دانشگاه اصفهان
دانشکده مهندسی کامپیوتر
گروه هوش مصنوعی

توسعه‌یک مدل زبانی پزشکی مبتنی بر استدلال درزبان فارسی

مهرداد قصابی

استاد راهنما

دکتر حمیدرضا برادران

بهمن ۱۴۰۴

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه هوش مصنوعی

هیأت داوران پروژهٔ کارشناسی آقای / خانم مهرداد قصابی به شمارهٔ دانشجویی ۴۰۲۳۶۱۴۰۲۹ در رشتهٔ هوش مصنوعی را در تاریخ ..... با عنوان «توسعه‌یک مدل زبانی پزشکی مبتنی بر استدلال درزبان فارسی» و نمره نهایی زیر ارزیابی کرد.

	به عدد	به حروف
با نمرهٔ نهایی		

نام و نام خانوادگی استاد راهنما:

تاریخ و امضا:

نام و نام خانوادگی استاد داور:

تاریخ و امضا:

تقدیم به:

پدرم که در طول تحصیل پشتیبانم بوده است

سپاس گزاری

سپاس و آفرین خداوندگار جان آفرین راست ، اوی که آدمی را به گوهر خرد آراست.

در آغاز دستان پدر و مادر نازنینم را به پاس مهر بیکرانشان به گرمی میفشارم، و از استاد راهنما خود جناب آقای دکتر حمیدرضا برادران بابت راهنمایی هایشان در طول انجام این پایان نامه سپاس گزاری میکنم.

مهرداد قصابی

بهمن ۱۴۰۴

چکیده

استفاده از هوش مصنوعی در پاسخگویی به سوالات پزشکی به عنوان یکی از حوزه‌های نوظهور و مهم در فناوری و بهداشت شناخته می‌شود که در سال‌های اخیر مورد توجه گسترده‌ای قرار گرفته است. این فناوری پیشرفته، با قابلیت‌های ویژه خود، می‌تواند کیفیت خدمات پزشکی ارائه‌شده به بیماران را به شکل چشمگیری ارتقا دهد. همچنین، با سرعت بخشیدن به فرآیند ارائه اطلاعات پزشکی و ارائه پاسخ‌های سریع و دقیق به سوالات پزشکان و بیماران، نقش مهمی در کاهش فشار کاری پزشکان ایفا می‌کند. به این ترتیب، هوش مصنوعی نه تنها موجب افزایش کارایی در سیستم‌های بهداشتی می‌شود، بلکه تجربه کلی بیماران را بهبود می‌بخشد و زمینه ارائه درمان‌های بهتر و مؤثرتر را فراهم می‌کند.

از طرف دیگر از آنجا که پزشکی مبتنی بر استدلال و تحلیل‌های منطقی است، توسعه یک مدل پزشکی که بر پایه زنجیره‌ای از افکار و استدلال‌های منطقی طراحی شده باشد، می‌تواند دقت و کارایی این مدل را به طور قابل توجهی افزایش دهد. چنین رویکردی امکان انجام فرآیندهای پیچیده تشخیصی و درمانی را به صورت ساختاریافته‌تر و هدفمندتر فراهم می‌کند. در این زمینه، هر مرحله از تشخیص و درمان باید مبتنی بر شواهد علمی و داده‌های معتبر باشد. به عنوان مثال، پزشکان در فرآیند تشخیص بیماری‌ها معمولاً از تاریخچه پزشکی، علائم بالینی و نتایج آزمایش‌ها بهره می‌گیرند. با طراحی یک مدل منطقی، این داده‌ها می‌توانند در قالب یک زنجیره منطقی به یکدیگر متصل شوند که به شناسایی الگوها و روابط میان علائم و بیماری‌ها کمک می‌کند.

واژگان کلیدی هوش مصنوعی در پزشکی، مدل های زبانی فارسی، مدل های زبانی پزشکی، پردازش زبان های طبیعی، زنجیره افکار

فهرست مطالب

فهرست تصاویر

فهرست جداول

فهرست الگوریتم‌ها

فهرست برنامه‌ها

فصل ۱

دیباچه

۱.۱ هدف پژوهش

هدف از این پژوهش، توسعه یک مدل زبانی پزشکی فارسی بر پایه استدلال [۱] است که قابلیت اجرا روی دستگاه‌های محلی را داشته باشد. اجرا روی دستگاه‌های محلی از آن جهت حائز اهمیت است که داده‌های پزشکی اغلب حساس و خصوصی هستند و ارسال آنها به سرورهای خارجی ممکن است خطرات جدی برای حریم خصوصی بیماران ایجاد کند.

۲.۱ کاربرد پژوهش

۱.۲.۱ کاربرد مدل های زبانی پزشکی

با معرفی معماری نوآورانه ترنسفورمر در مقاله Attention is All You Need [۱] تحولی بنیادین در حوزه پردازش زبان طبیعی [۲] ایجاد شد. این معماری زمینه‌ساز توسعه مدل‌های زبانی پیشرفته‌ای شده است که با استفاده از مکانیزم‌های توجه، توانایی درک و تولید زبان انسانی را با دقتی شگفت‌انگیز به دست آورده‌اند.

این پیشرفت‌ها منجر به افزایش چشمگیر کاربرد هوش مصنوعی [۳] در حوزه‌های مختلف، به‌ویژه در زمینه

[۱] reasoning
[۲] natural language processing
[۳] artificial intelligence

پزشکی، شده است.در حوزه پزشکی، مدل‌های زبانی مبتنی بر هوش مصنوعی نقش مهمی در تحلیل داده‌های پزشکی، بهبود دقت تشخیص بیماری‌ها، ارائه پیشنهادهای درمانی دقیق‌تر و افزایش کیفیت مراقبت از بیماران ایفا می‌کنند. علاوه بر این، این فناوری، به بهینه‌سازی سیستم‌های اداری و کاهش بار کاری کادر درمانی کمک شایانی کرده است. به عنوان مثال، مدل‌های هوش مصنوعی قادرند با تحلیل داده‌های حاصل از پرونده‌های پزشکی، الگوهای مرتبط با بیماری‌ها را شناسایی کنند و اطلاعات ارزشمندی را برای تصمیم‌گیری سریع‌تر و دقیق‌تر در اختیار پزشکان قرار دهند. این مدل‌ها همچنین می‌توانند نقش مهمی در تکمیل مشاوره‌های پزشکی ایفا کرده و به پزشکان در ارائه اطلاعات دقیق‌تر و سریع‌تر کمک کنند.و حتی شاید در آینده ای نه چندان دور بتوانند جای پزشکان را در مشاوره های پزشکی بگیرند.

این تحول نه تنها به افزایش کارایی و بهره‌وری در سیستم‌های درمانی منجر شده است، بلکه تجربه کلی بیماران را نیز بهبود بخشیده و امکان ارائه خدمات درمانی بهتر و مؤثرتر را فراهم کرده است. به همین دلیل، توسعه و استفاده از مدل های زبانی پزشکی [۴] ، همچنان مورد توجه پژوهشگران و متخصصان قرار دارد.

۲.۲.۱ کاربرد مدل های زبانی پزشکی فارسی

علیرغم پیشرفت‌های چشمگیر در توسعه مدل‌های زبانی پزشکی به زبان انگلیسی، در حوزه زبان فارسی هنوز کار چندانی صورت نگرفته است. این در حالی است که در سرتاسر جهان میلیون ها نفر تنها قادر به استفاده از این زبان هستند؛ بنابراین تلاش برای توسعه یک مدل زبانی پزشکی در زبان فارسی میتواند گامی رو به جلو در ارتباطات و خدمات درمانی باشد.

۳.۱ مراحل انجام پایان نامه

همانطور که در جدول ۱.۱ این پایان‌نامه در دو فاز اصلی طراحی و اجرا شده است. فاز نخست به جمع‌آوری دادگان پزشکی فارسی و توسعه مدلی با نام گائوکرنا-V اختصاص دارد که فاقد توانایی استدلال بوده و بیشتر بر درک سیستم یک [۵] زبان تمرکز دارد. از این فاز، مقاله‌ای با عنوان "اهرم قرار دادن داده‌های آنلاین برای بهبود دانش پزشکی یک مدل زبانی کوچک پزشکی فارسی" استخراج شده است که به تشریح فرآیند جمع‌آوری داده‌ها و نحوه بهینه‌سازی دانش پزشکی مدل می‌پردازد. در فاز دوم این پژوهش ابتدا تکنیک های جدیدی برای ارتقای توانایی

[۴] medical language models

[۵] در علم رفتارشناسی به درک سریع، شهودی و بدون نیاز به تفکر ژرف درک سیستم یک و به درک آهسته، غیر شهودی و نیازمند استدلال درک سیستم دو میگویند.

استدلال و درک سیستم دو مدل معرفی شده و سپس مدل گائوکرنا-R در این فاز توسعه داده میشود. از این فاز نیز، مقاله‌ای با عنوان "؟" استخراج شده است.

	gaokerena-V	**gaokerena-R**
مخزن گیت هاب	mehrdadghassabi/gaokerena-V	mehrdadghassabi/gaokerena-R
مخزن پارامتر ها	gaokerena/gaokerena-v1.0	gaokerena/gaokerena-r1.0
پیوند مقاله	https://arxiv.org/pdf/2505.16000	https://arxiv.org/pdf/0000.00000
هزینه	۳۰۰ دلار	۳۰۰ دلار
همکاران	دکتر حمیدرضا برادران، پدرام رستمی، میلاد توکلی، امیرحسین پورسینا و زهرا کاظمی	دکتر حمیدرضا برادران، پدرام رستمی و صدرا حکیم

جدول ۱.۱: اطلاعات دو فاز پایان نامه

۴.۱ ساختار پایان نامه

در این پایان‌نامه، ساختار فصل‌ها به گونه‌ای طراحی شده است که مراحل مختلف پژوهش به صورت منظم و هدفمند ارائه شوند. فصل دوم به بررسی کارهای پیشین اختصاص دارد که در آن مطالعات انجام‌شده در زمینه‌های مرتبط مرور خواهند شد. در فصل سوم، به دلیل عدم وجود دادگان پزشکی خاص در حوزه زبان فارسی، فرآیند جمع‌آوری و آماده‌سازی این دادگان به‌طور دقیق تشریح خواهد شد. سپس در فصل چهارم، با استفاده از دادگان معرفی‌شده در فصل سوم، مدل اولیه با نام گائوکرنا-V [۶] معرفی و تحلیل می‌شود. در فصل پنجم، توانایی‌های استدلال در مدل‌های هوش مصنوعی مورد بررسی قرار گرفته و چالش‌ها و راهکارهای مرتبط با این موضوع ارائه خواهند شد. در ادامه، در فصل ششم، با معرفی تکنیک هایی برای بهبود توانایی استدلال یک مدل زبانی مدل پیشرفته‌تری به نام گائوکرنا-R معرفی و ویژگی‌های آن به‌تفصیل شرح داده می‌شود. در نهایت، فصل پایانی به جمع‌بندی نتایج پژوهش و پیشنهاداتی برای تحقیقات آینده اختصاص دارد.

[۶] نام گائوکرنا از درختی افسانه‌ای الهام گرفته شده است که در روایات اساطیری زرتشتی به عنوان نماد شفادهی و جاودانگی شناخته می‌شود.

فصل ۲

بررسی کار های پیشین

۱.۲ مقدمه

همان‌طور که پیش‌تر اشاره شد، علیرغم پیشرفت‌های چشمگیر در توسعه مدل‌های زبانی پزشکی به زبان انگلیسی، مانند توسعه و معرفی مدل‌های MedPalm [۲] [۳] ،متأسفانه در حوزه زبان فارسی هنوز کار چندانی در این زمینه انجام نشده است. این مسئله بدین معناست که ما در حوزه زبان فارسی تقریباً با یک کاغذ سفید روبه‌رو هستیم. در این پایان‌نامه تلاش شده است تا قدمی رو به جلو در جهت توسعه مدل‌های زبانی پزشکی برای زبان فارسی برداشته شود.

در ادامه، به بررسی کارهای پیشین انجام‌شده، چه در حوزه زبان فارسی و چه در حوزه زبان انگلیسی، خواهیم پرداخت.

۲.۲ کار های پیشین در حوزه زبان انگلیسی

۱.۲.۲ مدل های Med-Palm

مدل های med-palm یکی از مدل‌های زبانی پزشکی بزرگ [۱] است که توسط تیم تحقیقاتی گوگل برای کاربرد های پزشکی توسعه داده شده است. این مدل با استفاده از داده‌های تخصصی پزشکی و بالینی آموزش

[۱] large medical language models

دیده است. هدف اصلی این خانواده از مدل های زبانی پزشکی پاسخ‌گویی به پرسش‌های پزشکی با دقت بالا، کمک به پزشکان در تصمیم‌گیری‌های بالینی، و تسهیل دسترسی به اطلاعات پزشکی برای کاربران است. نسخه‌های مختلف این مدل، مانند MedPaLM و MedPaLM2 ، توانایی‌های قابل توجهی در درک و تحلیل زبان تخصصی پزشکی نشان داده‌اند و به عنوان یک ابزار نوین در حوزه هوش مصنوعی پزشکی شناخته می‌شوند. این مدل‌ها با استفاده از آزمون‌های استاندارد پزشکی (مانند USMLE) ارزیابی شده و توانسته‌اند عملکردی نزدیک به سطح متخصصین پزشکی ارائه دهند. مدل MedPaLM2 به عنوان یک گام مهم در جهت توسعه مدل‌های زبان تخصصی در حوزه سلامت و پزشکی شناخته می‌شود.

۲.۲.۲ مدل ChatDoctor

مدل ChatDoctor [۴] یکی از برجسته‌ترین تلاش‌ها در حوزه توسعه مدل‌های زبانی پزشکی است که شباهت قابل توجهی به فاز نخست پایان‌نامه حاضر دارد. تیم توسعه‌دهنده این مدل، داده‌های آموزشی خود را از دو پلتفرم آنلاین پرسش و پاسخ پزشکی به نام‌های HealthcareMagic و iCliniq جمع‌آوری کرده‌اند. این تیم ابتدا بیش از دویست هزار جفت پرسش و پاسخ پزشکی از این منابع گردآوری کرده و سپس با اعمال فیلترهایی بر اساس طول و کیفیت پاسخ‌ها، مجموعه‌ای با کیفیت بالا شامل صد هزار جفت پرسش و پاسخ نهایی ایجاد کرده‌اند. داده‌های مذکور به‌عنوان پایه‌ای برای آموزش و تنظیم دقیق [۲] مدل LLaMa [۵] مورد استفاده قرار گرفته‌اند تا مدلی توانمند در تولید اطلاعات پزشکی دقیق و مرتبط ایجاد شود.

علاوه بر این، این مدل از رویکرد تولید مبتنی بر بازیابی اطلاعات [۳] بهره برده است. این رویکرد به مدل امکان می‌دهد تا به اطلاعات جدید و خارجی دسترسی پیدا کرده و آن‌ها را به‌طور مؤثر در پاسخ‌های خود ادغام کند. چنین رویکردی موجب ارتقای عملکرد کلی سیستم شده و توانایی مدل در تولید پاسخ‌هایی دقیق‌تر و مرتبط‌تر را به‌طور چشمگیری بهبود بخشیده است.

۳.۲.۲ مدل های Meerkat

مدل‌های Meerkat [۶] یکی دیگر از تلاش‌های برجسته در حوزه توسعه مدل‌های زبانی پزشکی است. این پروژه با استخراج زنجیره‌های تفکر [۴] از کتاب‌های درسی پزشکی و تنظیم دقیق یک مدل زبانی پایه با استفاده از این داده‌ها، همراه با مجموعه داده‌های مکمل دیگر، به وجود است.همانند فاز دوم پایان نامه حاضر هدف

[۲] fine-tuning
[۳] Retrieval-Augmented Generation (RAG)
[۴] chain of thought

اصلی Meerkat تمرکز بر فرآیندهای استدلالی است که در تصمیم‌گیری‌های پزشکی نقش دارند. این مدل تلاش کرده است تا نه تنها اطلاعات پزشکی دقیق ارائه دهد، بلکه فرآیندهای شناختی و تصمیم‌گیری متخصصان حوزه سلامت را شبیه‌سازی کند. به همین دلیل، Meerkat به عنوان مدلی برای تعاملات پیچیده‌تر و آگاهانه‌تر در حوزه پزشکی معرفی شده است.

۴.۲.۲ مدل MedMobile

MedMobile [۷] تلاشی دیگر در حوزه مدل‌های زبانی کوچک پزشکی است. برای توسعه این مدل زبانی کوچک، مدل Phi-3-mini [۸] به عنوان مدل پایه [۵] استفاده از ترکیبی از داده‌های مصنوعی و تولیدشده توسط انسان تنظیم دقیق شده است تا عملکردی بهینه و مناسب برای اجرا روی دستگاه‌های همراه مانند موبایل ارائه دهد. با تمرکز بر نیازهای خاص کاربران دستگاه های همراه، MedMobile تلاش کرده است مدلی کارآمد و مؤثر فراهم کند که دسترسی به اطلاعات پزشکی باکیفیت را در هر زمان و مکان به صورت محلی [۶] ممکن می‌سازد.

۳.۲ کار های پیشین در حوزه زبان فارسی

همان‌طور که پیش‌تر اشاره شد، تحقیقات محدودی بر روی مدل‌های زبانی پزشکی فارسی تمرکز داشته‌اند که این امر نشان‌دهنده شکاف قابل توجهی در منابع موجود برای جامعه پزشکی فارسی‌زبان است. علاوه بر این، پژوهش‌های بسیار اندک موجود در این زمینه ، به طور کامل در مورد مجموعه داده‌ها، مدل‌ها و کدهای خود متن بسته [۷] هستند.

از سوی دیگر، تمامی این تلاش‌ها عمدتا بر روی راهکارهای استخراجی [۸] متمرکز بوده‌اند که هدفشان بازیابی اطلاعات مرتبط از منابع از پیش تعریف شده است، به جای استفاده از رویکردهای تولیدی [۹] که قادر به تولید پاسخ‌های آگاه از زمینه باشند.

[۵] baseline model
[۶] local
[۷] closed-source
[۸] extractive
[۹] generative

۱.۳.۲ مدل Sina-bert

شاید اولین و برجسته ترین مدل زبانی پزشکی فارسی، Sina-BERT [۹] باشد که شامل آموزش یک مدل BERT [۱۰] با استفاده از یک پیکره خزش شده [۱۰] همراه با مجموعه‌داده پرسش و پاسخ پزشکی فارسی است که به طور خاص برای کاربردهای مختلف از جمله پاسخ به سوالات پزشکی، تحلیل احساسات پزشکی و بازیابی سوالات پزشکی توسعه یافته‌اند.

Sina-BERT در میان تلاش‌های متمرکز بر زبان فارسی، بیشترین شباهت را به فاز نخست پایان نامه حاضر دارد؛ با این تفاوت که از مدل برت [۱۱] یک مدل زبانی مبتنی بر رمزگذار [۱۲] به عنوان مدل پایه استفاده می‌کند. این انتخاب تولید پاسخ توسط این مدل را عملا ناممکن می سازد، چرا که برت عمدتا برای درک و استخراج اطلاعات طراحی شده است نه برای تولید پاسخ.

۲.۳.۲ سیستم پرسش و پاسخ پزشکی دکتر ویسی و همکاران

یکی از آثار برجسته در حوزه پردازش زبان طبیعی، سیستم پرسش و پاسخ پزشکی فارسی است که توسط دکتر ویسی و همکارانش [۱۱] طراحی و توسعه داده شده است. این سیستم به‌طور کلی شامل سه ماژول اصلی است: پردازش پرسش، بازیابی سند و استخراج پاسخ. ماژول پردازش پرسش وظیفه تحلیل و اصلاح پرسش‌های کاربران را برعهده دارد تا پرسش‌ها به شکل بهینه برای مراحل بعدی آماده شوند. سپس، ماژول بازیابی سند با استفاده از الگوریتم‌های پیشرفته، اسناد پزشکی مرتبط را از میان داده‌های از پیش تعیین‌شده پیدا می‌کند. در نهایت، ماژول استخراج پاسخ با شناسایی دقیق اطلاعات موجود در اسناد بازیابی‌شده، مناسب‌ترین پاسخ‌ها را استخراج کرده و به کاربران ارائه می‌دهد. این سیستم نه تنها به‌طور مؤثر به پرسش‌های پزشکی پاسخ می‌دهد، بلکه ساختار ماژولار آن امکان بهبود و توسعه در آینده را نیز فراهم می‌سازد.

۳.۳.۲ پایان نامه کارشناسی ارشد خانم لیلا دارابی

مشابه به این دو اثر، پیشین لیلا دارابی در پایان نامه ارشد خود [۱۲] از مدل‌هایی مانند Pars-BERT [۱۳] برای بازیابی پاسخ‌های مرتبط استفاده کرده است. رویکرد او شامل یافتن سوالات مشابه برای مدیریت پرسش‌های تکراری و به کارگیری استراتژی‌های ارزیابی دقیق و سهل‌گیرانه برای پاسخ‌های دقیق یا تقریبی می‌شود. علاوه بر

[۱۰] crawled
[۱۱] BERT
[۱۲] encoder-based

این، روش‌های طبقه‌بندی و شناسایی موجودیت‌های نامدار [۱۳] برای بهبود ارتباط پاسخ‌ها از طریق دسته‌بندی سوالات و شناسایی موجودیت‌های پزشکی مانند نام داروها و بیماری‌ها به کار گرفته می‌شوند.

[۱۳] Named Entity Recognition (NER)

فصل ۳

جمع آوری دادگان

۱.۳ مقدمه

همان‌طور که پیشتر اشاره شد، در حوزه زبان فارسی نه مدل‌های عمومی موجود هستند و نه مجموعه داده‌های مناسب برای استفاده در پژوهش‌های مرتبط. بنابراین، برای پیشبرد این پایان‌نامه، ناچار به جمع‌آوری دادگان اختصاصی بودیم تا بتوانیم نیازهای تحقیقاتی را برآورده کنیم. فرآیند جمع‌آوری دادگان شامل روش‌هایی مانند ترجمه [۱] داده‌های موجود از زبان‌های دیگر و خزش داده‌ها از منابع مختلف برای ایجاد یک مجموعه داده جامع و کاربردی بوده است.

۲.۳ معرفی پیکره پزشکی فارسی

عدم وجود یک پیکره پزشکی اختصاصی به زبان فارسی، چالشی قابل توجه برای پژوهشگران و توسعه‌دهندگانی ایجاد می‌کند که هدفشان توسعه مدل های پزشکی در زبان فارسی است. بدون داده‌های متنی باکیفیت و تخصصی که برای آموزش مدل‌های هوش مصنوعی ضروری است، این تلاش‌ها ممکن است با موانع روبه‌رو شوند و در نهایت بر توسعه فناوری‌ها و راه‌حل‌های پیشرفته پزشکی مناسب برای جمعیت فارسی‌زبان تاثیر بگذارند. برای حل این مشکل، ما یک مجموعه داده جامع شامل تقریبا نود میلیون توکن و حدود صد هزار مقاله گردآوری کرده‌ایم. [۲]

[۱] ترجمه می‌تواند به صورت ماشینی یا انسانی انجام شود.

[۲] برای بازدید از این پیکره میتوانید به آدرس huggingface.co/datasets/gaokerena/medical_corpus مراجعه کنید

گارسیا فررو و همکاران [۱۴] مجموعه‌ای از متون پزشکی را که به چهار زبان (انگلیسی، فرانسوی، اسپانیایی و ایتالیایی) اختصاص داشت، گردآوری کردند که می‌توان آن را همانطور که در جدول ۱.۳ نشان داده شده است با مجموعه ما مقایسه کرد.پیکره ای که ما گردآوری کرده ایم از مجله های آنلاین پزشکی خزش شده است که میتوانید سهم هر مجله در این پیکره را در تصویر ۱.۳ ببینید.

زبان	تعداد پرسش و پاسخ ها	گردآورنده
انگلیسی	1.1B	I. Garcia Ferrero et al.
اسپانیایی	950M	I. Garcia Ferrero et al.
فرانسوی	675M	I. Garcia Ferrero et al.
ایتالیایی	143M	I. Garcia Ferrero et al.
فارسی	90M	ما

جدول ۱.۳: مقایسه پیکره گردآوری شده با پیکره های گردآوری شده توسط I. Garcia Ferrero et al.

۳.۳ معرفی مجموعه داده MF3QA

گردآوری یک مجموعه داده واقعی از پرسش وپاسخ‌های پزشک و بیمار اهمیت بسیاری در ارتقا توانایی‌های مدل‌های زبانی در حوزه بهداشت و درمان دارد. چنین مجموعه داده‌ای به مدل‌ها امکان می‌دهد تا اطلاعات ارزشمندی را که از تعاملات واقعی میان ارائه‌دهندگان خدمات بهداشتی و بیماران به دست می‌آید، بیاموزند. با تحلیل این تعاملات واقعی، مدل‌های زبانی می‌توانند به درک جزئیات اصطلاحات پزشکی، نگرانی‌های بیماران، و زمینه پیرامون سؤالات بهداشتی دست یابند. علاوه بر این، این مجموعه داده مدل‌ها را قادر می‌سازد نه تنها محتوای دقیق پاسخ‌ها، بلکه ساختار و لحن مناسب برای پاسخ‌دهی به سوالات را نیز یاد بگیرند. این فرآیند دوگانه یادگیری از اهمیت بالایی برخوردار است، زیرا به مدل امکان می‌دهد پاسخ‌هایی دقیق، همدلانه و متناسب با زمینه ارائه دهد و در نهایت ارتباط و پشتیبانی از بیماران در محیط‌های پزشکی را بهبود بخشد.

در این زمینه، یانگ لیو در مقاله مروری [۳] خود [۲۱] به چندین مجموعه داده واقعی پرسش وپاسخ پزشک وبیمار اشاره کرده است، مقایسه‌ای میان این مجموعه دادگان و مجموعه داده ما در جدول ۲.۳ ارائه شده است.

[۳] survey

نام مجموعه داده	زبان	تعداد پرسش و پاسخ ها	گردآورنده
ChatDoctor	انگلیسی	100K	[۴] Yunxiang Li et al.
CMtMedQA	چینی	68K	[۱۵] Songhua Yang et al.
DISC-Med-SFT	چینی	465K	[۱۶] Zhijie Bao et al.
HuatuoGPT-sft-data-v1	چینی	226K	[۱۷] Hongbo Zhang et al.
Huatuo-26M	چینی	26M	[۱۸] Jianquan Li et al.
MedDialog	چینی و انگلیسی	3.66M	[۱۹]Guangtao Zeng et al.
Medical-Meadow	انگلیسی	160k	[۲۰] Tianyu Han et al.
MF3QA	فارسی	20k	ما

جدول ۲.۳: مقایسه مجموعه داده های پرسش و پاسخ آزاد پزشکی با مجموعه داده گردآوری شده

۱.۳.۳ منابع مجموعه داده MF3QA

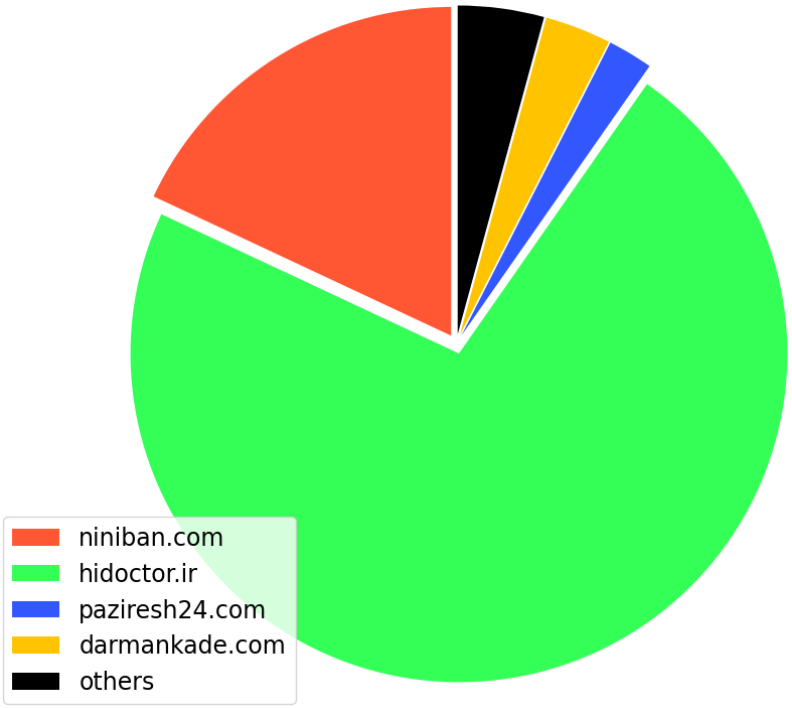
همان‌طور که در شکل ۲.۳ نشان داده شده است، برای گردآوری مجموعه داده MF3QA مراحل مختلفی طی شده است. در بخش آموزش، پرسش‌وپاسخ‌های بیمار و پزشک موجود در تالارهای گفت‌وگوی پزشکی فارسی [۴] "دکترهست" و "نی‌نی‌بان" را خزش کرده‌ایم. برای بخش اعتبارسنجی، تنها از داده‌های موجود در سایت "نی‌نی‌بان" استفاده کرده‌ایم تا انسجام بیشتری در این بخش حاصل شود. در بخش آزمایش نیز، از سایت‌های "دکترباب" و "ایزوویزیت" بهره برده‌ایم و به‌منظور اطمینان از تنوع داده‌ها، مجموعه داده پرسش‌وپاسخ K-QA [۲۲] را ترجمه کرده و به این بخش اضافه کرده‌ایم.

۲.۳.۳ فیلتر کردن رکورد های مجموعه داده MF3QA

در پایان‌نامه حاضر، بیش از صد و هشتاد هزار جفت پرسش‌وپاسخ از تالارهای گفت‌وگوی پزشکی فارسی گردآوری شده است. این جفت‌های پرسش‌وپاسخ، چه به صورت دستی [۵] و چه به صورت خودکار، مورد بررسی

[۴] Persian medical forums

[۵] فرآیند فیلتر کردن دستی توسط خانم زهرا کاظمی و آقای میلاد توکلی، از دانشجویان کارشناسی مهندسی کامپیوتر، انجام شده است.



شکل ۱.۳: سهم هر مجله در پیکره پزشکی فارسی گردآوری شده

قرار گرفته و جفت‌هایی که حاوی اطلاعات مفید نبودند، حذف شده‌اند. [۶]

این رویکرد مشابه کاری است که یونشیانگ لی و همکارانش برای توسعه مدل زبانی Chat Doctor انجام داده‌اند. [۴] آنها نیز داده‌ها را از تالارهای گفت‌وگوی پزشکی انگلیسی استخراج کرده و نیمی از جفت‌های پرسش‌وپاسخ را بر اساس طول پاسخ‌ها کنار گذاشته‌اند [۷] ، چراکه پاسخ‌های کوتاه‌تر معمولا حاوی اطلاعات مفیدی نیستند. با این حال، ما با چالش بزرگ‌تری مواجه بودیم؛ پزشکان فارسی‌زبان معمولا پاسخ‌های بسیار کوتاه‌تری نسبت به همتایان انگلیسی خود ارائه می‌دهند. این امر ما را مجبور کرد تا بیش از هشتاد درصد از رکوردهای پرسش‌وپاسخ خود را برای تضمین کیفیت کنار بگذاریم.

[۶] برای بازدید از مجموعه داده MF3QA به آدرس huggingface.co/datasets/gaokerena/MF3QA و برای بازدید از صد و هشتاد هزار جفت پرسش‌وپاسخ خزش شده به آدرس huggingface.co/datasets/gaokerena/MF3QA_uncleaned مراجعه کنید.

[۷] فیلتر کردن آنها صرفا بر اساس طول پاسخ بوده ولی همانطور که پیشتر اشاره شد ما برای فیلتر کردن از روش های دستی نیز استفاده کرده ایم.

شکل ۲.۳: سهم هر تالار گفتگو در مجموعه داده MF3QA

۱.۲.۳.۳ خزش از تالار گفتگو دکترهست

خزش از تالار گفتگوی "دکترهست"، که اصلی‌ترین منبع مجموعه داده MF3QA است، با چالش خاصی همراه بود. این تالار گفتگو تمام رکوردهای تعامل پزشک و بیمار خود را به صورت مستقیم در سایت ارائه نمی‌دهد و فقط به دو هزار رکورد آخر دسترسی می‌دهد. علاوه بر این، هر رکورد به صد رکورد مرتبط دیگر پیوند داده شده است.

برای حل این چالش، از الگوریتم ۱.۳ استفاده شد. در این روش، داده‌های تالار گفتگو به صورت یک گراف در نظر گرفته شده و با استفاده از جستجوی عرض-اول [۸] توانستیم حدود صد و بیست هزار رکورد از مجموع دویست هزار رکورد موجود در این تالار گفتگو را استخراج کنیم. این فرایند حدود دو هفته طول کشید.

۴.۳ ترجمه قسمت پزشکی مجموعه داده MMLU

۵.۳ گردآوری سوالات کنکور علوم پایه پزشکی ایران

۶.۳ ترجمه ماشینی مجموعه داده MedMCQA

[۸] breadth first search

الگوریتم ۱.۳ الگوریتم جستجو اول عرض برای استخراج رکورد های پرسش و پاسخ پزشکی

ورودی: گره های دارای دسترسی در تالار گفتگو (برگ ها)
خروجی: مجموعه‌ای از گره‌های بازدید شده

۱: یک پشته خالی S ایجاد کن
۲: یک مجموعه خالی $Visited$ ایجاد کن
۳: گره مبدأ v را به پشته S اضافه کن
۴: **تا زمانی که** پشته S خالی نیست **انجام بده**
۵: یک گره u را از پشته S بردار
۶: **اگر** گره u بازدید نشده است **آنگاه**
۷: گره u را به مجموعه $Visited$ اضافه کن
۸: **برای** هر همسایه n از گره u **انجام بده**
۹: **اگر** گره n بازدید نشده است **آنگاه**
۱۰: گره n را به پشته S اضافه کن
۱۱: **پایان شرط اگر**
۱۲: **پایان حلقهٔ برای**
۱۳: **پایان شرط اگر**
۱۴: **پایان حلقهٔ تا زمانی که**
۱۵: **بازگردان** نود های بازدید شده

فصل ۴

معرفی مدل گائوکرنا-V

فصل ۵

بررسی توانایی استدلال هوش مصنوعی

فصل ۶

معرفی مدل گائوکرنا-R

فصل ۷

نتیجه گیری

کتاب‌نامه

[۱] Vaswani, Ashish, et al. ”Attention is all you need.” Advances in neural information processing systems 30 (2017).

[۲] Singhal, Karan, et al. ”Toward expert-level medical question answering with large language models.” Nature Medicine (2025): 1-8.

[۳] Singhal, Karan, et al. ”Large language models encode clinical knowledge.” Nature 620.7972 (2023): 172-180.

[۴] Li, Yunxiang, et al. ”Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.” Cureus 15.6 (2023).

[۵] Touvron, Hugo, et al. ”Llama: Open and efficient foundation language models.” arXiv preprint arXiv:2302.13971 (2023).

[۶] Kim, Hyunjae, et al. ”Small language models learn enhanced reasoning skills from medical textbooks.” arXiv preprint arXiv:2404.00376 (2024).

[۷] Vishwanath, Krithik, et al. ”MedMobile: A mobile-sized language model with expert-level clinical capabilities.” arXiv preprint arXiv:2410.09019 (2024).

[۸] Abdin, Marah, et al. ”Phi-3 technical report: A highly capable language model locally on your phone.” / arXiv preprint arXiv:2404.14219 (2024).

[۹] Taghizadeh, Nasrin, et al. ”SINA-BERT: a pre-trained language model for analysis of medical texts in Persian.” arXiv preprint arXiv:2104.07613 (2021).

[۱۰] Koroteev, Mikhail V. ”BERT: a review of applications in natural language processing and understanding.” arXiv preprint arXiv:2103.11943 (2021).

[۱۱] Veisi, Hadi, and Hamed Fakour Shandi. ”A Persian medical question answering system.” International Journal on Artificial Intelligence Tools 29.06 (2020): 2050019.

[۱۲] Darabi, Leila. Medical Question Answering for Persian. Master’s thesis, LIACS, Leiden University, 2024.

[۱۳] Farahani, Mehrdad, et al. ”Parsbert: Transformer-based model for persian language understanding.” Neural Processing Letters 53 (2021): 3831-3847.

[۱۴] García-Ferrero, Iker, et al. ”Medical mT5: an open-source multilingual text-to-text LLM for the medical domain.” arXiv preprint arXiv:2404.07613 (2024).

[۱۵] Yang, Songhua, et al. ”Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue.” Proceedings of the AAAI conference on artificial intelligence. Vol. 38. No. 17. 2024.

[۱۶] Bao, Zhijie, et al. ”Disc-medllm: Bridging general large language models and real-world medical consultation.” arXiv preprint arXiv:2308.14346 (2023).

[۱۷] Zhang, Hongbo, et al. ”Huatuogpt, towards taming language model to be a doctor.” arXiv preprint arXiv:2305.15075 (2023).

[۱۸] Wang, Xidong, et al. ”Huatuo-26M, a Large-scale Chinese Medical QA Dataset.” Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

[۱۹] Zeng, Guangtao, et al. ”MedDialog: Large-scale medical dialogue datasets.” Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020.

[۲۰] Han, Tianyu, et al. ”MedAlpaca–an open-source collection of medical conversational AI models and training data.” arXiv preprint arXiv:2304.08247 (2023).

[۲۱] Liu, Yang, et al. ”Datasets for large language models: A comprehensive survey.” arXiv preprint arXiv:2402.18041 (2024).

[۲۲] Manes, Itay, et al. ”K-qa: A real-world medical q&a benchmark.” arXiv preprint arXiv:2401.14493 (2024).