

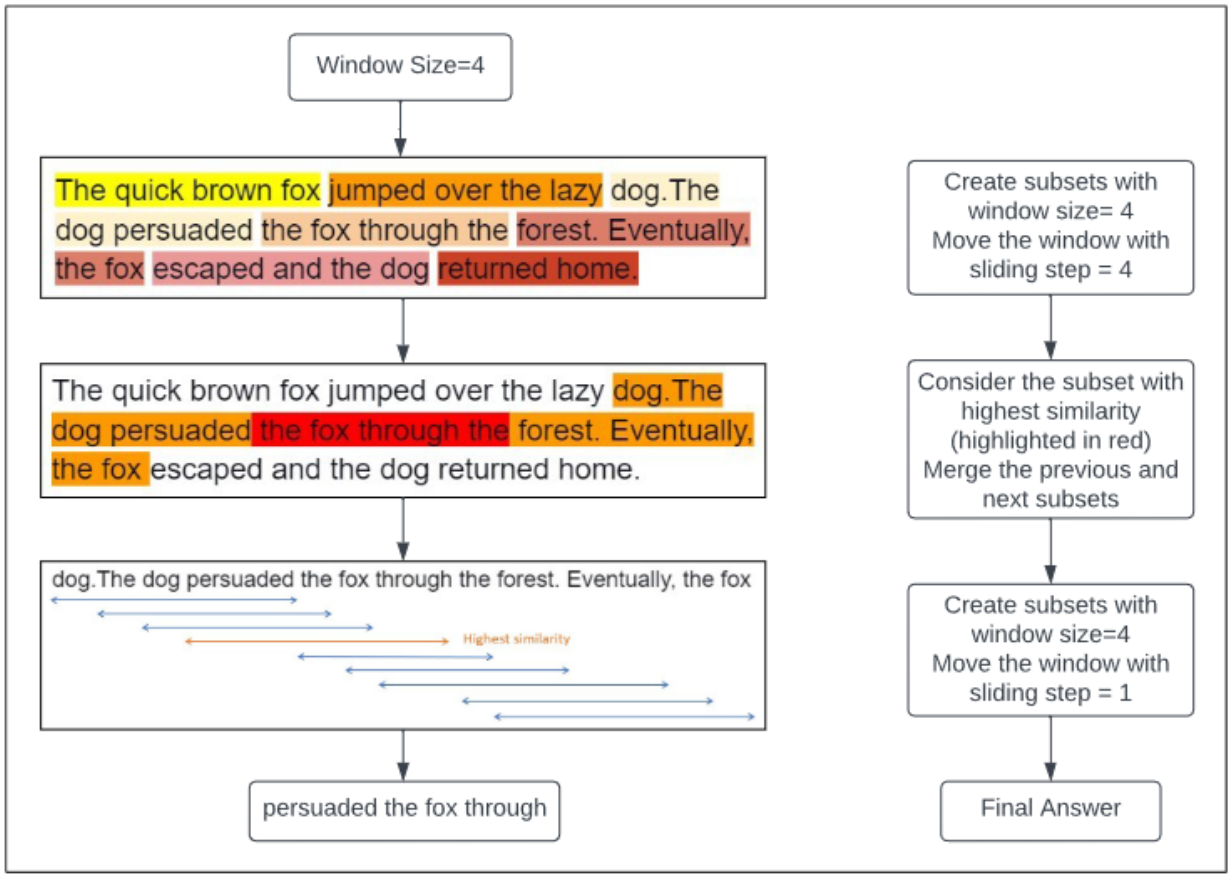
**زیرمجموعه قرار داشته باشد، از یک رویکرد ادغامی استفاده می‌کنیم. پس از شناسایی زیرمجموعه با بیشترین شباهت، آن را با زیرمجموعه‌های پیش‌رو و پس‌رو ادغام می‌کنیم. این ادغام به ما این امکان را می‌دهد که محدوده کاملی را که پاسخ ممکن است در آن قرار داشته باشد، مد نظر قرار دهیم. برای تعیین زیرمجموعه نهایی با بیشترین شباهت، ما با استفاده از یک پنجره با اندازه مشابه روش قبلی، با فاصله یکی به یک حرکت می‌کنیم. علاوه بر اندیس‌های کاراکتری، ما اندیس شروع و پایان توکن پاسخ را نیز محاسبه می‌کنیم که می‌تواند به اهداف گسترده‌تر در آموزش و توسعه مدل کمک کند. علاوه بر این، پس از به دست آوردن اندیس پاسخ در متن، متن پاسخ را با زیرمجموعه‌ای که بیشترین شباهت را نشان می‌دهد جایگزین می‌کنیم. این روش اطمینان از پوشش کامل پاسخ را از طریق فرآیند انتخاب مبتنی بر شباهت به ما می‌دهد.**




تصویر 3: تکنیک window size برای استخراج پاسخ

**برای توضیحات بیشتر، در زیر یک مثال به زبان انگلیسی آورده شده است:**

• **Context**: “The quick brown fox jumped over the lazy dog. The dog persuaded the fox through the forest. Eventually, the fox escaped and the dog returned home.”
• **Real Answer**: persuaded the fox through
• **Translated Answer**: chased the fox through
• **Window size**: length(Translated Answer) = 4




تصویر 4: مثالی به زبان انگلیسی از این متد

**شکل 4 توضیحی از تکنیک پنجره لغزان برای یافتن پاسخ نهایی در متن ارائه می‌دهد. این فرآیند برای کاهش منابع محاسباتی و زمان مورد نیاز برای یافتن اندیس در متن استفاده می‌شود. نیاز به این فرآیند به وجود می‌آید چرا که پاسخ در زمان ترجمه بازنویسی می‌شود و همیشه به همان شکل در متن نیست.**

## نحوه آموزش دادن و استفاده از مدل:

برای آموزش دادن مدل من Bert را پسندیدم و مدل خودم را بر بستر Bert تیون شده برای زبان فارسی موسوم به ParsBert انجام دادم. کاری که باید انجام می شد تیون کردن مدل با متن ها و سوالات فارسی بود که برای این امر آماده شود. من با کمک PyTorch و Transformer و سایت huggingFace مدل خودم را توسعه دادم.

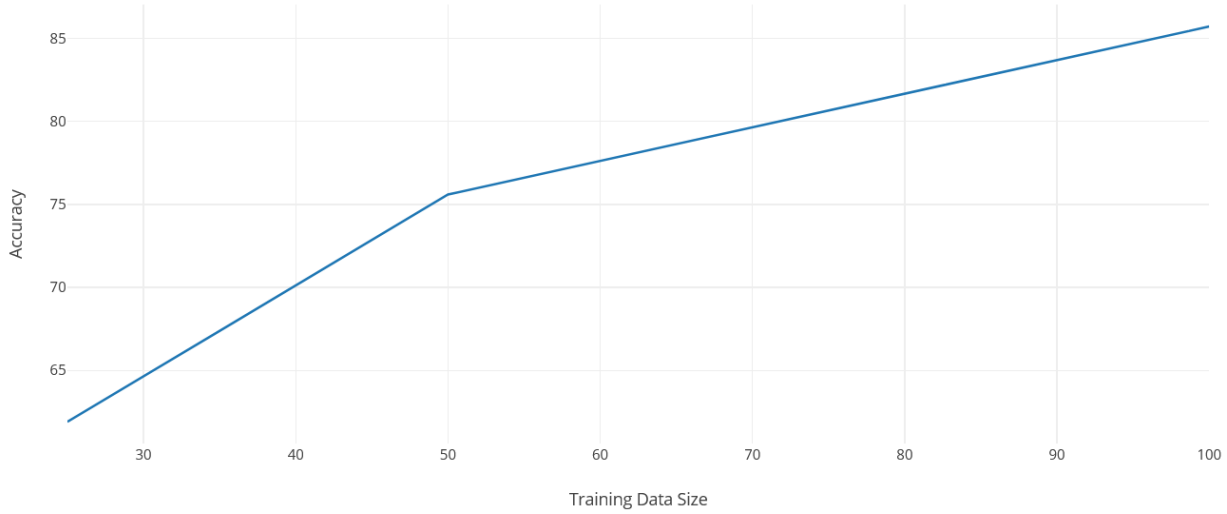
کد train کردن این مدل را از ریپازیتوری گیت هاب زیر می توانید مشاهده کنید.
گیت هاب

من برای این امر از دیتاست های فارسی مختلفی مانند Pquad و PersianQa استفاده کردم. رویه ساخت این دیتاست ها همانگونه هست که در قسمت های قبلی توضیح دادم به همین سبب به جای اینکه خودم دیتاست را از ابتدا تهیه کنم از این موارد آماده استفاده کردم.

من برای این امر چیزی که در مورد آن حساسیت داشتم این بود که با چه مقداری از دیتای train به چه دقتی می توانم برسم و میزان lossام به چه حدی می رسد. به همین خاطر train این مدل را چندین دفعه تکرار کردم و هر بار بخش بیشتری از دیتاست را در این train تاثیر دادم تا بتوانم نتیجه گیری مناسب تری داشته باشم.

در نمودار 1 من نشان داده ام که با دخیل کردم چند درصد از دیتاست خود به چه میزان دقتی از درصد رسیده ام.

برای استفاده کردن از این مدل می توانید از طریق سایت Hugging Face این مدل را دریافت و تست کنید.
لینک مدل




نمودار 1: تاثیر سایز دیتای train بر accuracy نهایی مدل

## نتیجه گیری:

**یکی از مهم ترین نتایجی که می توان از این تحقیق گرفت این است که نیاز نیست برای کار در حوزه مدل های زبانی بزرگ چرخ را از اول اختراع کرد و هم کار ها را خود کرد. کار های بسیاری انجام شده است که می توان با درک صحیح کار های انجام شده و استفاده از مدل های توسعه داده شده توسط ایشان چیزی را خلق کرد که ارزشی افزوده و ایده ای جدید داشته باشد.**

**در این کار عملی من با استفاده از دیتاست های فارسی توسعه داده شده و مدل bert فارسی از قبل توسعه داده شده و درک صحیح نحوه کاری این مدل موفق شدم آن را برای پاسخ دادن به سوالات و به صورت یک مدل پرسش و پاسخ تنظیم کنم. مدلی که با دیتاستی به طول بیش از 17 هزار نمونه و حدود 12 ساعت با بالاترین سیستم ها train شده است و اکنون می تواند کمکی شایان در راستای پیشبرد این علم برای زبان فارسی بکند.**

# منابع:


1. “.” 2022. [2308.16149] Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. https://arxiv.org/abs/2308.16149.
2. “[2202.06219] PQuAD: A Persian Question Answering Dataset.” 2022. arXiv. https://arxiv.org/abs/2202.06219.
3. “[2302.00321] An Evaluation of Persian-English Machine Translation Datasets with Transformers.” 2023. arXiv. https://arxiv.org/abs/2302.00321.
4 .“[2303.18223] A Survey of Large Language Models.” 2023. arXiv. https://arxiv.org/abs/2303.18223.
5. “[2308.09862].” 2022. Breaking Language Barriers: A Question Answering Dataset for Hindi and Marathi. https://arxiv.org/abs/2308.09862.
6. Ayoubi, Sajjad. n.d. “Persian (Farsi) Question Answering Dataset (+ Models).” GitHub. Accessed September 22, 20202123. https://github.com/SajjjadAyobi/PersianQA.

7. Farahani, Mohammad. 2020. “[2005.12515] ParsBERT: Transformer-based Model for Persian Language Understanding.” arXiv. https://arxiv.org/abs/2005.12515.
8. Mozafari, Jamshid. 2021. “(PDF) ParSQuAD: Persian Question Answering Dataset based on Machine Translation of SQuAD 2.0.” ResearchGate. https://www.researchgate.net/publication/356442081_ParSQuAD_Persian_Question_Answering_Dataset_based_on_Machine_Translation_of_SQuAD_20.
9. Tran, Ke. 2020. “[2002.07306] From English To Foreign Languages: Transferring Pre-trained Language Models.” arXiv. https://arxiv.org/abs/2002.07306.