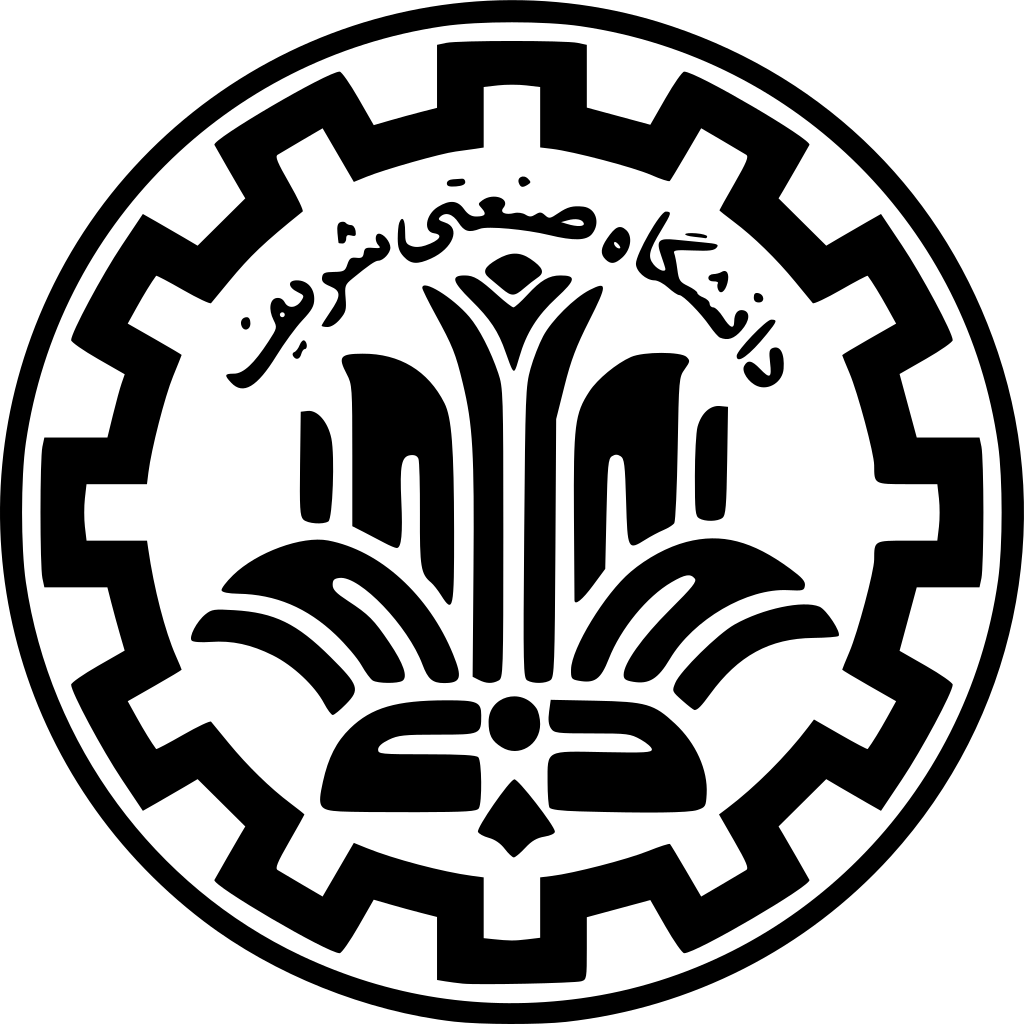
باسمه تعالی



دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

استخراج آراء قضایی و پیاده­سازی RAG حقوقی

بر روی آن با هدف پاسخ به پرسش[[1]](#footnote-2)

سید مجتبی ابطحی

402212203 Mojtaba.abtahi12@sharif.edu

امیرشکوری

402206437

Am.shakouri@sharif.edu

حامد جهانتیغ

401212324

Hamed.jahantigh96@sharif.edu

فاطمه رائیجیان

402203389

Fateme.raeejian22@sharif.edu

محمد مهدی رحیم سیرت

402206618 Mm.rahimsirat@ce.sharif.edu

زهرا ملکی

402206183 Zahramaleki@ce.sharif.edu

خلاصه

در این پژوهش تلاش شده است برپایه یک مدل زبانی، با کمک ایجاد یک مدل تولید تقویت شده با بازیابی[[2]](#footnote-3) بر روی پرونده­ها، پاسخ سوالات حقوقی افراد پیرامون پرونده­های موجود پاسخ داده شود. برای تحقق این امر در وهله اول داده­های لازم استخراج می­شود. در ادامه یک مدل بازیابی افزوده (RAG) بر روی آن پیاده می­شود. در نهایت با کمک مهندسی پرسش[[3]](#footnote-4) برپایه یک مدل بزرگ زبانی پاسخ مدنظر ارائه می­شود.

کلمات کلیدی: فناوری حقوقی، آراء قضایی (دادنامه)، مدل زبانی بزرگ (LLM)، تولید تقویت شده با بازیابی (RAG)، مهندسی پرسش (Prompt Engineering)

# مقدمه

در دوره کنونی با توسعه زیرساخت­های فناوری اطلاعات و گسترش آن در حوزه¬های دیگر و قوت گرفتن فضای میان رشته¬ای، فناوری حقوقی نیز به عنوان زمینه­ای مشترک میان فناوری و حقوق، با هدف ایجاد تحول در فضای سنتی قضایی و حقوقی شکل گرفته است. پیشرفت روزافزون فناوری هوش¬مصنوعی سبب شده تا بسیاری از فرآیندها و عملکردهای موجود از فضای انسان محور به سوی خودکار شدن توسط ماشین پیش رود. فناوری در عصر حاضر، می¬تواند به عنوان ابزاری قدرتمند در خدمت جوامع قضایی و حقوقی قرار بگیرد تا با استفاده صحیح و راهبردی از آن، بتوان جامعه را به سمت عدالت سوق دهد.

وکلا و حقوقدانان یک عضو مهم و تاثیرگذار در امور قضایی هستند که با کمک به آنان توسط ابزار فناوری می¬توان بسیاری از فرآیندهای قضایی را با دقت بالاتر از خطای انسانی تسریع کرد. در طول سال¬های گذشته فناوری حقوقی بیشتر در راستای همرسانی وکلا و مردم، احراز هویت، امضاء دیجیتال، هوشمندسازی قرارداد، تولید محتوای دیجیتال و ... فعالیت داشته است. در صورتی که ابزار سریع برای کمک به وکلا در امور وکالت با رویکرد صفر یا یک کلیک به صورت حداقلی و سطحی توسعه پیدا کرده است و جای کار در این راستا بسیار است.

یکی از موضوعات کمک کننده به جامعه وکالت و مردم در پیشبرد مشکلات و پرونده­های قضایی­شان، استفاده از پرونده­های گذشته، به ویژه آراء(دادنامه)های آن­هاست.

## تعریف مسئله

یکی از مسائل کلیدی در پیشبرد یک مسئله حقوقی و پرونده قاضی، آگاهی از نمونه­های گذشته به ویژه آراء پرونده­های پیشین می­باشد. در این راستا توسعه یک سیستم زبانی برای پاسخ به سوالات پیرامون مسائل حقوقی برپایه این دانش پیشین، می­تواند برخی از چالش­ها را تسهیل و بهبود دهد.

## تولید تقویت شده با بازیابی

RAG یا تولید تقویت شده با بازیابی، یک روش پیشرفته در پردازش زبان طبیعی است که به ترکیب قابلیت‌های بازیابی اطلاعات و تولید متنی می‌پردازد. در این روش، ابتدا از یک مدل بازیابی اطلاعات استفاده می‌شود تا اسناد یا داده‌های مرتبط با پرسش کاربر استخراج شود. سپس این اطلاعات بازیابی شده به عنوان ورودی به یک مدل تولید متنی مانند GPT داده می‌شود تا پاسخ نهایی تولید شود.

مزیت اصلی RAG این است که با ترکیب دانش و اطلاعات به‌روز از پایگاه‌های داده بزرگ، مدل‌های تولید متن می‌توانند پاسخ‌های دقیق‌تر و مبتنی بر اطلاعات واقعی ارائه دهند. این روش به ویژه در کاربردهایی مانند پاسخ به سؤالات پیچیده، تولید متون تخصصی یا پاسخ به سؤالات مرتبط با دانش به‌روز و دقیق بسیار مؤثر است.

## مهندسی پرسش

مهندسی پرسش (Prompt Engineering) یک حوزه نوظهور در پردازش زبان طبیعی (NLP) و هوش مصنوعی است که به بهینه‌سازی دستورات یا درخواست‌های متنی (پرامپت‌ها) برای تولید پاسخ‌های دقیق و کارآمد توسط مدل‌های زبان بزرگ مانند GPT می‌پردازد. هدف اصلی مهندسی پرامپت، طراحی و تنظیم متن‌هایی است که به طور مؤثر مدل‌های هوش مصنوعی را برای تولید نتایج مورد نظر راهنمایی می‌کنند.

این مهندسی شامل تکنیک‌های مختلفی است که بر اساس نوع مدل و هدف کاربردی مورد استفاده قرار می‌گیرد. به عنوان مثال، ممکن است پرامپت‌ها به شکلی خاص قالب‌بندی شوند تا مدل بتواند پاسخ‌های دقیق‌تری ارائه دهد. همچنین، مهندسی پرامپت ممکن است شامل تنظیماتی مانند تعیین طول پاسخ، استفاده از کلمات کلیدی خاص، یا حتی ترکیب چندین پرامپت برای هدایت مدل به سمت تولید خروجی‌های متنوع و کاربردی باشد.

# پیشینه پژوهش

فناوری حقوقی یا LegalTech، به کاربرد فناوری در حوزه حقوق و خدمات حقوقی اشاره دارد. پیشینه فعالیت در این حوزه به اوایل قرن بیستم و حتی قبل از آن بازمی‌گردد، اما توسعه و رشد چشمگیر آن عمدتاً در دو دهه گذشته رخ داده است.

در دهه ۱۹۷۰، شرکت‌های حقوقی بزرگ به تدریج شروع به استفاده از رایانه‌ها برای ذخیره و مدیریت اسناد کردند. این دوره را می‌توان به عنوان آغازین‌ترین مراحل استفاده از فناوری در حقوق دانست. در دهه ۱۹۹۰ و با گسترش اینترنت، ابزارهای بیشتری برای تحقیق حقوقی آنلاین و دسترسی به پایگاه‌های داده حقوقی توسعه یافتند که وکلا و مشاوران حقوقی را قادر ساخت تا سریع‌تر و با دقت بیشتری به اطلاعات حقوقی دسترسی پیدا کنند.

با ورود به قرن بیست و یکم و ظهور فناوری‌های پیشرفته‌تر مانند هوش مصنوعی، یادگیری ماشینی، و بلاکچین، حوزه فناوری حقوقی وارد مرحله‌ای جدید شد. ابزارهایی مانند نرم‌افزارهای مدیریت پرونده، سیستم‌های تحلیل داده‌های حقوقی، و پلتفرم‌های قراردادهای هوشمند به وکلا کمک کردند تا فرآیندهای حقوقی را بهینه‌سازی کرده و خدمات بهتری به مشتریان خود ارائه دهند. این فناوری‌ها نه تنها باعث افزایش کارایی و کاهش هزینه‌ها شدند، بلکه دسترسی به خدمات حقوقی را برای افراد بیشتری فراهم کردند.

در سال‌های اخیر، استفاده از پلتفرم‌های آنلاین برای مشاوره حقوقی، وکالت آنلاین، و حل اختلافات حقوقی نیز به شدت رواج یافته است. این روند نشان‌دهنده تحول بنیادین در نحوه ارائه خدمات حقوقی و حرکت به سوی دیجیتالیزه شدن این صنعت است.

در مجموع، فناوری حقوقی به یکی از مهم‌ترین و پرشتاب‌ترین زمینه‌های نوآوری در حوزه حقوق تبدیل شده و همچنان در حال گسترش و تحول است.

همچنین در تلاش­های گذشته دانشجویان این درس، با تمرکز بر روی قوانین فعالیت­هایی انجام شده است که برای دسترسی به آن می­توانید به لینک گیت­هاب <https://github.com/NLP-Final-Projects/IRI_LAW> مراجعه کرد.

# مجموعه داده

در حوزه­ی قضایی منابع، اسناد و مدارک از اهمیت ویژه­ای برخوردار هستند. می­توان منابع اصلی این حوزه را در دسته­های اصلی قوانین، مقررات، آراءقضایی (دادنامه­ها)، اساسنامه­ها، آراء وحدت رویه، نظریات مشورتی و طرح دعوی طبقه­بندی نمود. در این بخش توضیحات مربوط به نوع، نحوه استخراج، فرایند تمیزسازی و ویژگی­های داده­ها ارائه می­شود.

## نوع داده

با توجه به گستردگی زیاد منابع حقوقی، دو دسته اصلی برای حل مسئله تعریف شده از منابع انتخاب می­شوند:

1. قوانین اصلی: این قوانین تعریف حقوقی ندارند و بسته به میزان استفاده و کاربرد توسط حقوق­دانان به­کار گرفته می­شوند.
2. آراء قضایی (دادنامه­ها): آراء قضایی، که به آنها دادنامه نیز گفته می‌شود، به تصمیمات و احکام صادر شده توسط دادگاه‌ها و مقامات قضایی در مورد پرونده‌های حقوقی، کیفری یا اداری اشاره دارد. این آراء نتیجه بررسی و تحلیل دادگاه در مورد دعاوی مطرح شده است و بیانگر نظر نهایی دادگاه در خصوص موضوع مورد اختلاف می‌باشد.

## ویژگی­های داده

همان طور که توضیح داده شد، دو مجموعه داده توسط تیم توسعه یافته است که هرکدام دارای معماری و ساختار ویژه خود هستند:

1. **قوانین اصلی:**

قوانین به عنوان مجموعه‌ای از قواعد و مقررات که توسط مراجع قانون‌گذاری تصویب می‌شوند، دارای ساختار و اجزایی مشخص هستند. این ساختار و اجزاء به گونه‌ای تنظیم می‌شوند که قانون به صورت جامع، شفاف و قابل فهم باشد. در ادامه ساختار و اجزاء اصلی یک قانون توضیح داده شده است:

* *عنوان قانون:* به طور مختصر و دقیق موضوع و ماهیت قانون را مشخص می‌کند. عنوان باید به گونه‌ای انتخاب شود که محتوای قانون را به‌خوبی بیان کند. مثال: قانون مدنی، قانون مجازات اسلامی
* *مقدمه یا دیباچه:* مقدمه شامل توضیحاتی در مورد ضرورت و اهداف تصویب قانون است. این بخش ممکن است دلایل و اهداف پشت تصویب قانون را بیان کند و زمینه‌های تصویب آن را توضیح دهد. مثال: "به منظور تنظیم روابط میان شهروندان و حمایت از حقوق فردی و اجتماعی..."
* *مواد قانون:* اصلی‌ترین بخش قانون را تشکیل می‌دهند و هر ماده یک قاعده یا مقرره مشخص را تعیین می‌کند. مواد قانون به ترتیب شماره‌گذاری شده و می‌توانند دارای بندها و تبصره‌های مختلفی باشند.
* *فصل‌ها و بخش‌ها:* قوانین بزرگتر ممکن است به فصل‌ها و بخش‌های مختلف تقسیم شوند تا سازماندهی و دسترسی به مواد قانونی آسان‌تر شود. هر فصل یا بخش معمولاً موضوعات مرتبط را در بر می‌گیرد. مثال: فصل اول: تعاریف و کلیات، فصل دوم: حقوق و تکالیف، ...
* *پیوست‌ها:* در برخی قوانین، ممکن است پیوست‌هایی مانند جداول، نمودارها، فرم‌ها یا مستندات دیگر ضمیمه شود که برای توضیح بیشتر یا کاربرد عملی قانون لازم است.
* *تعاریف و اصطلاحات:* در این بخش از قانون، تعاریف دقیق و مشخصی از اصطلاحات و واژه‌های کلیدی ارائه می‌شود که در قانون به‌کار رفته‌اند. این تعاریف به منظور جلوگیری از سوءتعبیر و اختلاف نظر در تفسیر قانون است. مثال: "منظور از "شخص" در این قانون هر فرد حقیقی یا حقوقی است..."
* *موارد اجرایی و نحوه اجرا:* این بخش نحوه اجرای قانون، مسئولیت‌ها و اختیارات مراجع مختلف را مشخص می‌کند. همچنین ممکن است مقرراتی در مورد نظارت بر اجرای قانون و نحوه رسیدگی به تخلفات از قانون نیز در این بخش آمده باشد. مثال: "مسئول اجرای این قانون وزارت امور اقتصادی و دارایی است..."
* *تاریخ اجرا و نسخ:* در این بخش تاریخ دقیق اجرای قانون و همچنین قوانین یا مقررات قبلی که با تصویب این قانون نسخ می‌شوند، ذکر می‌شود. مثال: "این قانون از تاریخ 1 مهر 1403 لازم‌الاجرا است و کلیه قوانین و مقررات مغایر با آن نسخ می‌شوند."
* *امضاء و تصویب:* در انتهای قانون، اسامی و امضاهای مراجع تصویب‌کننده قانون (مانند نمایندگان مجلس، رئیس‌جمهور و...) درج می‌شود که نشان‌دهنده تأیید نهایی قانون است.

با توجه به مسئله تعریف شده و نیازمندی، در این پژوهش برای هر قانون نام، مواد قانون و مواد مکرر آورده شده است و بقیه اطلاعات در قوانین فیلتر شده­اند.





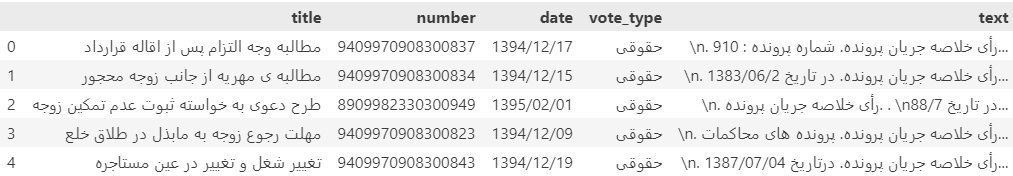
شكل ‏3‏.‏‌1 نمونه­ای از داده­های قوانین

1. **آراء قضایی:**

اجزاء اصلی یک رأی قضایی یا دادنامه معمولاً شامل موارد زیر است:

* *مقدمه:* شامل اطلاعات پایه‌ای مانند شماره پرونده، نام طرفین، و مرجع قضایی صادرکننده رأی.
* *موضوع دعوا:* شرح مختصری از موضوع اختلاف بین طرفین.
* *مبانی قانونی:* استناد به قوانین، مقررات و اصول حقوقی که مبنای تصمیم‌گیری دادگاه بوده‌اند.
* *استدلالات و دلایل دادگاه:* تحلیل و بررسی دادگاه از موضوع و دلایل انتخاب یک راه‌حل خاص.
* *نتیجه‌گیری و حکم:* تصمیم نهایی دادگاه که ممکن است شامل محکومیت، تبرئه، جبران خسارت، یا هر نوع دستور دیگری باشد.
* *امضاء*: رأی قضایی باید به امضای قاضی یا قضات صادرکننده آن برسد.

متناسب با نیاز تعریف شده، مجموعه داده مورد نیاز شامل شماره دادنامه، تاریخ، نوع پرونده، عنوان پرونده و رای پرونده آماده شده است.



شكل ‏3‏.‏‌2 نمونه­ای از داده­های آراء قضایی

## نحوه استخراج داده

برای هر دسته از داده­ها، نحوه استخراج متفاوت است. بدین ترتیب فرآیند استخراج برای هر کدام به ترتیب زیر شرح داده می­شود:

1. **قوانین اصلی:**

برای استخراج قوانین، منابع عمومی مختلفی از جمله سامانه ملی قوانین[[4]](#footnote-5)، سایت مهدی داوودآبادی[[5]](#footnote-6)، سایت پژوهش­های مجلس[[6]](#footnote-7) و غیره وجود دارد. در این پژوهش قوانین از سایت مهدی داوود آبادی استخراج شده­اند. در ابتدا 20 عنوان اصلی قوانین ارائه شده به عنوان قوانین اصلی در سایت انخاب می­شوند. این قوانین شامل *قانون اساسی جمهوری اسلامی ایران، قانون مالیاتهای مستقیم، قانون کار، قانون شوراهای حل اختلاف، قانون مدنی، قانون مسئولیت مدنی، قانون بیمه اجباری خسارات وارد شده به شخص ثالث در اثر حوادث ناشی از وسایل نقلیه، قانون تجارت، قانون دیوان عدالت اداری، قانون ثبت اسناد و املاک، تعزیرات و مجازاتهای بازدارنده، قانون آیین دادرسی کیفری، قانون مجازات اسلامی، قانون صدور چک، قانون حمایت خانواده، لایحه قانونی اصلاح قسمتی از قانون تجارت، قانون آیین دادرسی مدنی، انون نحوه اجرای محکومیت‌های مالی، قانون شهرداری، قانون اجرای احکام مدنی* می­باشد.

در ادامه متون این قوانین استخراج شده و با استفاده از کد model/law\_provider.py مجموعه داده اصلی شامل دو فایل dataset/law\_name.csv و dataset/madeh\_df.csv آماده می­شود. در فایل اول اطلاعات قوانین و در فایل دوم مواد به تفیکیک قرار گرفته ­اند. در مجموع 5،882 ماده و اصل استخراج شده است. برای عدم تداخل میان مواد مکرر در همان شماره ماده اصلی قرار گرفته­اند.

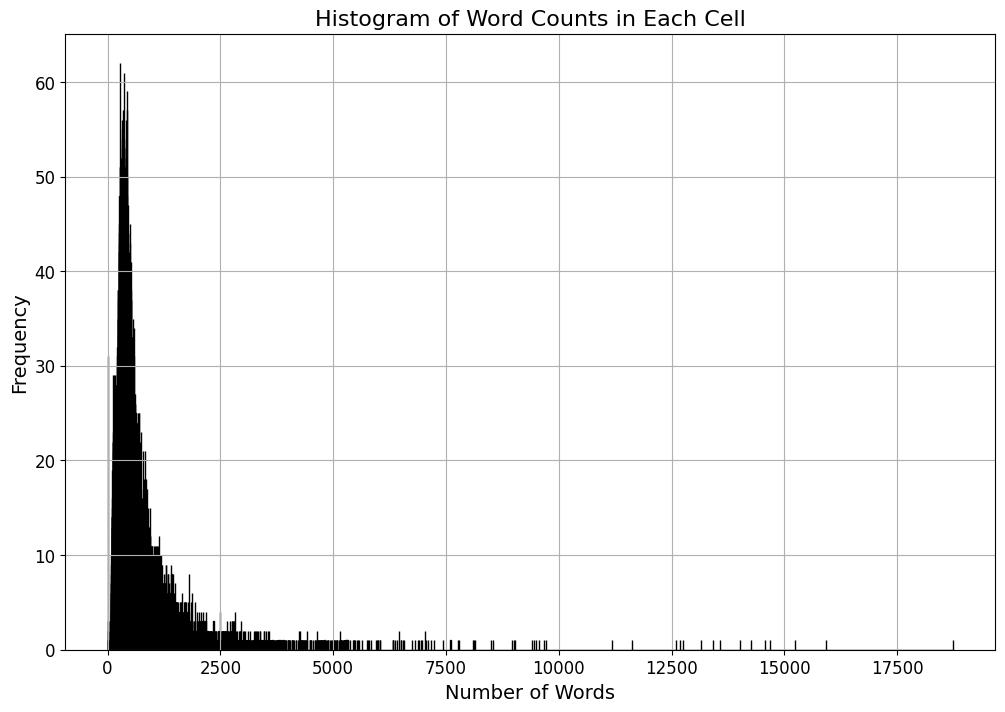
1. **آراء قضایی:**

در حال حاضر در حدود 30،000 دادنامه توسط پژوهشگاه قوه­قضاییه در سامانه ملی آراء[[7]](#footnote-8) به صورت گمنام قرار گرفته است. این آراء اساس این پژوهش قرار می­گیرد. برای استخراج داده­ها از این سایت، یک خزنده[[8]](#footnote-9) با نام model/case\_crawler.py نوشته شده است. این خزنده ابتدا اراء قضایی را از سامانه ملی آراء استخراج می­کند و متناسب با نیازمندی تعریف شده، داده­ها تمیز می­شود.

پس از اجرای این خزنده، بیش از 25،000 رای پس از مدت 7 ساعت و 30 دقیقه استخراج و در dataset/case\_df.csv آورده شده است.

## فرآیند تمیزسازی داده

پس از استخراج آراء، لازم است تا برخی پیش پردازش­ها بر رو داده­ها انجام شود. برای این امر ابتدا چند بررسی روی 25،048 داده انجام می­شود. از این تعداد تنها 622 داده دارای تعداد کلمات بیش از 2500 و 40 داده کمتر از 50 کلمه دارند. در شکل زیر توزیع تعداد کلمات اسناد در یک نمودار هیستوگرام نشان داده شده است.

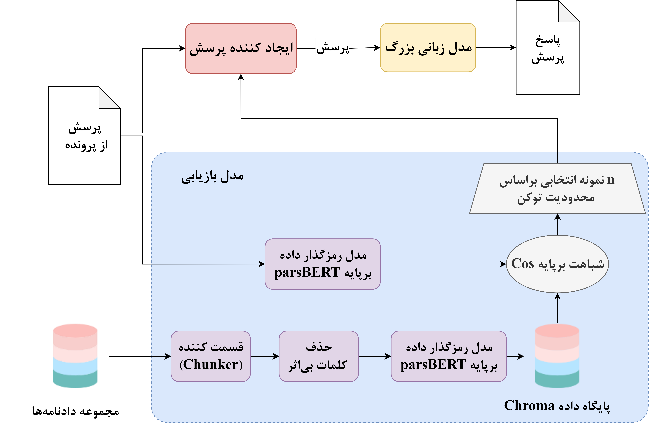


شكل ‏3‏.‏‌3 هیستوگرام تعداد کلمات هر پرونده

با توجه به فرآیند نهفته­سازی[[9]](#footnote-10) و محدودیت­های نمایش محتوا در بردار و همچنین محدودیت­های پرسش[[10]](#footnote-11) در مدل زبانی بزرگ، پیش پردازش­هایی بر روی داده­های پرونده­ها انجام می­شود. این پیش­پردازش­ها شامل حذف اسامی گمنام شده، تاریخ­ها، شماره­های بی اثر و ... می­باشد که دقیق­تر آن در فایل model/pre\_process.ipynb قرار گرفته است.

# روش تحقیق و پیاده­سازی

برای تحقق اهداف تعریف شده، از دو حوزه جدید در پردازش زبان طبیعی 1- مهندسی پرسش و 2- تولید تقویت شده با بازیابی در استفاده از مدل­های بزرگ زبانی استفاده شده است که در بخش ‏1 معرفی شدند.



شكل ‏4‏.‏‌1 مدل ارائه شده RAG بر روی دادنامه­های قضایی

همان طور که در شکل 1.4 نشان­داده شده است، مدل از بخش­های مختلفی تشکیل شده است که هر یک به تفضیل ارائه می­شود:

1. **مجموعه داده دادنامه­ها:**

این مجموعه داده شامل بیش از 25 هزار دادنامه­های استخراجی می­باشد که در بخش ‏3 به تفضیل به آن پرداخته شده است.

1. **بخش بازیابی اطلاعات:**

این بخش از دو قسمت اصلی تشکیل می­شود.

*در بخش اول* به صورت منفعل برپایه مدل parsBERT[[11]](#footnote-12) مجموعه داده­های پرونده­ها کدگذاری می­شوند. با توجه به ظرفیت محدود مدل­­ها هر پرونده ابتدا کلمات اضافی از پرونده ها حذف می­شود و با محدودیت 250 کلمه جدا می­شود. برای هر کدام از بخش­ها[[12]](#footnote-13) یک اشتراک 50 کلمه­های از بخش قبل و بعد در نظر گرفته شده است تا بخشی از پیوستگی حفظ شود. در نهایت کدگذاری های بخش­های هر پرونده بایکدیگر میانگین گرفته می­شود و کدگذاری کل پرونده را تشکیل میدهد. برای دسترسی بهتر و استفاده در زمان پاسخگویی به پرسش­ها، بردارهای رمزگذاری شده پرونده ها در پایگاه­داده Chroma ذخیره می­شود.

*در بخش دوم* در زمانی که یک پرسش مطرح می­شود، به صورت فعال با کمک مدل زبانی parsBERT کدگذاری می­شود و براساس اطلاعات ذخیره شده در پایگاه داده Chroma از دادنامه­ها، بیشترین امتیاز به صورت شباهت کسینوس محاسبه می­شود. متن دادنامه­ها براساس محدودیت توکن تا حداکثر 5 سند بازیابی شده استخراج می­شود و به بخش بعدی ارجاع داده می­شود.

1. **ایجاد کننده پرسش:**

با استفاده از دادنامه­های بازیابی شده در بخش قبل و پرسش مطرح شده، این بخش پرسش (Prompt) مناسب برای مدل زبانی بعد را ایجاد می­کند. در حال حاضر این پرسش از متون دادنامه­ها، عنوان دادنامه­ها و پرسش مطرح شده استفاده می­کند.

1. **مدل زبانی:**

برای دریافت پاسخ پرسش براساس دانش استخراج شده از RAG، از فراخانی یک مدل زبانی در بستر API Call استفاده شده است. مدل انتخابی در این بخش از سری مدل­های ارائه شده توسط شرکت OpenAI مدل 4o-mini می­باشد که با توجه به شرایط تحریمی، برای دسترسی به API از سایت gilas.io استفاده شد.

# ارزیابی و اعتبارسنجی

برای ارزیابی نتایج، در ابتدا به صورت انسانی اقدام به تشکیل یک مجموعه داده از پرسش­ و پاسخ­ بر روی دادنامه­های استخراج شده می­شود. برای بهبود پژوهش این مجموعه ارزیابی شامل سه دسته به شرح زیر می­شود.

1. آسان: سوال ها عموما از عنوان­ها دریافت شده و پاسخ در متن موجود می­باشد.
2. متوسط: سوالات و پاسخ ها هر دو از متن دریافت شده است.
3. سخت: سوالات از مجموعه ایی از پرونده های مشابه و جواب ها نیز متن آن استخراج شده است.

برای ارزیابی از دو معیار Bleu و Bert Score استفاده می­شود که هر یک در ابتدا به شرح زیر معرفی می­شوند:

*BLEU Score (Bilingual Evaluation Understudy)* یک معیار خودکار برای ارزیابی کیفیت ترجمه‌های ماشینی است. این معیار با مقایسه n-gramهای تولید شده توسط مدل با n-gramهای مرجع (ترجمه‌های انسانی) کار می‌کند. BLEU نمره‌ای بین 0 تا 1 ارائه می‌دهد که هرچه به 1 نزدیک‌تر باشد، کیفیت ترجمه بهتر است. این معیار به دلیل سادگی و سرعت محاسبه، به‌طور گسترده‌ای در ارزیابی سیستم‌های ترجمه استفاده می‌شود، اما ممکن است در شناسایی معانی عمیق‌تر یا ساختارهای پیچیده ناتوان باشد.

*BERT Score* یک معیار جدیدتر است که از مدل‌های زبان پیشرفته مانند BERT برای ارزیابی کیفیت متن استفاده می‌کند. این معیار به جای مقایسه n-gramها، به بررسی شباهت‌های معنایی بین متن تولید شده و متن مرجع می‌پردازد. BERT Score با استفاده از embeddingهای تولید شده توسط BERT، شباهت‌های کلمه به کلمه را محاسبه می‌کند و نمره‌ای بین 0 تا 1 ارائه می‌دهد. این رویکرد به دلیل توجه به معانی و زمینه‌های کلمات، معمولاً دقت بیشتری نسبت به BLEU Score دارد.

به‌طور خلاصه، BLEU Score بیشتر بر روی دقت ساختاری تمرکز دارد، در حالی که BERT Score به درک معنایی متن‌ها می‌پردازد.

با توجه به آنکه تمرکز پروژه بر روی RAG می­باشد، برای آزمون از مقایسه جواب­های تولید شده در دو دسته با استفاده از RAG و بدون استفاده از آن مقایسه با پاسخ های ارزیابی انجام می­شود. در نهایت با اجرای معیارهای ارزیابی، دقت­های استخراجی به شرح زیر بدست می­آید.

جدول ‏5‏.‏1 معیارهای ارزیابی Bleu

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **مدل پاسخ** | **Bleu-4** | **Bleu-3** | **Bleu-2** | **Bleu-1** |
| با RAG | 10.6% | 14.8% | 20.4% | 32.5% |
| بدون RAG | 6.1% | 8.9% | 13.3% | 21.2% |

جدول ‏5‏.‏2 معیارهای ارزیابی BERT

|  |  |  |  |
| --- | --- | --- | --- |
| **مدل پاسخ** | **F1** | **Recall** | **Precision** |
| با RAG | 91.5% | 92.3% | 91.5% |
| بدون RAG | 90.4% | 91.0% | 90.4% |

همان طور که مشاهده می­شود نتایج در BLEU بهبود چشمگیری را نشان میدهد.

# جمع­بندی

در این پژوهش با ارائه یک مدل برپایه تولید تقویت شده با بازیابی این امکان فراهم آمد تا یک مدل زبانی بزرگ با کمک دانش افزوده از آن بتواند به سوالات حقوقی مردم متناسب با دادنامه­ها و نمونه­های مشابه پاسخ دهد و بهبود تقریبا 70 درصدی در معیار BLEU ایجاد نماید. همچنین برای استفاده از مدل توسعه داده شده می­توان از نشانی <https://huggingface.co/spaces/mojtabaa4/Legal_RAG_Chatbot> استفاده کرد.

1. این پژوهش در راستای پروژه نهایی درس پردازش زبان طبیعی نیمسال دوم 03-1402 جناب دکتر حسام­الدین عسگری در دانشگاه صنعتی شریف انجام شده است. فایل­های پژوهش در نشانی <https://github.com/HamedJahantigh-git/legal_case_rag> قرار گرفته است. [↑](#footnote-ref-2)
2. Retrieval-Augmented Generation [↑](#footnote-ref-3)
3. Prompt Engineering [↑](#footnote-ref-4)
4. <https://qavanin.ir> [↑](#footnote-ref-5)
5. <https://davoudabadi.ir> [↑](#footnote-ref-6)
6. <https://rc.majlis.ir/fa/law> [↑](#footnote-ref-7)
7. <https://ara.jri.ac.ir> [↑](#footnote-ref-8)
8. Crawler [↑](#footnote-ref-9)
9. Embedding [↑](#footnote-ref-10)
10. Prompt [↑](#footnote-ref-11)
11. <https://huggingface.co/HooshvareLab/bert-base-parsbert-uncased> [↑](#footnote-ref-12)
12. Chunk [↑](#footnote-ref-13)