دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

توسعه یک مدل زبانی پزشکی مبتنی بر استدلال در زبان فارسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته هوش مصنوعی و رباتیک

مهرداد قصابی

استاد راهنما

دکتر حمیدرضا برادران

بهمن ۱۴۰۴

دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

توسعه یک مدل زبانی پزشکی مبتنی بر استدلال در زبان فارسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته هوش مصنوعی و رباتیک

مهرداد قصابی

استاد راهنما

دکتر حمیدرضا برادران

بهمن ۱۴۰۴

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گواهی دفاع از پایان‌نامهٔ کارشناسی ارشد

هیأت داوران پایان‌نامهٔ کارشناسی ارشد آقای مهرداد قصابی به شمارهٔ دانشجویی ۴۰۲۳۶۱۴۰۲۹ در رشته هوش مصنوعی و رباتیک را در تاریخ ..... با عنوان «توسعه یک مدل زبانی پزشکی مبتنی بر استدلال در زبان فارسی»

	به عدد	به حروف
با نمرهٔ نهایی		

و درجه [ ] ارزیابی کرد.

ردیف	مشخصات هیات داوران	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
۱	استاد راهنما	دکتر حمیدرضا برادران	استادیار	دانشگاه اصفهان	
۲	استاد داور داخلی	دکتر داور داخلی	دانشیار	دانشگاه اصفهان	
۳	استاد داور خارجی	دکتر داور خارجی	دانشیار	دانشگاه اصفهان	

تعهدنامهٔ اصالت اثر

باسمه تعالی

اینجانب مهرداد قصابی تأیید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آن‌ها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتری ارائه نشده است.

نام و نام خانوادگی دانشجو: مهرداد قصابی

تاریخ و امضای دانشجو:

تقدیم به:

پدرم که در طول تحصیل پشتیبانم بوده است

قدردانی

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، توانایی استدلال هوش مصنوعی

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

مدل‌های زبانی در سال‌های اخیر با استفاده از داده‌های بسیار گسترده‌تر و معماری‌های پیشرفته‌تر به پیشرفت‌های چشمگیری دست یافته‌اند. این مدل‌ها توانایی درک بهتر مفاهیم، تولید متن‌های طبیعی‌تر و پاسخ‌دهی دقیق‌تر به سؤالات را پیدا کرده‌اند.

این پیشرفت‌ها منجر به افزایش چشمگیر کاربرد هوش مصنوعی [۲] در حوزه‌های مختلف، به‌ویژه در زمینه

---

[۱] reasoning

[۲] artificial intelligence

پزشکی، شده است.امروزه در حوزه پزشکی، مدل‌های زبانی مبتنی بر یادگیری ژرف [۳] 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 کشور های فارسی زبان 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

[۳] deep learning

[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.

بهینه‌سازی دانش پزشکی مدل می‌پردازد. در فاز دوم این پژوهش ابتدا تکنیک های جدیدی برای ارتقای توانایی استدلال و درک سیستم دو مدل معرفی شده و سپس مدل گائوکرنا−R در این فاز توسعه داده شده است. 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
توانایی استدلال	خیر	بله
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 در حوزه زبان فارسی، فرآیند گردآوری 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 استدلال ارائه خواهند شد. در ادامه، در فصل ششم، با معرفی تکنیک هایی برای بهبود توانایی استدلال یک مدل زبانی مدل پیشرفته‌تری به

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

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.

فصل ۲

ادبیات موضوع

۱.۲ مقدمه

در این فصل به بررسی مفاهیم مدل‌های زبانی [۱] انواع آن و چگونگی سنجش آن‌ها خواهیم پرداخت. همچنین، انواع سیستم‌های پرسش و پاسخ [۲] را مورد تحلیل قرار خواهیم داد تا درک بهتری از عملکرد و کاربردهای مختلف این سیستم‌ها به دست آوریم.

۲.۲ مدل های زبانی

مدل‌های زبانی ابزارهای پیشرفته‌ای هستند که برای پردازش و تولید زبان طبیعی طراحی شده‌اند. این مدل‌ها با استفاده از یادگیری ماشین [۳] و به ویژه یادگیری عمیق [۴] ، توانایی درک و تولید متن را دارند.

در حقیقت، وظیفه مدل‌های زبانی پیش بینی توکن بعدی [۵] بر اساس متنی است که تاکنون تولید شده است.

این مدل ها دارای انواع مختلفی است که در ادامه آنها را بررسی میکنیم.

---

[۱] language models
[۲] question answering systems
[۳] machine learning
[۴] deep learning
[۵] next token prediction

۱.۲.۲ مدل های زبانی آماری

مدل‌های زبانی n-gram ابتدایی ترین مدل های زبانی هستند که در دهه نود میلادی به عنوان جایگزینی برای ترجمه ماشینی مبتنی بر قانون [۶] معرفی شدند [۱] همانطور که در فرمول ۱.۲.۲ مشاهده میکنید این مدل‌ها برای پیش‌بینی توکن بعدی از فراوانی n-gram ها در پیکره [۷] موجود استفاده میکنند. به عنوان مثال در مدل زبانی 2-gram ، احتمال وقوع یک کلمه تنها بر اساس کلمه قبلی محاسبه می‌شود.

$$P(w_۱, w_۲, ..., w_n) = \prod_{i=۱}^{n} P(w_i | w_{i-n+۱}, w_{i-n+۲}, ..., w_{i-۱})$$

۲.۲.۲ مدل های زبانی بازگشتی

شبکه‌های عصبی بازگشتی نوعی از شبکه‌های عصبی مصنوعی [۸] هستند که به طور خاص برای پردازش توالی‌ها [۹] طراحی شده‌اند. [۲] از آنجا که زبان را نیز می‌توان به صورت یک توالی از توکن‌ها تعریف کرد، بنابراین می‌توان از شبکه‌های عصبی بازگشتی به عنوان مدل‌های زبانی استفاده کرد.

همانطور که در فرمول ۲.۲.۲ که فرمول پایه شبکه های عصبی است می بینید ویژگی اصلی شبکه‌های عصبی بازگشتی این است که دارای حلقه‌های بازگشتی است که به آن اجازه می‌دهد اطلاعات را از مراحل قبلی به مراحل بعدی منتقل کند. این ساختار باعث می‌شود شبکه‌های عصبی بازگشتی بتواند وابستگی‌های زمانی و ترتیبی داده‌ها را مدل‌سازی کند.

شبکه های عصبی بازگشتی مانند نوع ساده آن [۱۰] معمولا در مدل کردن وابستگی های بلند مدت [۱۱] با مشکل مواجه میشوند، این وابستگی ها که در زبان های طبیعی به وفور یافت میشوند باعث شده اند که مدل های زبانی مبتنی بر شبکه های عصبی بازگشتی مدل های چندان خوبی نباشند. هر چند در انواع دیگر این شبکه ها مانند LSTM [۳] برای حل این مشکل تلاش شده است اما این مشکل هنوز در این نوع از شبکه های عصبی وجود دارد.

$$h_t = f(W_h h_{t-۱} + W_x x_t + b)$$

[۶] rule-based machine translation
[۷] corpus
[۸] artificial neural networks
[۹] sequences
[۱۰] vanilla recurrent neural networks
[۱۱] long term dependencies

۳.۲.۲ مدل های زبانی مبتنی بر ترنسفورمر

همان‌طور که در قسمت پیشین اشاره شد، مدل‌های زبانی بازگشتی در مدل‌سازی وابستگی‌های بلندمدت با مشکل مواجه هستند و به دلیل ماهیت توالی‌گونه خود، سرعت پردازش آنها نیز بسیار پایین است.

مقاله Attention is All You Need [۴] با معماری نوآورانه خود توانست هر دوی این مشکلات را حل کند. این معماری از مکانیزم توجه [۱۲] استفاده می‌کند که به مدل اجازه می‌دهد تا به طور همزمان به تمامی ورودی‌ها توجه کند و وابستگی‌های بلندمدت را به راحتی شناسایی کند. به این ترتیب، سرعت پردازش به طور قابل توجهی افزایش می‌یابد.

علاوه بر این، معماری ترنسفورمر [۱۳] که در شکل ۱.۲ میتوانید آن را ببینید ، قابلیت پردازش موازی را دارد، زیرا معماری آن ماهیت توالی‌گونه ندارد. به همین دلیل، با استفاده از پردازنده‌های گرافیکی [۱۴] می‌توان پردازش‌های موازی را انجام داد که سرعت پردازش را به طور چشمگیری بهبود می‌بخشد.

مدل های مبتنی بر معماری ترنسفورمر سه دسته هستند که در ادامه به بررسی این سه دسته خواهیم پرداخت.

۱.۳.۲.۲ مدل‌های زبانی فقط رمزگذار

مدل‌های زبانی فقط رمزگذار [۱۵] ، تنها از بخش رمزگذار معماری ترنسفورمر که در سمت چپ شکل ۱.۲ قابل مشاهده است، استفاده می‌کنند. این مدل‌ها به‌طور ویژه برای پردازش و درک متن طراحی شده‌اند و ورودی‌ها را به یک نمایش داخلی تبدیل می‌کنند که شامل اطلاعات معنایی و ساختاری متن است.

این نوع مدل‌ها عمدتا در وظایفی مانند تحلیل احساسات [۱۶] ، دسته‌بندی متن [۱۷] و استخراج ویژگی‌ها [۱۸] کاربرد دارند. آن‌ها به‌خوبی می‌توانند الگوهای زبانی و معنایی را شناسایی کرده و اطلاعات مفیدی از داده‌های متنی استخراج کنند، اما به‌تنهایی توانایی تولید متن جدید را ندارند.

به‌عنوان مثال، می‌توان از BERT [۵] [۶] ، که یکی از معروف‌ترین مدل‌های زبانی فقط رمزگذار است، یاد کرد.

---

[۱۲] Attention Mechanism
[۱۳] transformer
[۱۴] graphical processing unit
[۱۵] encoder only language models
[۱۶] sentiment analysis
[۱۷] text classification
[۱۸] feature extraction

۲.۳.۲.۲ مدل‌های زبانی فقط رمزگشا

مدل‌های زبانی فقط رمزگشا [۱۹] نوعی از مدل‌های زبانی هستند که به طور خاص برای تولید متن و پیش‌بینی توکن‌های بعدی در یک توالی طراحی شده‌اند. این مدل‌ها تنها از ساختار رمزگشای معماری ترنسفورمر که در سمت راست شکل ۱.۲ قابل مشاهده است، استفاده می‌کنند و به صورت تک‌جهته عمل می‌کنند، به این معنا که برای تولید هر توکن، تنها به توکن‌های قبلی خود در توالی دسترسی دارند.

در این مدل‌ها، هدف اصلی پیش‌بینی توکن بعدی بر اساس توکن‌های قبلی است. به عنوان مثال، اگر ورودی مدل یک جمله باشد، مدل سعی می‌کند کلمه بعدی را پیش‌بینی کند. این نوع از مدل‌ها در کاربردهایی مانند تولید متن، چت‌بات‌ها و ترجمه ماشینی بسیار موثر هستند.

مدل‌های رمزگشا معمولا با استفاده از داده‌های متنی بزرگ آموزش دیده و توانایی بالایی در تولید متن‌های معنادار و مرتبط دارند. یکی از معروف‌ترین نمونه‌های این دسته از مدل‌ها، مدل GPT [۲۰] [۷] است که توسط OpenAI توسعه یافته است.

۳.۳.۲.۲ مدل‌های زبانی رمزگذار-رمزگشا

دل‌های زبانی رمزگذار-رمزگشا [۲۱] نوعی از مدل‌های زبانی هستند که هر دو قسمت معماری ترنسفورمر که در شکل ۱.۲ مشاهده میکنید استفاده میکند.

این معماری که در مدل هایی مانند T5 [۸] به کار گرفته شده در وظایف پیچیده‌ای مانند ترجمه ماشینی، خلاصه‌سازی متن و تولید گفتار کاربرد دارد. به عنوان مثال، در ترجمه ماشینی، بخش رمزگذار جمله‌ای را به زبان مبدا تحلیل کرده و آن را به یک نمایش معنایی تبدیل می‌کند، سپس بخش رمزگشا این نمایش را به زبان مقصد ترجمه می‌کند.

۳.۲ سیستم های پرسش و پاسخ

سیستم‌های پرسش و پاسخ [۲۲] فناوری‌های هوشمندی هستند که با استفاده از تکنیک‌های پردازش زبان طبیعی [۲۳] ، به کاربران این امکان را می‌دهند تا سوالات خود را مطرح کرده و پاسخ‌های دقیق و مرتبط دریافت کنند. این

[۱۹] Decoder-Only Language Models
[۲۰] Generative Pre-trained Transformer
[۲۱] Encoder-Decoder Language Models
[۲۲] Question Answering Systems
[۲۳] Natural Language Processing

سیستم ها بر دو نوع هستند استخراجی [۲۴] و تولیدی [۲۵] که در ادامه به بررسی آنها خواهیم پرداخت. [۹]

۱.۳.۲ سیستم های پرسش و پاسخ استخراجی

سیستم‌های پرسش و پاسخ استخراجی به منظور پاسخ‌گویی به سوالات کاربران،به جستجوی اطلاعات در پایگاه‌های داده یا مستندات می‌پردازند و پاسخ‌ها را از متن استخراج می‌کنند. این سیستم‌ها معمولا از یک مدل زبانی فقط رمزگذار مانند bert استفاده میکنند.

۲.۳.۲ سیستم های پرسش و پاسخ تولیدی

با توسعه و گسترش هوش مصنوعی تولید کننده [۲۶] سیستم‌های پرسش و پاسخ‌هایی پدید آمدند که در آن برای پاسخ دادن به پرسش کاربر به دانش مدل زبانی تکیه میشود. یعنی مدل زبانی با توجه به دانشی که در زمینه پرسش مطرح شده دارد بایستی پاسخ را تولید کند. از آنجایی که در این سیستم‌ها بایستی چیزی تولید شود بنابراین در آن‌ها از مدل‌های زبانی فقط رمزگشا یا مدل‌های زبانی رمزگذار-رمزگشا استفاده میگردد.

با بهبود دانش پزشکی و توانایی استدلال یک مدل پایه ما نیز در پایان نامه حاضر اقدام به طراحی یک سیستم پرسش و پاسخ تولیدی کرده ایم.

۴.۲ شیوه های سنجش مدل های زبانی

سنجش دانش یک مدل زبانی به ویژه در زمینه پزشکی از اهمیت بالایی برخوردار است، زیرا این فرآیند به شناخت ما از میزان دانش یک مدل زبانی، کمک شایانی می‌کند. در ادامه چندین روش سنجش کیفیت پاسخ‌های مدل های زبانی را مطرح خواهیم کرد، ضمنا ما از روش اول و دوم برای سنجش دانش مدل های خود استفاده کرده ایم.

---

[۲۴] extractive
[۲۵] generative
[۲۶] inteligence artificial generative

۱.۴.۲ سنجش بر اساس میزان پاسخگویی به پرسش و پاسخ های چند گزینه ای

یکی از معیارهای مهم برای سنجش عملکرد مدل‌های زبانی، ارزیابی توانایی آن‌ها در پاسخگویی به پرسش‌های چهار گزینه‌ای است که از پیش آماده شده‌اند. این نوع ارزیابی به دلیل ساختار مشخص و استاندارد پرسش‌ها، امکان مقایسه دقیق‌تری بین مدل‌های مختلف را فراهم می‌آورد. یکی از مجموعه‌های داده‌ای که به طور گسترده در این زمینه مورد استفاده قرار می‌گیرد، مجموعه داده MMLU [۲۷] [۱۰] است. این مجموعه شامل پرسش‌های متنوعی است که درباره موضوعات مختلفی مانند علوم،پزشکی، ریاضیات، تاریخ، و ادبیات طراحی شده‌اند.

مجموعه داده MMLU به عنوان یک استاندارد در ارزیابی مدل‌های زبانی، به محققان و توسعه‌دهندگان این امکان را می‌دهد که عملکرد مدل های زبانی خود را در زمینه‌های مختلف بسنجند و نقاط قوت و ضعف آن‌ها را شناسایی کنند. پرسش‌های چهار گزینه‌ای در MMLU به گونه‌ای طراحی شده‌اند که نیاز به درک عمیق و تحلیل دقیق متن دارند. این ویژگی، مدل‌ها را به چالش می‌کشد تا نه تنها اطلاعات را بازیابی کنند، بلکه توانایی استدلال و تحلیل خود را نیز به نمایش بگذارند. با استفاده از این معیار، می‌توان به راحتی مقایسه‌هایی بین مدل‌های مختلف انجام داد و پیشرفت‌های حاصل شده در زمینه هوش مصنوعی و پردازش زبان طبیعی را ارزیابی کرد.

۲.۴.۲ سنجش بر اساس نظر مدل داور

یک روش دیگر برای سنجش عملکرد یک مدل زبانی، استفاده از یک مدل زبانی دیگر به عنوان داور است. در این رویکرد، یک مدل زبانی مستقل به عنوان مرجع برای ارزیابی کیفیت پاسخ‌های تولید شده توسط مدل اصلی مورد استفاده قرار می‌گیرد. این روش به دلیل قابلیت‌های بالای مدل‌های زبانی در پردازش و درک زبان طبیعی، می‌تواند به طور موثری به ارزیابی دقت و کیفیت پاسخ‌ها کمک کند.

در این فرآیند، پاسخ‌های تولید شده توسط مدل اصلی به مدل قاضی ارائه می‌شود. مدل قاضی می‌تواند با استفاده از معیارهای مختلفی مانند شباهت معنایی با پاسخ اصلی [۲۸] ، صحت اطلاعات، و سازگاری با زمینه، کیفیت پاسخ‌ها را ارزیابی کند. به عنوان مثال، مدل قاضی می‌تواند با بررسی تطابق پاسخ‌ها با اطلاعات موجود در متون معتبر یا داده‌های آموزشی، نمره‌ای برای هر پاسخ تولید کرده [۱۱] یا پاسخی را بر پاسخ دیگر ترجیح دهد.

---

[۲۷] Massive Multitask Language Understanding

[۲۸] ground truth

۳.۴.۲ سنجش بر اساس استنتاج زبان طبیعی

با داشتن یک مجموعه داده مانند K-QA [۱۲] ، شامل پاسخ‌های تولید شده توسط انسان که به همراه توضیحات دقیقی دسته‌بندی شده‌اند، این پاسخ‌ها به عنوان "الزامی" [۲۹] یا "مفید" [۳۰] مشخص گشته‌اند. این دسته‌بندی نشان می‌دهد که آیا توضیحات باید به طور ضروری در پاسخ گنجانده شوند یا اینکه اضافی و مفید هستند.

این حقایق اتمی می‌توانند برای به کارگیری یک روش ارزیابی مبتنی بر استنتاج زبان طبیعی [۳۱] استفاده شوند. در این روش، پاسخ مدل به عنوان "مقدمه" [۳۲] و هر یک از توضیحات انسانی به عنوان "فرضیه" [۳۳] در نظر گرفته می‌شود. سپس یک مدل زبانی توانا در حوزه استنتاج زبان طبیعی تعیین خواهد کرد که آیا مقدمه مستلزم [۳۴] ، متناقض [۳۵] یا خنثی [۳۶] با فرضیه است.

با انجام این کار روی همه رکوردهای مجموعه داده، دو امتیاز کامل بودن [۳۷] و حقیقت داشتن [۳۸] به صورت زیر به دست می‌آید.

$$S_{comp}(r_i, A'_i) = \frac{۱[r_i \quad entails \quad a]}{|A'_i|}$$

$$S_{fact}(r_i, A'_i) = \begin{cases} ۰ & \text{if} \quad \exists a \in A_i \quad s.t. \quad r_i \quad contradicts \quad a \\ ۱ & \text{if} \quad otherwise \end{cases}$$

۴.۴.۲ سنجش بر اساس امتیاز bert

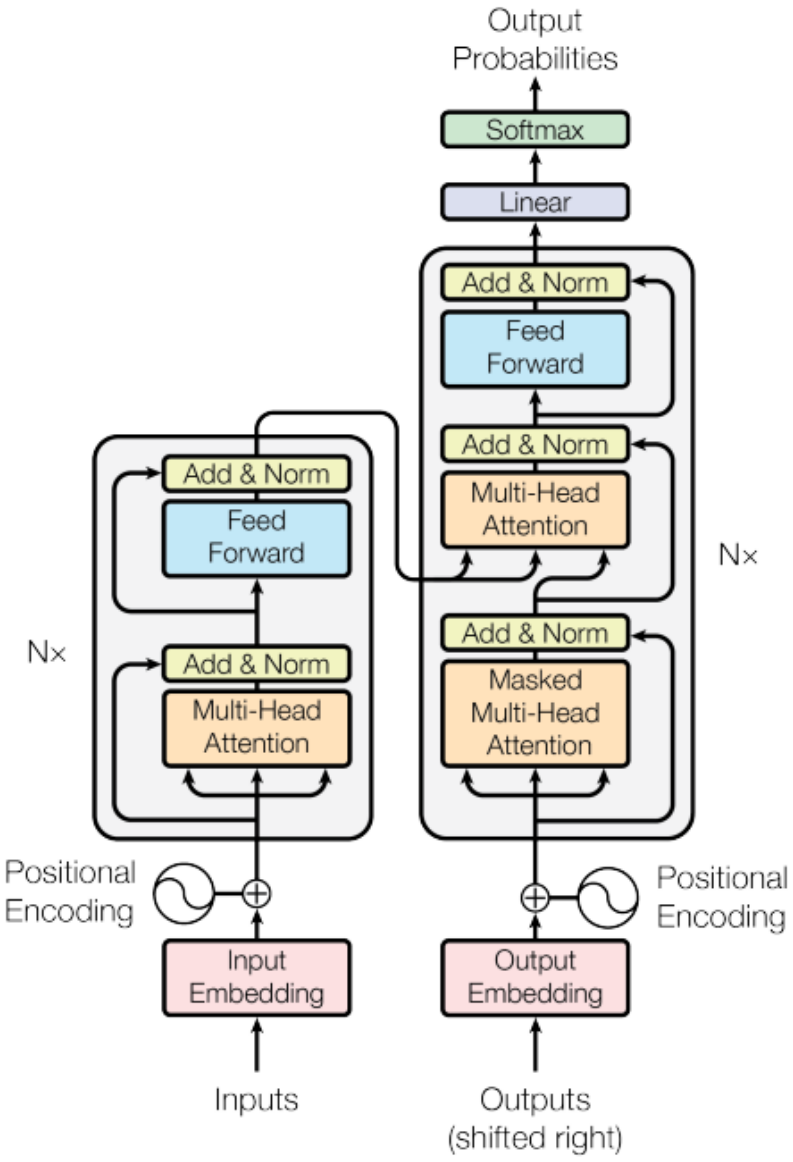
در این معیار، کیفیت پاسخ‌های تولید شده توسط یک مدل زبانی فقط رمزگذار مانند bert ارزیابی می‌شود. در این فرآیند، پاسخ داده شده با پاسخ صحیح [۳۹] موجود در مجموعه داده مقایسه می‌گردد.

[۲۹] must have
[۳۰] nice to have
[۳۱] natural language inference
[۳۲] premise
[۳۳] hypothesis
[۳۴] entailment
[۳۵] contradiction
[۳۶] neutral
[۳۷] completeness
[۳۸] factuality
[۳۹] ground truth

مدل زبانی پاسخ فقط رمزگذار را به دو بردار تبدیل می‌کند: یکی برای پاسخ داده شده و دیگری برای پاسخ صحیح. سپس برای محاسبه امتیاز bert ، کافی است میزان شباهت این دو بردار را با یک روش شباهت‌سنجی مانند شباهت کسینوسی [۴۰] محاسبه کنیم.

پس از محاسبه امتیازها برای تمامی پاسخ‌ها در مجموعه داده، این امتیازات جمع‌آوری و میانگین‌گیری می‌شوند. میانگین امتیازها نمای کلی از عملکرد مدل را ارائه می‌دهد و به شناسایی نقاط قوت و ضعف آن کمک می‌کند.

[۴۰] cosine similarity



شکل ۱.۲: معماری ترنسفورمر

فصل ۳

بررسی کار های پیشین

۱.۳ مقدمه

همان‌طور که پیش‌تر اشاره شد، علیرغم پیشرفت‌های چشمگیر در توسعه مدل‌های زبانی پزشکی به زبان انگلیسی، مانند توسعه و معرفی مدل‌های MedPalm [۱۳] [۱۴] یا مدل Med-Gemini [۱۵] ،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.

۲.۳ کار های پیشین در حوزه زبان انگلیسی

۱.۲.۳ 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.

۲.۲.۳ مدل ChatDoctor

مدل ChatDoctor [۱۶] 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 [۱۷] 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.

۳.۲.۳ مدل های Meerkat

مدل‌های Meerkat [۱۸] 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.

۴.۲.۳ مدل MedMobile

MedMobile [۱۹] تلاشی دیگر در حوزه مدل‌های زبانی کوچک پزشکی است. برای توسعه این مدل زبانی کوچک، مدل Phi-3-mini [۲۰] به عنوان مدل پایه [۵] استفاده از ترکیبی از داده‌های مصنوعی و تولیدشده توسط انسان تنظیم دقیق [۶] 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 [۷] ممکن می‌سازد.

۳.۳ 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 [۸] هستند.

از سوی دیگر، تمامی این تلاش‌ها عمدتا بر روی راهکارهای استخراجی [۹] متمرکز بوده‌اند که هدفشان بازیابی اطلاعات مرتبط از منابع از پیش تعریف شده است، به جای استفاده از رویکردهای تولیدی [۱۰] که قادر به تولید پاسخ‌های آگاه از زمینه باشند.

---

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

۱.۳.۳ مدل Sina-bert

شاید اولین و برجسته ترین مدل زبانی پزشکی فارسی، Sina-BERT [۲۱] باشد که شامل آموزش یک مدل BERT [۲۲] با استفاده از یک پیکره خزش شده [۱۱] 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 [۱۲] یک مدل زبانی مبتنی بر رمزگذار [۱۳] 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.

۲.۳.۳ 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 [۲۳] 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.

۳.۳.۳ پایان نامه کارشناسی ارشد خانم لیلا دارابی

مشابه به این دو اثر، پیشین لیلا دارابی در پایان نامه ارشد خود [۲۴] از مدل‌هایی مانند Pars-BERT [۲۵] 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

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

این، روش‌های طبقه‌بندی و شناسایی موجودیت‌های نامدار [۱۴] برای بهبود ارتباط پاسخ‌ها از طریق دسته‌بندی سوالات و شناسایی موجودیت‌های پزشکی مانند نام داروها و بیماری‌ها به کار گرفته می‌شوند.

[۱۴] Named Entity Recognition (NER)

فصل ۴

جمع آوری دادگان

۱.۴ 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.

۲.۴ 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 [۲۶] 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 ۱.۴ 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 ۱.۴ 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 ۱.۴: مقایسه پیکره گردآوری شده با پیکره های گردآوری شده توسط I. Garcia Ferrero et al.

۳.۴ 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 [۲۷] به چندین مجموعه داده واقعی پرسش وپاسخ پزشک وبیمار اشاره کرده است، مقایسه‌ای میان این مجموعه دادگان و مجموعه داده ما در جدول ۲.۴ ارائه شده است.

[۳] survey

نام مجموعه داده	زبان	تعداد پرسش و پاسخ ها	گردآورنده
ChatDoctor	انگلیسی	100K	[۱۵] Yunxiang Li et al.
CMtMedQA	چینی	68K	[۲۸] Songhua Yang et al.
DISC-Med-SFT	چینی	465K	[۲۹] Zhijie Bao et al.
HuatuoGPT-sft-data-v1	چینی	226K	[۳۰] Hongbo Zhang et al.
Huatuo-26M	چینی	26M	[۳۱] Jianquan Li et al.
MedDialog	چینی و انگلیسی	3.66M	[۳۲]Guangtao Zeng et al.
Medical-Meadow	انگلیسی	160k	[۳۳] Tianyu Han et al.
MF3QA	فارسی	20k	ما

جدول ۲.۴: مقایسه مجموعه داده های پرسش و پاسخ آزاد پزشکی با مجموعه داده گردآوری شده

۱.۳.۴ منابع مجموعه داده MF3QA

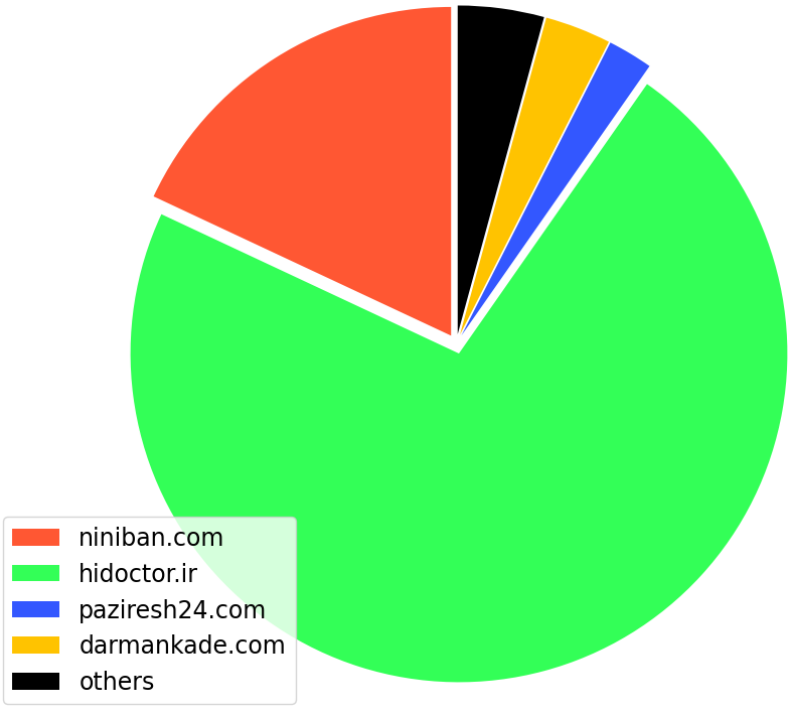
همان‌طور که در شکل ۲.۴ نشان داده شده است، برای گردآوری مجموعه داده MF3QA مراحل مختلفی طی شده است. در بخش آموزش، پرسش‌وپاسخ‌های بیمار و پزشک موجود در تالارهای گفت‌وگوی پزشکی فارسی [۴] "دکترهست" و "نی‌نی‌بان" را خزش کرده‌ایم. برای بخش اعتبارسنجی، تنها از داده‌های موجود در سایت "نی‌نی‌بان" استفاده کرده‌ایم تا انسجام بیشتری در این بخش حاصل شود. در بخش آزمایش نیز، از سایت‌های "دکتریاب" و "ایزوویزیت" بهره برده‌ایم و به‌منظور اطمینان از تنوع داده‌ها، مجموعه داده پرسش‌وپاسخ K-QA [۱۲] را ترجمه کرده و به این بخش اضافه کرده‌ایم.

۲.۳.۴ 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 ۱.۴: 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. [۱۶] 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 ۲.۴: سهم هر تالار گفتگو در مجموعه داده MF3QA

۱.۲.۳.۴ 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 ۱.۴ 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.

۴.۴ ترجمه قسمت پزشکی مجموعه داده MMLU

مجموعه داده MMLU [۱۰] [۹] یکی از معتبرترین مجموعه داده‌ها برای ارزیابی توانایی مدل‌های زبانی در درک و پاسخ‌دهی به سوالات چندوظیفه‌ای است. این مجموعه شامل سوالاتی در موضوعات مختلف از جمله علوم پزشکی، مهندسی، علوم انسانی و دیگر حوزه‌ها است که به‌صورت چندگزینه‌ای طراحی شده‌اند.

در پروژه ما، برای ارزیابی مدل زبانی پزشکی توسعه‌یافته، بخش پزشکی این مجموعه داده را به زبان فارسی ترجمه کردیم. [۱۰] هدف از این کار، تطبیق داده‌های ارزیابی با زبان مورد استفاده در مدل و بررسی توانایی مدل در پاسخ‌دهی دقیق به سوالات تخصصی پزشکی در زبان فارسی بود. [۱۱]

[۸] breadth first search
[۹] Massive Multitask Language Understanding
[۱۰] ترجمه توسط آقای امیرحسین پورسینا دانشجوی پزشکی انجام شده است.
[۱۱] برای بازدید از ترجمه این مجموعه داده به آدرس huggingface.co/datasets/gaokerena/FA_MED_MMLU مراجعه کنید.

الگوریتم ۱.۴ 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

۵.۴ گردآوری سوالات کنکور علوم پایه پزشکی ایران

آزمون علوم پایه پزشکی یک آزمون سراسری در ایران است که دانشجویان پزشکی موظف هستند پس از گذراندن دروس علوم پایه معمولا در پنج ترم در آن شرکت کنند. این آزمون به منظور سنجش میزان آموخته‌های دانشجویان از دروس علوم پایه و آمادگی آنها برای ورود به مراحل بالینی برگزار می‌شود. در صورتی که دانشجو پس از سه مرتبه در این آزمون قبول نشود [۱۲] از ادامه تحصیل در رشته پزشکی محروم میشود.

برای سنجش دانش پزشکی مدل زبانی خود ما سوالات این آزمون را از pdf سوالاتی که سازمان سنجش برای آن منتشر میکند استخراج کرده ایم. [۱۳] [۱۴]

[۱۲] نمره قبولی در این آزمون در سالهای مختلف متفاوت است و معمولا حدود سی و شش درصد میباشد
[۱۳] برای این کار از کتابخانه fitz پایتون استفاده شده است.
[۱۴] برای بازدید از این مجموعه داده به آدرس https://huggingface.co/datasets/gaokerena/KOPP مراجعه کنید.

۶.۴ ترجمه ماشینی مجموعه داده MedMCQA

فصل ۵

معرفی مدل گائوکرنا-V

۱.۵ مقدمه

در این فصل با استفاده از دادگانی که گردآوری کرده ایم یک مدل پایه [۱] را تمرین میدهیم تا مدل جدید مدل گائوکرنا-V را معرفی کنیم که همانطور که در فصل نخست درباره آن صحبت شد دارای توانایی استدلال نیست.

در ادامه با استفاده از ترجمه قسمت پزشکی مجموعه داده MMLU به مقایسه مدل جدید خود با مدل پایه و بقیه جایگزین ها خواهیم پرداخت.

۲.۵ مدل پایه

به دلیل عدم وجود یک مدل زبانی پزشکی فارسی متن باز [۲] ما مجبور به انتخاب یک مدل زبانی همه منظوره [۳] به عنوان مدل پایه هستیم.

گزینه‌های متعددی مانند Qwen2 [۳۴]، aya-expanse [۳۵]، Gemma2 [۳۶] و PersianMind [۳۷] برای انتخاب در دسترس بودند، که به دو دلیل زیر مدل aya-expanse-8b انتخاب شده است.

دلیل نخست این است که داده‌های آموزشی سایر مدل‌ها عمدتا شامل زبان‌های غیر فارسی هستند، که این امر می‌تواند منجر به ایجاد سوگیری‌هایی در مدل شود که حتی در صورت دستور صریح به استفاده از زبان فارسی،

[۱] baseline model
[۲] open source
[۳] general purpose

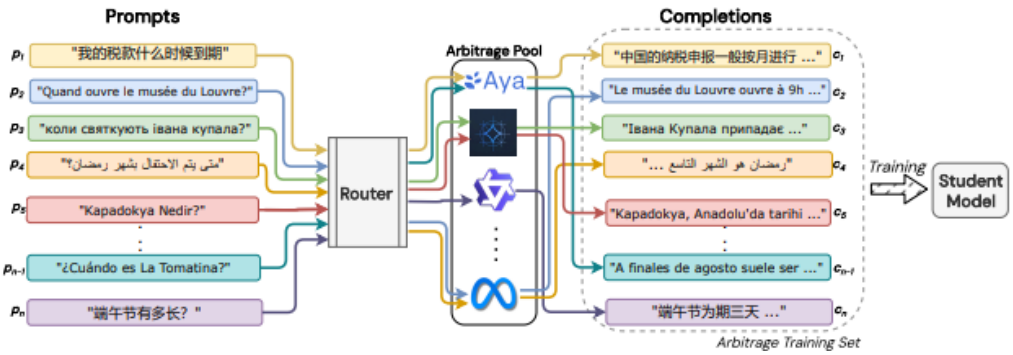
باعث تولید کاراکترهای غیر فارسی می‌شود. در مقابل، aya-expanse درک قوی‌ای از دستور زبان فارسی نشان می‌دهد و متنی غنی و دستوری صحیح به زبان فارسی تولید می‌کند، که آن را به گزینه‌ای بهتر برای پژوهش ما تبدیل می‌کند.

علاوه بر این، اگر ما پارامترهای به‌روزرسانی‌شده خود را به جای aya-expanse در مدل دیگری از خانواده aya ، یعنی aya-vision [۳۸] ادغام کنیم، این امکان را به دست می‌آوریم که تصاویر پزشکی مانند MRI و CT scan را به عنوان ورودی بپذیریم. این امر باعث افزایش قابلیت کاربرد مدل ما در حوزه پزشکی خواهد شد.

۱.۲.۵ ویژگی های مدل aya-expanse

همانطور که پیشتر از آن یاد شد مدل زبانی aya-expanse قابلیت تولید متن های فارسی غنی با دستور زبانی صحیح را داراست [۴] ویژگی که بقیه مدل های زبانی تنها برای برخی از زبان های دارای منابع غنی [۵] مانند انگلیسی و چینی دارا هستند.

این ویژگی زمانی قابل دسترس است که علاوه بر در دسترس بودن منابع غنی برای تمرین منابع به صورت مساوی بین زبان هایی که مدل زبانی پشتیبانی میکند تقسیم شده باشد؛ aya-expanse با تولید دادگان مصنوعی [۶] بر مشکل نبود دادگان فائق آمده است، البته تمرین دادن یک مدل زبانی با دادگان مصنوعی تولید شده توسط یک مدل زبانی دیگر باعث فروپاشی مدل [۷] [۳۹] میشود که aya-expanse با معرفی مکانیسم آربیتاژ داده [۴۰] که در شکل ۱.۵ میبینید از رخ دادن این اتفاق جلوگیری کرده است.



شکل ۱.۵: مکانیسم آربیتاژ داده

[۴] به جز زبان فارسی این مدل از بیست و دو زبان دیگر نیز پشتیبانی میکند. برای اطلاعات بیشتر به پیوند https://huggingface.co/CohereLabs/aya-expanse-8b مراجعه کنید.
[۵] Resource-rich languages
[۶] synthetic data
[۷] model collapse

۳.۵ تنظیم دقیق روی پیکره پزشکی

برای دستیابی به مدل مدل گائوکرنا-V ابتدا مدل پایه روی شصت درصد از پیکره پزشکی گردآوری شده تنظیم دقیق [۸] شده است [۹] ،برای این کار از اندازه دسته [۱۰] برابر با دو استفاده کردیم تا نیاز به حافظه در طول آموزش کاهش یابد. علاوه بر این، از تجمع گرادیان [۱۱] با شانزده مرحله استفاده کردیم که به طور مؤثر اندازه‌ی کلی دسته را به سی و دو افزایش داد و دینامیک پایدار آموزش را فراهم کرد.

برای کاهش بیشتر مصرف حافظه در فرآیند تنظیم دقیق، از روش LoRA [۴۱] [۱۲] بهره بردیم تا تعداد پارامترهای قابل آموزش به‌طور چشمگیری کاهش یابد. برای این کار از رتبه [۱۳] برابر با هشت، مقدار آلفا برابر با شانزده، نرخ حذف [۱۴] برابر با پنج درصد و نرخ پوسیدگی وزن [۱۵] ده درصد استفاده کرده و وزن‌های LoRA را به تمام پارامترهای قابل آموزش در هر لایه ترنسفورمر اختصاص دادیم.

برای بهینه‌سازی بیشتر این فرآیند، از تکنیک‌های کارآمد توکن‌سازی [۱۶] و تکنیک های آموزش مبتنی بر مدیریت حافظه [۱۷] بهره بردیم. فرآیند توکن‌سازی متن ورودی را به دنباله‌های قابل مدیریت توکن تقسیم کرد، و با کوتاه کردن، پر کردن، و مدیریت توکن‌های اضافی، ساختار ورودی و برچسب‌ها را ثابت نگه داشت تا یکپارچگی مفهومی در طول ثابت زمینه حفظ شود. این آماده‌سازی ساده‌شده، همراه با تنظیم دقیق مبتنی بر LoRA ،وزاستفاده از Flash Attention 2 [۴۲] بهبود یافت.

Flash Attention 2 با کاهش سربار حافظه، امکان مدیریت طول‌های زمینه بلندتر و اندازه‌های دسته بزرگ‌تر را به صورت کارآمد فراهم میکند، که به تنظیم دقیق موثر کمک کرده و تعادل بین کارایی محاسباتی و عملکرد مدل را برقرار میسازد.

---

[۸] fine tune

[۹] برای دسترسی به کد تنظیم دقیق مدل پایه روی پیکره پزشکی گرد آوری شده میتوانید به آدرس https://github.com/Mehrdadghassabi/Gaokerena-V/blob/main/fine-tuning/Pretraining.ipynb مراجعه کنید

[۱۰] batch size

[۱۱] gradient accumulation

[۱۲] low rank adaptation

[۱۳] rank

[۱۴] drop out

[۱۵] weight decay

[۱۶] efficient tokenization

[۱۷] memory-aware training techniques

۴.۵ تنظیم دستورالعملی روی مجموعه داده MF3QA

پس از تنظیم دقیق روی پیکره پزشکی، بایستی فرآیند تنظیم دستورالعملی [۱۸] را با استفاده از مجموعه داده MF3QA که فصل پیشین معرفی شد انجام دهیم. [۱۹] برای این کار به طور مشخص، باز هم از روش LoRA استفاده کرده ایم ولی این بار با رتبه [۲۰] برابر با دو، آلفا برابر با دو، نرخ حذف برابر با چهل درصد و نرخ پوسیدگی وزن [۲۱] پنجاه درصد استفاده کرده ایم. فرآیند تنظیم دستورالعمل تنها برای یک دوره [۲۲] انجام شد تا مدل شیوه پاسخگویی درست را بهتر درک کند.

۵.۵ رد پای کربن مدل گائوکرنا-V

اثر کربنی حاصل از بهینه‌سازی مدل گائوکرنا-V که شامل مراحل تنظیم دقیق و تنظیم دستورالعملی میشود، بر اساس مشخصات سخت‌افزاری و مدت زمان اجرا تخمین زده شده است. فرآیند تمرین به مدت نوزده ساعت بر روی کارت گرافیک NVIDIA A100 PCIe 40GB که در منطقه شرق آسیا [۲۳] پلتفرم ابری گوگل میزبانی می‌شدند، اجرا شده است.

با توجه به مصرف برق معمولی هر پردازنده گرافیکی که برابر با دویست و پنجاه وات است، کل انرژی مصرف‌شده در این مدت برابر 4.750 کیلووات‌ساعت است. با در نظر گرفتن ضریب شدت کربن شبکه شرق آسیا که برابر با 560 گرم کربن دی اکسید معادل به ازای هر کیلووات‌ساعت است، میزان انتشار کربن در طول این فرآیند به 2660 گرم می‌رسد. [۴۳]

---

[۱۸] instruction tuning

[۱۹] برای دسترسی به کد تنظیم دستورالعملی روی مجموعه داده MF3QA میتوانید به آدرس https://github.com/Mehrdadghassabi/Gaokerena-V/blob/main/fine-tuning/InstructionTuning.ipynb مراجعه کنید

[۲۰] rank

[۲۱] weight decay

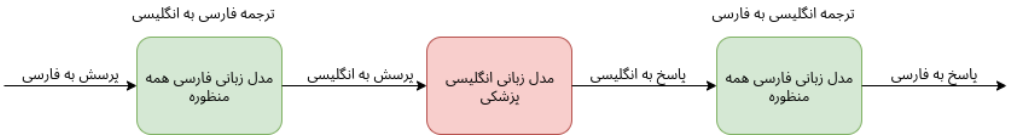
[۲۲] epoch

[۲۳] asia-east1

۶.۵ نتایج

در نبود مدل های پزشکی فارسی برای مقایسه ما مدل گائوکرنا-V را با مدل های زبانی فارسی همه منظوره [۲۴] و جایگزین های خط لوله ای [۲۵] مقایسه کرده ایم.

همانطور که در شکل ۲.۵ میبینید جایگزین های خط لوله ای شامل یک سری مراحل است ابتدا، یک مترجم پرسش کاربر را از فارسی به انگلیسی تبدیل می‌کند، سپس این پرسش انگلیسی به یک مدل زبانی پزشکی انگلیسی داده می‌شود، و در نهایت، پاسخ تولیدشده توسط مدل انگلیسی دوباره از انگلیسی به فارسی ترجمه می‌شود.



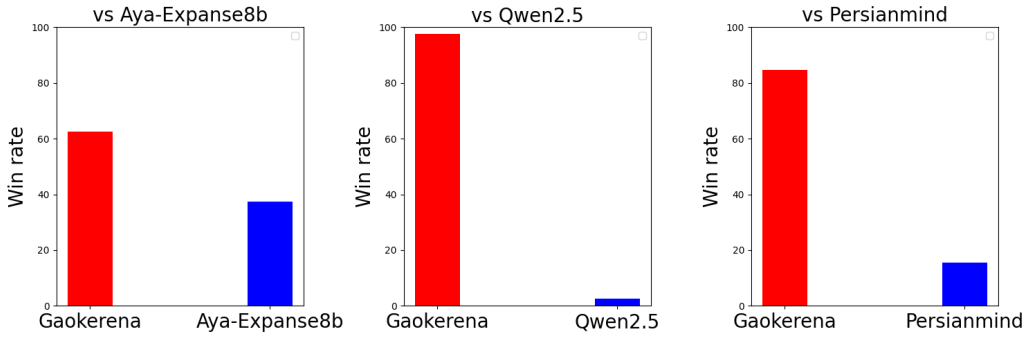
شکل ۲.۵: مکانیسم جایگزین خط لوله ای

۱.۶.۵ مقایسه با مدل های زبانی فارسی همه منظوره

همان‌طور که در جدول ۱.۵ مشاهده می‌کنید، مدل گائوکرنا-V توانست با موفقیت کنکور علوم پایه پزشکی شهریور 1402 ، را پشت سر بگذارد [۲۶] و به اولین مدل زبان فارسی با کمتر از هشت میلیارد پارامتر تبدیل شد که این آزمون را با موفقیت پشت سر گذاشته است. علاوه بر این، مدل ما در قسمت پزشکی مجموعه داده [۲۷] نیز بهبودهایی را نشان داد و نه تنها میانگین نمرات بالاتری کسب کرد، بلکه در اکثر زیرشاخه‌ها عملکرد برجسته‌ای داشت و توانایی خود را در درک و تولید دانش پزشکی به زبان فارسی به نمایش گذاشت.

قابل توجه است که دلیل سرعت بالای پاسخگویی PersianMind ، همان‌طور که در جدول ۱.۵ نشان داده شده، این است که این مدل تمایل دارد پاسخ‌های بسیار کوتاه‌تری نسبت به مدل‌های دیگر تولید کند و توکن پایان پاسخ را زودتر ایجاد نماید. علاوه بر ارزیابی پرسش و پاسخ چندگزینه‌ای، ما از GPT-4o [۴۴] به عنوان داور برای پرسش و پاسخ آزاد نیز استفاده کردیم. قسمت تست مجموعه داده MF3QA به مدل زبان رقیب و مدل ما ارائه شد. همان‌طور که در شکل ۳.۵ نشان داده شده است، GPT-4o به طور عمده پاسخ‌های تولیدشده توسط مدل ما را نسبت به سه مدل زبان دیگر ترجیح داده است.

[۲۴] general purpose language models
[۲۵] pipeline alternatives
[۲۶] .
[۲۷] MMLU



شکل ۳.۵: نرخ پیروزی گائوکرنا-V در رقابت با بقیه مدل های زبانی فارسی همه منظوره

۲.۶.۵ مقایسه با جایگزین های خط لوله ای

همان‌طور که قبلا اشاره شد، یکی از گزینه‌های جایگزین برای توسعه یک مدل زبانی پزشکی فارسی، استفاده از جایگزین های خط لوله ای [۲۸] است. با این حال، یکی از مشکلات عمده این سیستم‌ها سرعت پایین آن‌ها است. این سیستم‌ها زمان استنتاج بالایی دارند، زیرا خروجی یک مدل باید به مدل دوم منتقل شود و سپس خروجی مدل دوم دوباره توسط مدل اول پردازش شود. این فرآیند تکراری به طور قابل توجهی کارایی سیستم را کاهش می‌دهد.

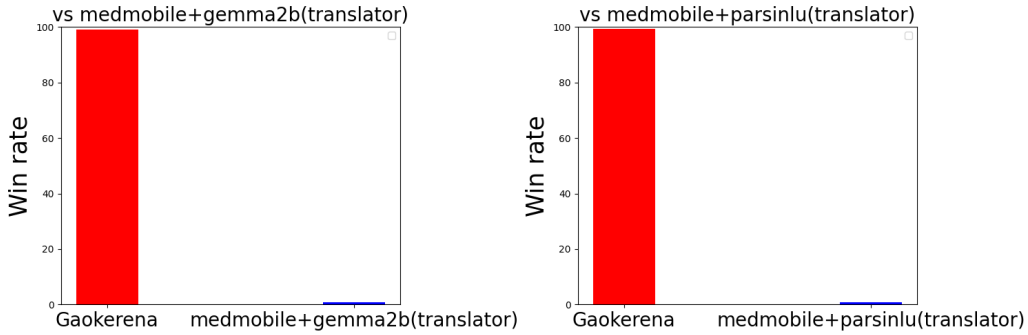
برای رفع مشکل سرعت پایین سیستم‌های مرحله‌ای، ما تمام پارامترها شامل پارامترهای مربوط به مترجم‌ها و مدل زبانی پزشکی را به طور همزمان بارگذاری کرده‌ایم. آزمایش‌های ما با مدل‌هایی مانند Medmobile همراه با gemma-2b-it به عنوان مترجم، و Medmobile همراه با مدل‌های parsinlu [۴۵] [۴۶] به عنوان مترجم انجام گرفته است؛ همان‌طور که در جدول ۲.۵ مشاهده میکنید جایگزین های خط لوله ای دقت و سرعت بسیار پایینی از خود نشان دادند.

یکی دیگر از مشکلات مهم جایگزین های خط لوله ای، عملکرد ضعیف آن‌ها در شناسایی و ترجمه دقیق اصطلاحات پزشکی است. این محدودیت چالشی جدی ایجاد می‌کند، زیرا دقت در استفاده از اصطلاحات تخصصی برای ارتباط موثر در محیط‌های مراقبت‌های بهداشتی حیاتی است. علت اصلی این ضعف احتمالاً به این دلیل است که مترجم‌های استفاده‌شده در این سیستم‌ها به طور خاص برای ترجمه پزشکی توسعه نیافته‌اند. برخلاف مدل‌های ترجمه عمومی، ترجمه پزشکی نیازمند درک دقیق واژگان تخصصی، زمینه و پیچیدگی‌های زبان پزشکی است.

در حال حاضر، هیچ مدلی برای ترجمه پزشکی به زبان فارسی طراحی نشده است، که این امر باعث می‌شود سیستم‌های موجود توانایی کافی برای مدیریت پیچیدگی‌های اصطلاحات پزشکی نداشته باشند. همان‌طور که در

[۲۸] pipeline alternatives

شکل ۴.۵ نشان داده شده است، این محدودیت‌ها منجر به نرخ پیروزی بسیار پایین جایگزین های خط لوله ای در رقابت با مدل گائوکرنا-V شده است.



شکل ۴.۵: نرخ پیروزی گائوکرنا-V در رقابت با جایگزین های خط لوله ای

PersianMind	**Qwen2.5**	**aya-expanse-8b** (baseline)	**Gaokerena-V** (ours)	
25.18	41.48	40.74	**48.14**	MMLU-anatomy(fa)
34.0	52.0	49.0	**53.0**	MMLU-medical genetics(fa)
20.23	43.35	**44.51**	43.93	MMLU-college medicine(fa)
25.28	47.92	52.07	**55.47**	MMLU-clinical knowledge(fa)
23.89	43.01	45.58	**47.05**	MMLU-professional medicine(fa)
32.63	44.85	45.14	**47.22**	MMLU-college biology(fa)
25.89	45.17	46.64	**49.31**	MMLU(avg)
19.64	33.33	34.52	**38.69**	IBMSEE Sept 2023
6.8b	7.6b	8b	8b	Number of parameters
حدود 2s	حدود 15s	حدود 8s	حدود 10s	inference time

جدول ۱.۵: مقایسه مدل گائوکرنا-V با بقیه مدل های زبانی فارسی همه منظوره

MedMobile parsinlu +	**MedMobile + gemma2 -2b-it**	**Gaokerena** (ours)	
25.18	14.07	**48.14**	MMLU-anatomy(fa)
35.0	20.0	**53.0**	MMLU-medical-genetics(fa)
27.17	19.08	**43.93**	MMLU-college-medicine(fa)
31.70	27.54	**55.47**	MMLU-clinical-knowledge(fa)
33.82	17.27	**47.05**	MMLU-professional-medicine(fa)
31.25	18.75	**47.22**	MMLU-college-biology(fa)
30.99	20.11	**49.31**	MMLU(avg)
32.73	24.40	**38.69**	IBMSEE Sept2023
3.8b+1.2b+1.2b	3.8b+2b	8b	Number of parameters
حدود 30s	حدود 20s	حدود 10s	inference time

جدول ۲.۵: مقایسه مدل گائوکرنا-V با جایگزین های خط لوله ای

فصل ۶

بررسی توانایی استدلال هوش مصنوعی

فصل ۷

معرفی مدل گائوکرنا-R

فصل ۸

نتیجه گیری

کتاب‌نامه

[۱] Brown, Peter F., et al. ”A statistical approach to machine translation.” (1990): 79-85.

[۲] Williams, Ronald J., and David Zipser. ”A learning algorithm for continually running fully recurrent neural networks.” Neural computation 1.2 (1989): 270-280.

[۳] Hochreiter, Sepp, and Jürgen Schmidhuber. ”Long short-term memory.” Neural computation 9.8 (1997): 1735-1780.

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

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

[۶] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.

[۷] Yenduri, Gokul, et al. ”Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions.” arXiv preprint arXiv:2305.10435 (2023).

[۸] Raffel, Colin, et al. ”Exploring the limits of transfer learning with a unified text-to-text transformer.” Journal of machine learning research 21.140 (2020): 1-67.

[۹] Luo, Man, et al. ”Choose your QA model wisely: A systematic study of generative and extractive readers for question answering.” arXiv preprint arXiv:2203.07522 (2022).

[۱۰] Hendrycks, Dan, et al. ”Measuring massive multitask language understanding.” arXiv preprint arXiv:2009.03300 (2020).

[۱۱] Liu, Yang, et al. ”G-eval: NLG evaluation using gpt-4 with better human alignment.” arXiv preprint arXiv:2303.16634 (2023).

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

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

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

[۱۵] Saab, Khaled, et al. ”Capabilities of gemini models in medicine.” arXiv preprint arXiv:2404.18416 (2024).

[۱۶] 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).

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

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

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

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

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

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

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

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

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

[۲۶] 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).

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

[۲۸] 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.

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

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

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

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

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

[۳۴] Yang, An, et al. “Qwen2 Technical Report.” arXiv Preprint arXiv:2407.10671, 2024.

[۳۵] Dang, John, et al. ”Aya expanse: Combining research breakthroughs for a new multilingual frontier.” arXiv preprint arXiv:2412.04261 (2024).

[۳۶] Team, Gemma, et al. ”Gemma 2: Improving open language models at a practical size, 2024.” URL https://arxiv. org/abs/2408.00118 1.3 (2024).

[۳۷] Rostami, Pedram, Ali Salemi, and Mohammad Javad Dousti. ”Persianmind: A cross-lingual persian-english large language model.” arXiv preprint arXiv:2401.06466 (2024).

[۳۸] Dash, Saurabh, et al. ”Aya Vision: Advancing the Frontier of Multilingual Multimodality.” arXiv preprint arXiv:2505.08751 (2025).

[۳۹] Shumailov, Ilia, et al. ”AI models collapse when trained on recursively generated data.” Nature 631.8022 (2024): 755-759.

[۴۰] Odumakinde, Ayomide, et al. ”Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress, 2024.” URL https://arxiv.org/abs/2408.14960.

[۴۱] Hu, Edward J., et al. ”Lora: Low-rank adaptation of large language models.” ICLR 1.2 (2022): 3.

[۴۲] Dao, Tri. ”Flashattention-2: Faster attention with better parallelism and work partitioning.” arXiv preprint arXiv:2307.08691 (2023).

[۴۳] Lacoste, Alexandre, et al. ”Quantifying the carbon emissions of machine learning.” arXiv preprint arXiv:1910.09700 (2019).

[۴۴] Hurst, Aaron, et al. ”Gpt-4o system card.” arXiv preprint arXiv:2410.21276 (2024).

[۴۵] Khashabi, Daniel, et al. ”Parsinlu: a suite of language understanding challenges for persian.” Transactions of the Association for Computational Linguistics 9 (2021): 1147-1162.

[۴۶] Kashefi, Omid. ”MIZAN: a large persian-english parallel corpus.” arXiv preprint arXiv:1801.02107 (2018).

Abstract

The use of artificial intelligence in answering medical questions is recognized as one of the emerging and critical fields in technology and healthcare, which has garnered widespread attention in recent years. This advanced technology, with its unique capabilities, can significantly enhance the quality of medical services provided to patients. Additionally, by accelerating the process of delivering medical information and providing quick and accurate responses to the questions of doctors and patients, it plays a vital role in reducing the workload of physicians. As such, artificial intelligence not only increases efficiency in healthcare systems but also improves the overall patient experience and paves the way for better and more effective treatments. On the other hand, since medicine is based on reasoning and logical analysis, developing a medical model designed on a chain of logical thoughts and reasoning can significantly enhance the accuracy and efficiency of such a model. This approach facilitates the execution of complex diagnostic and therapeutic processes in a more structured and purposeful manner. In this regard, every stage of diagnosis and treatment must be based on scientific evidence and reliable data. For instance, in the process of diagnosing diseases, doctors typically rely on medical history, clinical symptoms, and test results. By designing a logical model, these data can be interconnected in a logical chain, helping identify patterns and relationships between symptoms and diseases.

Keywords Artificial Intelligence in medicine,Persian language models,Medical language models,Natural language processing,Artificial Intelligence reasoning

University of Isfahan
Faculty of Computer Engineering

Developing a medical language model based on reasoning in Persian language

A Thesis submitted to the Graduate Studies Office
In partial fulfillment of the requirements for
The degree of Master of Science
in Computer Engineering - Artificial Intelligence and robotic

By:
Mehrdad Ghassabi

Supervisors:
First Supervisor and Second Supervisor

Advisor:
First Advisor

January 2026