باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

استخراج آراء قضایی و پیاده‌سازی RAG حقوقی بر روی آن با هدف پاسخ به پرسش[1]

حامد جهانتیغ
۴۰۱۲۱۲۳۲۴
Hamed.jahantigh96@sharif.edu

سید مجتبی ابطحی
۴۰۲۲۱۲۲۰۳
Mojtaba.abtahi12@sharif.edu

امیرشکوری
۴۰۲۲۰۶۴۳۷
Am.shakouri@sharif.edu

زهرا ملکی
۴۰۲۲۰۶۱۸۳
Zahramaleki@ce.sharif.edu

محمد مهدی رحیم سیرت
۴۰۲۲۰۶۶۱۸
Mm.rahimsirat@ce.sharif.edu

فاطمه رائیجیان
۴۰۲۲۰۳۳۸۹
Fateme.raeejian22@sharif.edu

خلاصه

در این پژوهش تلاش شده است برپایه یک مدل زبانی، با کمک ایجاد یک مدل تولید تقویت شده با بازیابی[2] 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] 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)

۱ 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

[1] این پژوهش در راستای پروژه نهایی درس پردازش زبان طبیعی نیمسال دوم ۰۳-۱۴۰۲ 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.

[2] Retrieval-Augmented Generation

[3] Prompt Engineering

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.

۱.۱ 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.

۲.۱ 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.

۳.۱ 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.

۲ 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.

۳ 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.

۱.۳ نوع داده

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

۱- قوانین اصلی: این قوانین تعریف حقوقی ندارند و بسته به میزان استفاده و کاربرد توسط حقوق‌دانان به‌کار گرفته می‌شوند.

۲- 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.

۲.۳ 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:

۱- 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 ۱ مهر ۱۴۰۳ 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.

	law_index	law_name
0	0	قانون مالیاتهای مستقیم
1	1	قانون کار
2	2	قانون شوراهای حل اختلاف
3	3	قانون مدنی
4	4	قانون مسئولیت مدنی

	law_index	madeh_index	madeh_text
0	0	1	...:اشخاص زیر مشمول پرداخت مالیات میباشندn\ماده 1
1	0	2	...اشخاص زیر مشمول پرداخت مالیات‌های موضn\ماده 2
2	0	3	...به موجب ماده (2) قانون اصلاح موادی ازn\ماده 3
3	0	4	... به موجب ماده (2) قانون اصلاح موادی ازn\ماده 4
4	0	5	...به موجب ماده (2) قانون اصلاح موادی ازn\ماده 5

شکل ۱.۳ نمونه‌ای از داده‌های قوانین

۲- 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.

	text	vote_type	date	number	title
0	...رأی خلاصه جریان پرونده. شماره پرونده : 910 .\n\	حقوقی	1394/12/17	9409970908300837	مطالبه وجه التزام پس از افاله قرارداد
1	...رأی خلاصه جریان پرونده. در تاریخ 1383/06/2 .\n\	حقوقی	1394/12/15	9409970908300834	مطالبه ی مهریه از جانب زوجه محجور
2	...در تاریخ \n88/7\ . .رأی خلاصه جریان پرونده .\n\	حقوقی	1395/02/01	8909982330300949	طرح دعوی به خواسته ثبوت عدم تمکین زوجه
3	...رأی خلاصه جریان پرونده. پرونده های محاکمات .\n\	حقوقی	1394/12/09	9409970908300823	مهلت رجوع زوجه به مابذل در طلاق خلع
4	...رأی خلاصه جریان پرونده. درتاریخ 1387/07/04 .\n\	حقوقی	1394/12/19	9409970908300843	تغییر شغل و تغییر در عین مستاجره

شکل ۲.۳ نمونه‌ای از داده‌های آراء قضایی

۳.۳ 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:

۱- قوانین اصلی:

برای استخراج قوانین، منابع عمومی مختلفی از جمله سامانه ملی قوانین[۴]، سایت مهدی داوودآبادی[۵]، سایت پژوهش‌های مجلس[۶] و غیره وجود دارد. در این پژوهش قوانین از سایت مهدی داوود آبادی استخراج شده‌اند. در ابتدا ۲۰ عنوان اصلی قوانین ارائه شده به عنوان قوانین اصلی در سایت انتخاب 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 تفکیک قرار گرفته اند. در مجموع ۵،۸۸۲ ماده و اصل استخراج شده است. برای عدم تداخل میان مواد مکرر در همان شماره ماده اصلی قرار گرفته‌اند.

۲- آراء قضایی:

در حال حاضر در حدود ۳۰،۰۰۰ دادنامه توسط پژوهشگاه قوه‌قضاییه در سامانه ملی آراء[۷] به صورت گمنام قرار گرفته است. این آراء اساس این پژوهش قرار می‌گیرد. برای استخراج داده‌ها از این سایت، یک خزنده[۸] 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.

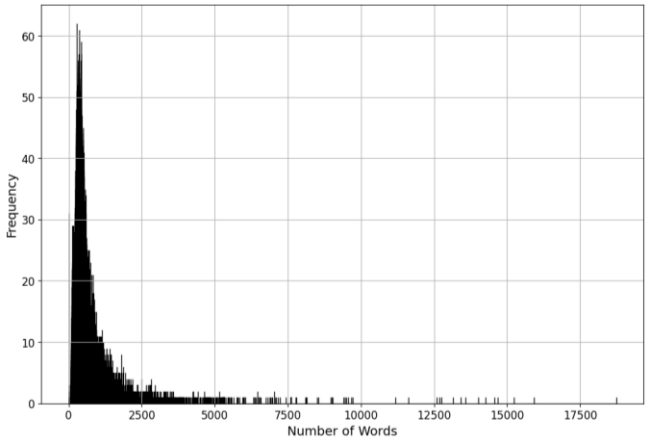
پس از اجرای این خزنده، بیش از ۲۵،۰۰۰ رای پس از مدت ۷ ساعت و ۳۰ دقیقه استخراج و در dataset/case_df.csv آورده شده است.

۴.۳ 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 ۲۵،۰۴۸ داده انجام

[4] https://qavanin.ir
[5] https://davoudabadi.ir
[6] https://rc.majlis.ir/fa/law
[7] https://ara.jri.ac.ir
[8] Crawler

می‌شود. از این تعداد تنها ۶۲۲ داده دارای تعداد کلمات بیش از ۲۵۰۰ و ۴۰ داده کمتر از ۵۰ کلمه دارند. در شکل زیر توزیع تعداد کلمات اسناد در یک نمودار هیستوگرام نشان داده شده است.

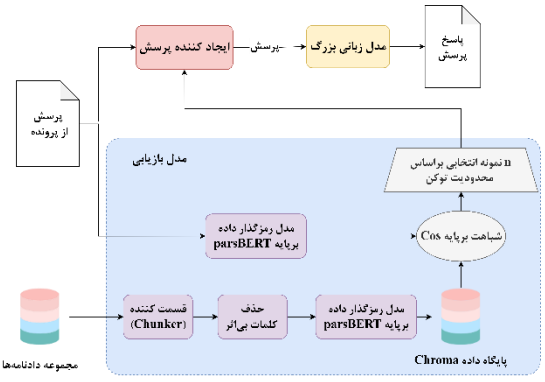


شکل ۳.۳ هیستوگرام تعداد کلمات هر پرونده

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

۴ روش تحقیق و پیاده‌سازی

برای تحقق اهداف تعریف شده، از دو حوزه جدید در پردازش زبان طبیعی ۱- مهندسی پرسش و ۲- تولید تقویت شده با بازیابی در استفاده از مدل‌های بزرگ زبانی استفاده شده است که در بخش ۱ معرفی شدند.



شکل ۱.۴ مدل ارائه شده RAG بر روی دادنامه‌های قضایی

همان طور که در شکل ۱.۴ نشان‌داده شده است، مدل از بخش‌های مختلفی تشکیل شده است که هر یک به تفضیل ارائه می‌شود:

۱- مجموعه داده دادنامه‌ها:

این مجموعه داده شامل بیش از ۲۵ هزار دادنامه‌های استخراجی می-باشد که در بخش ۳ به تفضیل به آن پرداخته شده است.

۲- بخش بازیابی اطلاعات:

این بخش از دو قسمت اصلی تشکیل می‌شود.

در بخش اول به صورت منفعل برپایه مدل parsBERT[11] مجموعه داده‌های پرونده‌ها کدگذاری می‌شوند. با توجه به ظرفیت محدود مدل‌ها هر پرونده ابتدا کلمات اضافی از پرونده ها حذف می‌شود و با محدودیت ۲۵۰ کلمه جدا می‌شود. برای هر کدام از بخش‌ها[12] یک اشتراک ۵۰ کلمه‌ای 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 ۵ سند بازیابی شده استخراج می‌شود و به بخش بعدی ارجاع داده می‌شود.

۳- 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.

۴- 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

[9] Embedding
[10] Prompt
[11] https://huggingface.co/HooshvareLab/bert-base-parsbert-uncased
[12] Chunk

OpenAI مدل 4o-mini می‌باشد که با توجه به شرایط تحریمی، برای دسترسی به API از سایت gilas.io استفاده شد.

۵ 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.

۱- آسان: سوال ها عموما از عنوان‌ها دریافت شده و پاسخ در متن موجود می‌باشد.

۲- متوسط: سوالات و پاسخ ها هر دو از متن دریافت شده است.

۳- 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 ۰ تا ۱ ارائه می‌دهد که هرچه به ۱ 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 ۰ تا ۱ 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 ۱.۵ معیارهای ارزیابی Bleu

مدل پاسخ	Bleu-4	Bleu-3	Bleu-2	Bleu-1
با RAG	۱۰.۶٪	۱۴.۸٪	۲۰.۴٪	۳۲.۵٪
بدون RAG	۶.۱٪	۸.۹٪	۱۳.۳٪	۲۱.۲٪

جدول ۲.۵ معیارهای ارزیابی BERT

مدل پاسخ	F1	Recall	Precision
با RAG	۹۱.۵٪	۹۲.۳٪	۹۱.۵٪
بدون RAG	۹۰.۴٪	۹۱.۰٪	۹۰.۴٪

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

۶ 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 ۷۰ درصدی در معیار BLEU ایجاد نماید. همچنین برای استفاده از مدل توسعه داده شده می‌توان از نشانی https://huggingface.co/spaces/parsi-ai-nlpclass/Legal_RAG استفاده کرد.