


پیشنهاد رساله پایان نامه کارشناسی ارشد رشته هوش مصنوعی  
دانشکده مهندسی کامپیوتر

عنوان پژوهش:

1-فارسی:	توسعه یک مدل زبانی پزشکی مبتنی بر استدلال در زبان فارسی
2-انگلیسی:	Developing a medical language model based on reasoning in persian language

مشخصات دانشجو:

نام و نام خانوادگی	شماره دانشجویی	رشته و گرایش	امضا
مهرداد قصابی	4023614029	هوش مصنوعی	

مشخصات استادان راهنما و مشاور:

ردیف	نام و نام خانوادگی	دانشگاه و گروه آموزشی / سایر مؤسسات	تخصص	راهنما یا مشاور	امضا
	دکتر حمیدرضا برادران	دانشگاه اصفهان گروه هوش مصنوعی و رباتیک	گروه هوش مصنوعی و رباتیک	راهنما	

ثبت پیشنهاد در ایرانداک:

نوع ثبت	شماره نامه گواهی ثبت	تاریخ نامه گواهی ثبت
همانندجویی		
ثبت نهایی		

## هدفمندی سازی پایان نامه و رساله:

نوع هدفمندی (بر اساس شیوه نامه هدفمندی)	سازمان حمایت کننده یا عنوان هسته پژوهشی*

## شناسه اخلاق در پژوهش:

☐ نیاز دارد ☒ نیاز ندارد

## نوع تحقیق:

☐ 1- بنیادی ☒ 2- توسعه‌ای ☒ 3- کاربردی

## کلید واژه‌ها:

### فارسی ( انگلیسی )

1- پردازش زبان‌های طبیعی (natural language processing)

2 – مدل های زبانی پزشکی (medical language models)

3- زنجیره افکار (chain of thoughts)

4- مدل های زبانی کوچک (small language models)

## مساله پژوهش:

استفاده از هوش مصنوعی در پاسخگویی به سوالات پزشکی به عنوان یکی از زمینه‌های نوظهور و بسیار مهم در حوزه فناوری و بهداشت و درمان شناخته می‌شود که در سال‌های اخیر توجه زیادی را به خود جلب کرده است. این فناوری پیشرفته، با توانایی‌های منحصر به فرد خود، می‌تواند به طور قابل توجهی به بهبود کیفیت خدمات پزشکی ارائه شده به بیماران کمک کند، همچنین با تسریع در فرآیند ارائه اطلاعات پزشکی و فراهم آوردن پاسخ‌های سریع و دقیق به سوالات پزشکان و بیماران، نقش بسزایی در کاهش بار کاری پزشکان ایفا کند. بدین ترتیب، هوش مصنوعی نه تنها می‌تواند باعث افزایش کارایی در سیستم‌های بهداشتی شود، بلکه می‌تواند تجربه کلی بیماران را نیز بهبود بخشد و به ارائه درمان‌های بهتر و مؤثرتر کمک کند.

با توجه به این که پزشکی به عنوان یک علم مبتنی بر استدلال و تحلیل‌های منطقی شناخته می‌شود، توسعه یک مدل پزشکی که بر اساس زنجیره‌ای از افکار و استدلال‌های منطقی بنا شده باشد، می‌تواند به طور قابل توجهی دقت و کارایی این مدل را افزایش دهد. این رویکرد به ما این امکان را می‌دهد که فرآیندهای پیچیده تشخیصی و درمانی را به گونه‌ای ساختاریافته‌تر و هدفمندتر انجام دهیم. در واقع، هر مرحله از تشخیص و درمان باید بر اساس شواهد علمی و داده‌های معتبر قرار گیرد. به عنوان مثال، در تشخیص بیماری‌ها، پزشکان معمولاً از تاریخچه پزشکی، علائم بالینی و نتایج آزمایش‌ها استفاده می‌کنند. با ایجاد یک مدل منطقی، می‌توان این داده‌ها را به صورت یک زنجیره منطقی به هم متصل کرد که به شناسایی الگوها و روابط میان علائم و بیماری‌ها کمک می‌کند.

مدل‌های زبانی بزرگ به دلیل نیاز به منابع محاسباتی و حافظه بالا، قابلیت اجرای مستقیم بر روی دستگاه‌های کوچک خانگی را ندارند. بنابراین، این مدل‌ها باید حتماً بر روی سرورهای قدرتمند و دستگاه‌های بزرگ اجرا شوند و سپس نتایج حاصل از پردازش برای کاربران ارسال گردد. این موضوع به‌ویژه در زمینه پزشکی که حریم خصوصی اطلاعات بیماران از اهمیت و حساسیت بالایی برخوردار است، می‌تواند مشکلات عدیده‌ای را ایجاد کند. عدم امکان پردازش محلی داده‌ها ممکن است منجر به نگرانی‌های مربوط به امنیت اطلاعات و حفظ حریم خصوصی بیماران شود، چرا که انتقال داده‌های حساس به سرورهای خارجی می‌تواند در معرض خطرات امنیتی قرار گیرد.

بنابراین، در این زمینه، مدل‌های زبانی کوچک به عنوان ابزارهای کلیدی و مؤثر شناخته می‌شوند که می‌توانند به طور ویژه در حوزه پزشکی به کار گرفته شوند. از این رو، توسعه یک مدل زبانی پزشکی کوچک نه تنها از نظر حریم خصوصی حائز اهمیت است، بلکه می‌تواند به بهبود دسترسی به خدمات پزشکی، افزایش دقت تشخیص‌ها و حفظ حریم خصوصی بیماران نیز کمک شایانی نماید. این امر ضرورت توجه به طراحی و پیاده‌سازی چنین مدل‌هایی را در راستای ارتقا کیفیت خدمات بهداشتی و درمانی نشان می‌دهد.

## پیشینه پژوهش:

در دهه‌های گذشته، مدل‌های زبانی عمدتاً به استفاده از روش‌های آماری محدود می‌شدند، مانند مدل‌های n-gram، که به دلیل محدودیت‌های ذاتی خود در تحلیل و تولید زبان طبیعی، توانایی پاسخ‌گویی به پرسش‌های پیچیده و تخصصی کاربران در زمینه پزشکی را نداشتند. اما با ظهور معماری ترنسفورمر [1]، تحولی شگرف در حوزه مدل‌های زبانی رخ داد. این پیشرفت به هوش مصنوعی این امکان را داد که با بهره‌گیری از ساختارهای پیچیده‌تر و قابلیت‌های یادگیری عمیق، به‌طور مؤثر و دقیق‌تری به پرسش‌های کاربران پاسخ دهد و در نتیجه، توانایی ارائه پاسخ‌های مناسب و مرتبط در زمینه‌های تخصصی مانند پزشکی را پیدا کند.

## الف - انواع سیستم‌های پرسش و پاسخ پزشکی

سیستم‌های پرسش و پاسخ پزشکی را میتوان به دو گروه تقسیم‌بندی کرد. [2] در این قسمت توضیح کوتاهی در مورد این دو سیستم خواهیم داد.

### الف ۱- سیستم‌های پرسش و پاسخ استخراجی (extractive)

سیستم‌های پرسش و پاسخ استخراجی به منظور پاسخ‌گویی به سوالات کاربران، پاسخ کاربر را از طریق استخراج از یک متن منبع از پیش تأیید شده میدهد. این سیستم‌ها معمولاً از یک مدل زبانی مبتنی بر (encoder) استفاده میکنند

### الف ۲- سیستم‌های پرسش و پاسخ تولید کننده (generative)

با توسعه و گسترش هوش مصنوعی تولید کننده (generative) سیستم‌های پرسش و پاسخ‌هایی پدید آمدند که در آن برای تولید پاسخ به پرسش کاربر به دانش مدل زبانی تکیه میشود. از آنجایی که در این سیستم‌ها بایستی چیزی تولید شود بنابراین در آن‌ها از معمولاً از مدل‌های زبانی دارای (decoder) استفاده میشود.

## ب - شیوه سنجش کیفیت پاسخ‌های مدل‌های زبانی

سنجش صحت پاسخ‌های یک مدل زبانی به ویژه در زمینه پزشکی از اهمیت بالایی برخوردار است، زیرا این فرآیند به ارزیابی و تضمین کیفیت و دقت اطلاعاتی که مدل ارائه می‌دهد، کمک شایانی می‌کند. در ادامه چندین روش سنجش کیفیت پاسخ‌های مدل‌های زبانی را مطرح خواهیم کرد و مزایای و معایب هر یک را بررسی خواهیم کرد.

### ب ۱- معیار امتیاز برت

در این معیار، کیفیت پاسخ‌های تولید شده توسط یک مدل زبانی (BERT) [3] ارزیابی می‌شود. در این فرآیند، پاسخ تولید شده با پاسخ صحیح موجود در مجموعه داده، که به عنوان “ground truth” شناخته می‌شود، مقایسه می‌گردد.

مدل BERT به دلیل توانایی بالای خود در درک زبان طبیعی، یک امتیاز عددی برای میزان شباهت بین این دو پاسخ تولید می‌کند. این امتیاز نشان‌دهنده دقت و کیفیت پاسخ تولید شده است. 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. این روش ارزیابی به توسعه‌دهندگان این امکان را می‌دهد که با بهبود مستمر مدل‌های زبانی، به کیفیت بالاتری در تولید پاسخ‌های منطقی و مرتبط دست یابند و در نتیجه، تجربه کاربری و اعتماد کاربران به سیستم‌های هوش مصنوعی را افزایش دهند.

## ب-۲. معیار پاسخگویی به پرسش‌های چهار گزینه‌ای

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 [4] است. این مجموعه شامل پرسش‌های متنوعی است که در موضوعات مختلفی مانند علوم، پزشکی، ریاضیات، تاریخ، و ادبیات طراحی شده‌اند. MMLU به عنوان یک استاندارد در ارزیابی مدل‌های زبانی، به محققان و توسعه‌دهندگان این امکان را می‌دهد که عملکرد مدل‌های 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. به این ترتیب، ارزیابی عملکرد مدل‌ها با استفاده از مجموعه داده‌هایی مانند MMLU نه تنها به بهبود کیفیت و دقت مدل‌ها کمک می‌کند، بلکه به ارتقا دانش علمی در زمینه توسعه فناوری‌های هوش مصنوعی نیز می‌انجامد.

## ب-۳. سنجش کیفیت بر اساس استلزام زبان طبیعی (natural language entailment)

مجموعه داده K-QA شامل پاسخ‌های تولید شده توسط انسان است که به همراه توضیحات دقیقی دسته‌بندی شده و به عنوان “الزامی” یا “مفید” مشخص گشته‌اند. [5] این دسته‌بندی نشان می‌دهد که آیا این 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 (NLI) استفاده شوند. که در آن بایستی پاسخ مدل را به عنوان “مقدمه” و هر یک از توضیحات انسانی به عنوان “فرضیه” در نظر گرفته شود. سپس یک مدل زبانی بزرگ مانند GPT-4، قضاوت خواهد کرد که آیا مقدمه شامل یا متناقض با هر فرضیه است.

با انجام این کار روی همه رکورد های مجموعه داده دو امتیاز کامل بودن (completeness) و حقیقت داشتن (factuality) به صورت زیر بدست می آید.

$$S_{\text{comp}}(r_i, \mathcal{A}'_i) = \sum_{a \in \mathcal{A}'_i} \frac{1[r_i \text{ entails } a]}{|\mathcal{A}'_i|},$$

$$S_{\text{fact}}(r_i, \mathcal{A}_i) = \begin{cases} 0 & \text{if } \exists a \in \mathcal{A}_i \text{ such that } r_i \text{ contradicts } a \\ 1 & \text{otherwise,} \end{cases}$$

در اینجا  $r$  پاسخ مدل به سؤال است،  $A$  لیست تمامی توضیحات مربوط را شامل می شود،  $A'$  لیست توضیحات الزامی است و  $a$  یک توضیح خاص است. تابع شاخص  $[cond]$  1 در صورتی که شرط برقرار باشد، مقدار 1 را برمی گرداند و در غیر این صورت مقدار 0 را ارائه می دهد. نمرات کامل بودن و واقعیت به طور میانگین بر روی تمامی سوالات موجود در مجموعه داده محاسبه می شوند.

ب- ۴. سنجش توسط یک مدل زبانی دیگر

یک روش دیگر برای سنجش عملکرد یک مدل زبانی، استفاده از یک مدل زبانی دیگر به عنوان قاضی ارزیابی است. در این رویکرد، یک مدل زبانی مستقل به عنوان مرجع برای ارزیابی کیفیت پاسخ های تولید شده توسط مدل اصلی مورد استفاده قرار می گیرد. این روش به دلیل قابلیت های بالای مدل های زبانی در پردازش و درک زبان طبیعی، می تواند به طور مؤثری به ارزیابی دقت و کیفیت پاسخ ها کمک کند.

در این فرآیند، پاسخ های تولید شده توسط مدل اصلی به مدل قاضی ارائه می شود. مدل قاضی می تواند با استفاده از معیارهای مختلفی مانند شباهت معنایی، دقت اطلاعات، و سازگاری با زمینه، کیفیت پاسخ ها را ارزیابی کند. به عنوان مثال، مدل قاضی می تواند با بررسی تطابق پاسخ ها با اطلاعات موجود در متون معتبر یا داده های آموزشی، نمره ای برای هر پاسخ تولید کند.

### ج- پیشینه پژوهش در زبان های مختلف

در این قسمت به بررسی پیشینه پژوهش در این زمینه در دو زبان انگلیسی و فارسی خواهیم پرداخت.

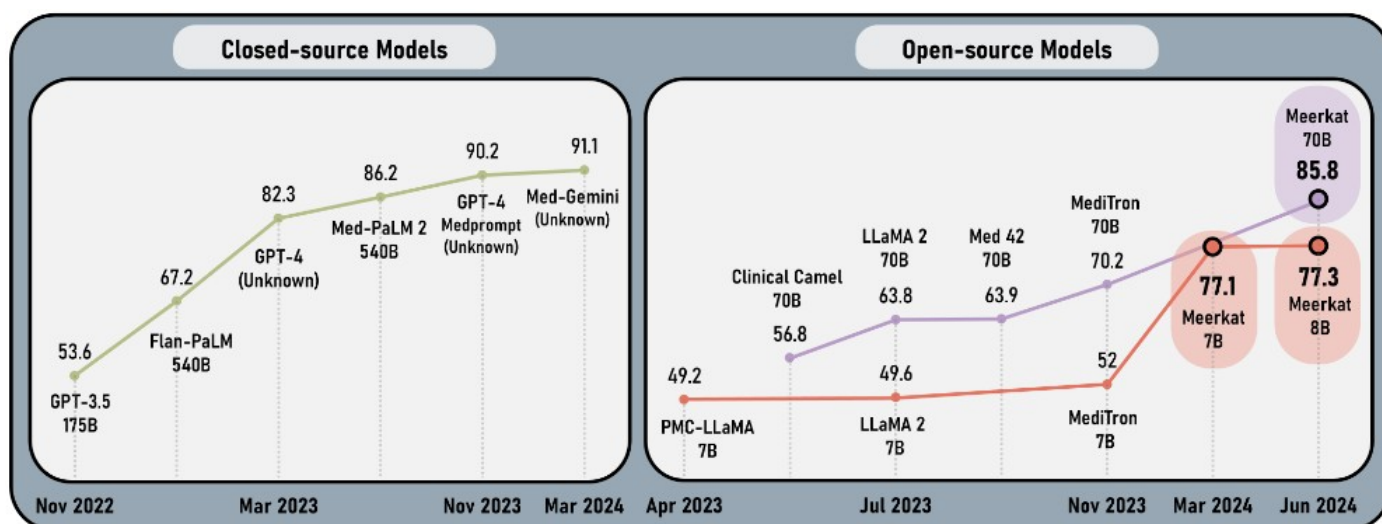
ج- ۱. زبان انگلیسی

در حوزه زبان انگلیسی، تحقیقات گسترده ای در زمینه مدل های زبانی بزرگ پزشکی صورت گرفته است. این تحقیقات به دلیل اهمیت بالای حوزه پزشکی و نیاز به پردازش و تحلیل داده های پیچیده و متنوع در این زمینه، به سرعت در حال گسترش است. مدل های زبانی بزرگ، به ویژه آن هایی که با داده های پزشکی آموزش دیده اند، می توانند به طور مؤثری در تسهیل فرآیندهای تشخیص، درمان و مدیریت بیماری ها کمک کنند.

یکی از این پژوهش ها، ارائه مدل Med-Gemini بوده است. [6] این مدل که به صورت متن بسته طراحی شده است، قابلیت منحصر به فردی دارد که به آن اجازه می دهد تا هم از تصاویر و هم از متن به عنوان ورودی استفاده

کند. این ویژگی به Med-Gemini این امکان را می‌دهد که اطلاعات را از منابع مختلف و به صورت چندرسانه‌ای تحلیل کند و به نتایج دقیق‌تری دست یابد. همانطور که در شکل ۱ مشاهده می‌شود، Med-Gemini توانسته است در مقایسه با سایر مدل‌های متن بسته، بهترین نتایج را به دست آورد. این موفقیت به دلیل استفاده از الگوریتم‌های پیشرفته یادگیری عمیق و معماری‌های بهینه‌سازی شده است که به مدل اجازه می‌دهد تا الگوها و روابط پیچیده‌ای را در داده‌های پزشکی شناسایی کند.

مدل Med-Gemini به ویژه در زمینه‌های مختلف پزشکی، از جمله تشخیص بیماری‌ها، تحلیل تصاویر پزشکی و پردازش متن‌های پزشکی، کاربردهای فراوانی دارد. به عنوان مثال، این مدل می‌تواند به تشخیص زودهنگام بیماری‌ها از طریق تحلیل تصاویر رادیولوژی یا MRI کمک کند و در عین حال، اطلاعات مربوط به تاریخچه پزشکی بیماران را از متون پزشکی استخراج کند.



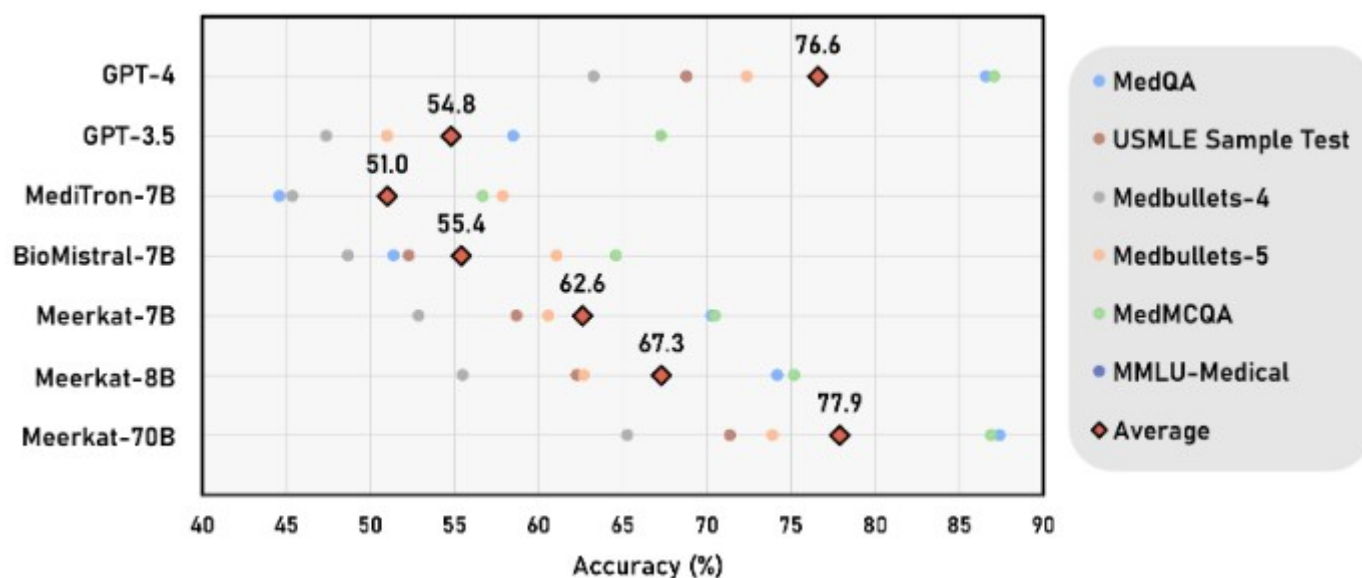
شکل ۱: نتایج مدل‌های زبانی متن بسته و متن باز انگلیسی روی مجموعه داده MedQA

اما همانطور که پیش‌تر اشاره شد، مدل‌های زبانی کوچک به دلیل حفظ حریم خصوصی و امکان اجرای آن‌ها بر روی دستگاه‌های محلی و کوچک، در حوزه پزشکی از اهمیت ویژه‌ای برخوردارند. این مدل‌ها به کاربران این امکان را می‌دهند که بدون نیاز به انتقال داده‌های حساس به سرورهای خارجی، اطلاعات پزشکی خود را پردازش و تحلیل کنند.

یکی از مدل‌های زبانی کوچک انگلیسی در حوزه پزشکی، مدل زبانی **Meerkat** است. [7] در این پژوهش، برای توسعه این مدل از یک مدل زبانی بزرگ بهره‌برداری شده است. این مدل بزرگ به منظور تولید داده‌های زنجیره افکار (Chain of Thought) مورد استفاده قرار گرفته و سپس یک مدل زبانی کوچک‌تر بر اساس این داده‌ها تمرین داده شده است.

این فرآیند نه تنها به بهبود دقت و کارایی مدل Meerkat کمک کرده، بلکه امکان تحلیل‌های عمیق‌تری از زبان و ساختارهای آن را نیز فراهم آورده است. با استفاده از داده‌های زنجیره افکار، مدل قادر است تا فرآیندهای منطقی و استدلالی را بهتر درک کند و به این ترتیب، پاسخ‌های دقیق‌تری به سوالات پزشکی ارائه دهد..

در شکل ۲، میزان پاسخگویی به سؤالات چهار گزینه‌ای و در شکل ۳، دو امتیاز کامل بودن (completeness) و حقیقت داشتن (factuality) این مدل را در مقایسه با مدل‌های دیگر مشاهده می‌کنید.



شکل ۲: نتایج مدل‌های زبانی مختلف روی چندین مجموعه داده [7]

Model	Size	Completeness	Factuality
GPT-4	Unknown	81.0	92.5
GPT-3.5	175B	71.4	92.0
ChatDoctor [34]	7B	63.0	89.1
Mistral-Instruct [18]	7B	62.4	88.1
Med-Alpaca [35]	13B	6.8	-
PMC-LLaMA [20]	13B	49.8	90.0
Meerkat (Ours)	7B	70.3	89.6
	8B	72.2	90.0
	70B	75.4	89.6

شکل ۳: نتایج مدل‌های زبانی مختلف روی مجموعه داده K-QA [7ع]



در این حوزه بر خلاف زبان انگلیسی در زبان فارسی پژوهش های انگشت شماری در این زمینه انجام شده است که در جملگی آنان سیستم نهایی یا دادگان به صورت عمومی منتشر نشده است.

یکی از این پژوهش ها پایان نامه کارشناسی ارشد خانم لیلا دارابی بوده است [8]؛ وی بر جمع آوری یک مجموعه داده برای پاسخگویی به سؤالات پزشکی به زبان فارسی تمرکز داشته است.

او در مورد سابقه پژوهش در مورد سیستم های پرسش و پاسخ پزشکی در زبان فارسی می گوید:

(در حوزه پاسخگویی به سؤالات پزشکی، در زبان فارسی تحقیقات به دلیل کمبود مستندات و وبسایت های فارسی محدود بوده است. با این حال، ویسی و همکاران یک سیستم پاسخگویی به سؤالات پزشکی به زبان فارسی با سه ماژول اصلی ارائه کردند [9]: پردازش پرسش، بازیابی مستندات و استخراج پاسخ. این سیستم با استفاده از ابزارهای پردازش زبان سفارشی و الگوریتم های تشخیص شباهت، توانست دقت ۸۳.۶ درصدی در پاسخ به سؤالات مربوط به بیماری ها و داروها کسب کند.

تقی زاده و همکاران مدل زبانی SINA-BERT را معرفی کردند [10] که بر پایه BERT ساخته شده است تا جای خالی یک مدل زبان قابل اعتماد به زبان فارسی در حوزه پزشکی را پر کند. این مدل پیش آموزش را بر روی مجموعه وسیعی از محتوای پزشکی از منابع رسمی و غیررسمی در اینترنت انجام داد تا عملکرد خود را در وظایف مرتبط با سلامت بهبود بخشد. SINA-BERT در وظایفی مانند دسته بندی سؤالات پزشکی، تحلیل احساسات پزشکی و بازیابی سؤالات پزشکی به کار گرفته شد و با معماری یکسان خود در این وظایف، عملکرد برتری نسبت به مدل های مبتنی بر BERT قبلی که در زبان فارسی موجود بودند، نشان داد.)

## د - استدلال در تولید پاسخ توسط مدل های زبانی

مدل های زبانی بزرگ به دلیل توانایی خود در تولید متن های طبیعی و پاسخ به سؤالات، در بسیاری از کاربردها از جمله مشاوره های هوش مصنوعی، تولید محتوا و حتی تحلیل داده ها به کار می روند. اما یکی از چالش های مهم در این زمینه، توانایی این مدل ها در استدلال منطقی برای تولید پاسخ های معنادار و مرتبط است. در ادامه به بررسی توانایی دلیل آوری مدل های زبانی به ویژه در زمینه پزشکی خواهیم پرداخت.

### د ۱- استفاده از زنجیره افکار در تولید پاسخ توسط مدل های زبانی

زنجیره افکار (Chain of Thought یا COT) در تولید پاسخ ها توسط مدل های زبانی، بهبود دقت و صحت پاسخ ها را فراهم می آورد. این رویکرد به مدل ها کمک می کند تا مراحل استدلال را به وضوح دنبال کرده و تحلیل های عمیق تری از سؤالات پیچیده ارائه دهند.

به عنوان مثال، فرض کنید که شخصی از مدل می پرسد: «بهرام ده سیب دارد چهار تای آن را به شهرام میدهد، هفت سیب از دلارام میگیرد و دوتای آن را میخورد بهرام چند سیب دارد؟» با استفاده از زنجیره افکار، مدل می تواند مراحل زیر را دنبال کند:

یک: بهرام ده سیب دارد چهار تای آن را میخورد پس اکنون شش سیب دارد.

دو: بهرام شش سیب دارد هفت سیب از دلارام میگیرد پس اکنون سیزده سیب دارد.

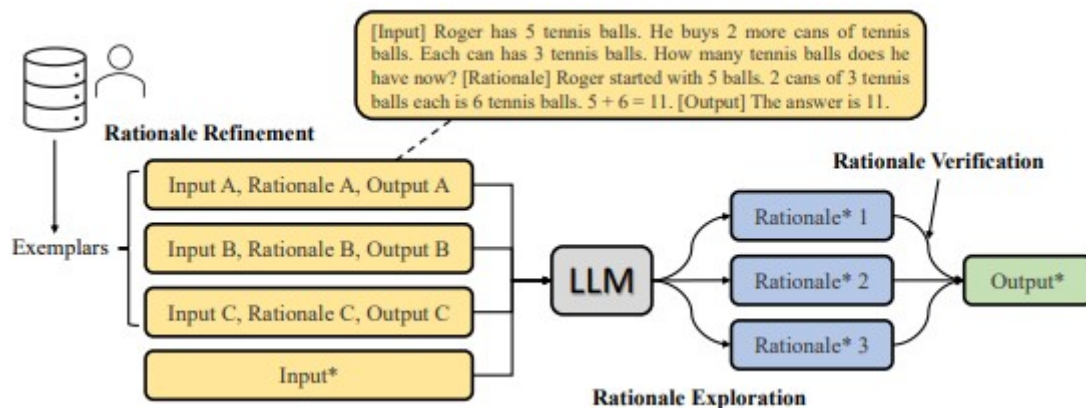
سه: بهرام سیزده سیب دارد دوتای آن را میخورد پس اکنون یازده سیب دارد.

با دنبال کردن این زنجیره مدل میتواند جواب دهد که بهرام یازده سیب دارد ولی در صورتی که به صورت مستقیم این پرسش را از مدل زبانی میپرسیدیم ممکن بود پاسخ اشتباهی تولید کند.

بنابراین زنجیره افکار باعث افزایش شفافیت و اعتماد به مدلها می شود و به آنها این امکان را می دهد که به سؤالات چندمرحله ای پاسخ دهند و از بروز تعصبها و خطاهای منطقی جلوگیری کنند. به طور کلی، زنجیره افکار به بهبود کیفیت پاسخها و ایجاد سیستمهای هوشمندتر و قابل اعتمادتر کمک می کند.

اما استفاده از این زنجیره بدین معنا نیست که مدل واقعا استدلال کردن را یاد گرفته است Jason Wei در این مورد در مقاله خود میگوید [11]:

(ما تأکید می کنیم که هرچند زنجیره افکار فرآیندهای تفکر استدلال کنندگان انسانی را شبیه سازی می کند، این موضوع به این سؤال پاسخ نمی دهد که آیا شبکه عصبی واقعاً در حال استدلال است یا خیر.)



شکل ۴: شیوه استدلال توسط زنجیره افکار [12]

## د ۲- استفاده از زنجیره افکار در زمینه پزشکی

همانطور که پیشتر اشاره شد، مدل Meerkat برای آموزش از داده های زنجیره افکار تولید شده توسط GPT-4 استفاده می کند. این رویکرد به مدل Meerkat این امکان را می دهد که از توانایی های پیشرفته GPT-4 بهره برداری کند. GPT-4 به دلیل اندازه و ساختار پیچیده اش، قادر به پردازش و تحلیل حجم بالایی از اطلاعات است و می تواند به تولید استدلال های منطقی پزشکی با استفاده از کتب مرجع پردازد. [7]

همانطور که در شکل های ۲ و ۳ مشاهده می شود، این مدل پس از فرایند آموزش، توانسته است نتایج قابل توجهی را در پاسخ به پرسش های چند گزینه ای و همچنین پرسش های فرم آزاد کسب کند. به ویژه، مدل Meerkat با هفت میلیارد پارامتر، در زمینه امتیاز کامل بودن (completeness score) عملکردی بهتر از مدل GPT-3.5، که دارای صد و هفتاد و پنج میلیارد پارامتر است، نشان داده است. این نتایج به وضوح اهمیت فرآیند استدلال در مدل های زبانی را در حوزه پزشکی تأیید می کند و نشان می دهد که حتی مدل های کوچکتر نیز می توانند با استفاده از

روش‌های نوین یادگیری، به عملکردهای برتری دست یابند. این پیشرفت‌ها نه تنها به بهبود دقت و کیفیت پاسخ‌ها کمک می‌کند، بلکه قابلیت اطمینان و کاربرد این مدل‌ها در زمینه‌های پزشکی را نیز افزایش می‌دهد.

### د- ۳- بهسازی زنجیره افکار توسط یادگیری تقویتی

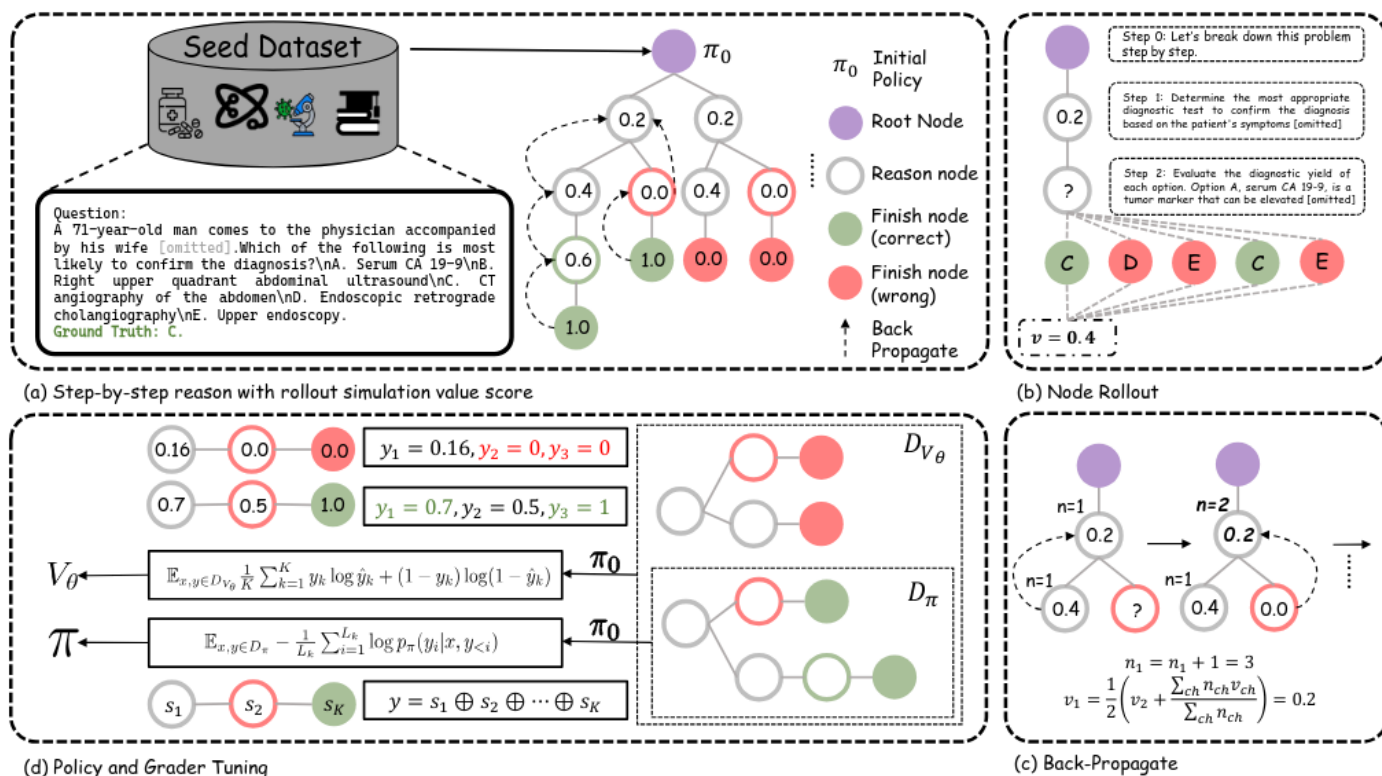
درست بودن زنجیره افکار تولید شده توسط یک مدل زبانی دارای اهمیت زیادی است، زیرا حتی مدل‌های زبانی بزرگ نیز ممکن است زنجیره‌های استدلال نادرستی تولید کنند و به نتایج اشتباهی برسند. این مسئله به ویژه در زمینه‌های حساس مانند پزشکی می‌تواند عواقب جدی و خطرناکی به همراه داشته باشد. به عنوان مثال، یک تشخیص نادرست یا پیشنهاد درمان نادرست می‌تواند به آسیب به بیماران منجر شود و اعتماد به سیستم‌های هوش مصنوعی را در محیط‌های پزشکی کاهش دهد. بنابراین، نیاز داریم تا به گونه‌ای صحت و درستی این افکار را تأیید کنیم و از ایجاد خطاهای جدی جلوگیری کنیم.

بدین منظور ما نیازمند مجموعه داده استدلال‌های صحیح تولید شده توسط انسان هستیم تا با استفاده از آن دادگان یک مدل یاداش را آموزش داده و بدین گونه بسنجیم که استدلال تولید شده توسط مدل زبانی تا چه حد درست است؛ اما این داده‌ها بسیار محدود بوده و جمع‌آوری آن با توجه به ماهیتش بسیار زمانبر می‌باشد بنابراین اگر شیوه‌ای بیابیم که این دادگان را به طور خودکار تولید کند گام بزرگی در صرفه جویی در زمان برداشته ایم.

در این راستا، جیانگ و همکاران مدل زبان پزشکی جدیدی به نام MedS3 را معرفی کردند [13] که با هدف بهبود استدلال توسط مدل زبانی در حوزه پزشکی طراحی شده است. این مدل از یادگیری تقویتی و جستجوی درخت مونت کارلو (MCTS) بهره می‌برد تا صرفاً با داشتن پرسش‌های چهار گزینه‌ای زنجیره‌های استدلال قابل تأیید تولید کند. استفاده از یادگیری تقویتی به این معناست که مدل می‌تواند از تجربیات گذشته خود یاد بگیرد و به تدریج بهبود یابد، در حالی که MCTS به آن اجازه می‌دهد تا به طور مؤثری زنجیره‌های استدلال را بررسی و ارزیابی کند.

همانطور که در شکل ۴ می‌بینید مدل MedS3 از یک پارادایم خودتکاملی استفاده می‌کند که به آن اجازه می‌دهد با یادگیری از تجربیات خود، به تدریج بهبود یابد. این مدل با یک مجموعه داده اولیه شامل حدود ۸۰۰۰ نمونه از پنج حوزه مختلف شروع می‌کند و از طریق روش جستجوی درخت مونت کارلو، زنجیره‌های استدلال قابل تأیید را ایجاد می‌کند. این رویکرد به مدل کمک می‌کند تا زنجیره‌های منطقی و معتبرتری تولید کند و در نتیجه، دقت و کیفیت پاسخ‌ها را افزایش دهد.

در مرحله استنتاج، مدل بازیگر چندین پاسخ تولید می‌کند و سپس مدل یاداش که توسط دادگان تولیدی تمرین داده شده است بهترین پاسخ را بر اساس نمره یاداش انتخاب می‌کند. این فرآیند به بهبود دقت و کیفیت پاسخ‌ها کمک می‌کند و از تولید زنجیره‌های نادرست جلوگیری می‌کند. به این ترتیب، MedS3 صرفاً با داشتن پرسش‌های چهارگزینه‌ای می‌تواند استدلال‌های صحیح تولید کرده و پاسخ‌های مناسبی به پرسش‌های پزشکی کاربران دهد.



شکل ۴ استفاده مدل medS3 از جستجو درخت مونته کارلو [13]

## اهداف پژوهش:

با توجه به گسترش هوش مصنوعی در حوزه پزشکی و نیاز به ابزارهای هوشمند، خلا وجود یک مدل زبانی فارسی که بتواند به پرسش‌های کاربران پاسخ‌های دقیق و معتبر ارائه دهد، به وضوح احساس می‌شود. این مدل باید قابلیت اجرا بر روی دستگاه‌های محلی را داشته باشد تا دسترسی و استفاده از آن آسان‌تر باشد. در این پایان‌نامه، تلاش می‌شود تا با استفاده از تکنیک‌های پیشرفته یادگیری ماشین و پردازش زبان طبیعی، این خلا پر شود و مدلی توسعه یابد که به نیازهای خاص جامعه پزشکی و کاربران فارسی‌زبان پاسخ دهد و به بهبود کیفیت خدمات بهداشتی و درمانی کمک کند.

## فرضیه‌ها یا سوال‌های پژوهش:

- زنجیره افکار چه نقشی در فرآیند تولید متن پزشکی و پاسخ به سوالات پزشکی ایفا می‌کند؟
- چه روش‌هایی برای ارزیابی کیفیت و دقت مدل زبانی پزشکی در زبان فارسی می‌توان استفاده کرد؟
- چگونه می‌توان داده‌های آموزشی پرسش و پاسخ تولید شده توسط یک مدل زبانی بزرگ را بهبود داد؟
- تأثیر استفاده از زنجیره افکار بر کیفیت نتایج تولید شده توسط مدل زبانی چگونه است؟

## روش تحقیق:

در این پژوهش قرار است یک مدل زبانی کوچک در حوزه پزشکی معرفی شود که از زنجیره افکار برای ارائه پاسخ‌های خود استفاده میکند که مراحل آن به صورت زیر است.

- جمع‌آوری مجموعه داده
- از آنجایی که تا به امروز در حوزه پزشکی به زبان فارسی هیچ مجموعه داده‌ای وجود ندارد، این کمبود به عنوان یک مانع جدی در توسعه مدل‌های زبانی مؤثر در این حوزه شناخته می‌شود. بنابراین، در مرحله اول، به گردآوری داده‌های پزشکی به زبان فارسی خواهیم پرداخت تا یک پایگاه داده جامع و معتبر ایجاد کنیم. این داده‌ها شامل مقالات علمی، گزارش‌های بالینی و پرسش و پاسخ‌های پزشکی خواهد بود و می‌تواند به آموزش و بهبود عملکرد مدل‌های زبانی کمک کند. این اقدام زمینه را برای پژوهش‌های آینده و توسعه ابزارهای هوش مصنوعی در حوزه پزشکی فارسی‌زبان فراهم خواهد کرد.

- طراحی و پیاده‌سازی

پس از شناسایی چالش‌های موجود در مرحله گسترش پیشینه تحقیق، در این مرحله ضروری است که یک معماری کلی برای توسعه مدل زبانی پزشکی مبتنی بر زنجیره افکار طراحی شود تا به این چالش‌ها پاسخ دهد. این معماری باید شامل چارچوب‌های نظری و عملیاتی باشد که به ما کمک کند تا به طور مؤثر به

مسائل شناسایی شده در زمینه پردازش زبان طبیعی و تحلیل داده‌های پزشکی پردازیم. همچنین، باید به نیازهای خاص حوزه پزشکی و ویژگی‌های زبانی متون پزشکی توجه شود تا بتوانیم راه‌حل‌های مناسبی ارائه دهیم. با طراحی یک ساختار منسجم و هدفمند که زنجیره افکار را در فرآیند تولید و تحلیل متن در نظر بگیرد، می‌توانیم به بهبود کیفیت مدل و دستیابی به نتایج معتبرتر و کاربردی‌تر در تشخیص بیماری‌ها و تولید محتواهای پزشکی کمک کنیم. این رویکرد نه تنها به افزایش دقت و کارایی مدل زبانی کمک می‌کند، بلکه می‌تواند به ارتقاء تجربه پزشکان و بیماران نیز منجر شود.

• نگارش پایان نامه

در نهایت پس از پیاده‌سازی مدل نتایج پژوهش در پایان نامه نگاشته میشود.

### جدول زمانی مراحل اجرا:

زمان مورد نیاز بر حسب ماه												مراحل پژوهش
۱۲	۱۱	۱۰	۹	۸	۷	۶	۵	۴	۳	۲	۱	
								■	■	■	■	بررسی پیشینه پژوهش
								■	■	■		جمع آوری دادگان پزشکی
					■	■	■	■	■			تولید دادگان پزشکی توسط مدل زبانی
					■	■	■	■	■			پیاده سازی مدل زبانی پزشکی
		■	■	■	■	■						نگارش پایان نامه و مقاله

[1]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30

[2]

Luo, M., Hashimoto, K., Yavuz, S., Liu, Z., Baral, C. and Zhou, Y., 2022. Choose your qa model wisely: A systematic study of generative and extractive readers for question answering. arXiv preprint .arXiv:2203.07522

[3]

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In 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) (pp. 4171-4186)

[4]

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J., 2020. Measuring .massive multitask language understanding. arXiv preprint arXiv:2009.03300

[5]

Manes, I., Ronn, N., Cohen, D., Ber, R.I., Horowitz-Kugler, Z. and Stanovsky, G., 2024. K-qa: A real-world .medical q&a benchmark. arXiv preprint arXiv:2401.14493

[6]

Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, .E. and Chaves, J.Z., 2024. Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416

[7]

Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., Yoon, C., Sohn, J., Choi, D. and Kang, J., 2024. Small language models learn enhanced reasoning skills from medical textbooks. arXiv preprint .arXiv:2404.00376

[8]

Darabi, L. (2023). Medical question answering for Persian (Master's thesis). Leiden Institute of Advanced Computer Science. theses.liacs.nl/pdf/2023-2024-DarabiLeila.pdf

[9]

Veisi, H. and Shandi, H.F., 2020. A Persian medical question answering system. International Journal on .Artificial Intelligence Tools, 29(06), p.2050019

[10]

Taghizadeh, N., Doostmohammadi, E., Seifossadat, E., Rabiee, H.R. and Tahaei, M.S., 2021. SINA-BERT: .a pre-trained language model for analysis of medical texts in Persian. arXiv preprint arXiv:2104.07613

[11]

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint .arXiv:2206.07682

[12]

Huang, J. and Chang, K.C.C., 2022. Towards reasoning in large language models: A survey. arXiv preprint .arXiv:2212.10403

[13]

Jiang, S., Liao, Y., Chen, Z., Zhang, Y., Wang, Y. and Wang, Y., 2025. MedS  $\hat{\text{S}}$ : Towards Medical .Small Language Models with Self-Evolved Slow Thinking. arXiv preprint arXiv:2501.12051





به نام خدا

## منشور اخلاق پژوهش

با استعانت از خدای سبحان و با اعتقاد راسخ به اینکه عالم محضر خداست و او همواره ناظر بر اعمال ماست و به منظور انجام شایسته پژوهش‌های اصیل، تولید دانش جدید و بهسازی زندگانی بشر، ما دانشجویان و اعضای هیات علمی دانشگاه‌ها و پژوهشگاه‌های کشور:

تمام تلاش خود را برای کشف حقیقت و فقط حقیقت به کار خواهیم بست و از هر گونه جعل و تحریف در فعالیت‌های علمی پرهیز می‌کنیم.

حقوق پژوهشگران، پژوهیدگان (انسان، حیوان، گیاه و اشیا)، سازمان‌ها و سایر صاحبان حقوق را به رسمیت می‌شناسیم و در حفظ آن می‌کوشیم.

به مالکیت مادی و معنوی آثار پژوهشی ارج مینهیم، برای انجام پژوهشی اصیل اهتمام ورزیده از سرقت علمی و ارجاع نامناسب اجتناب می‌کنیم.

ضمن پایبندی به انصاف و اجتناب از هر گونه تبعیض و تعصب، در کلیه فعالیت‌های پژوهشی رهیافتی نقادانه اتخاذ خواهیم کرد.

ضمن امانت‌داری، از منابع و امکانات اقتصادی، انسانی و فنی موجود استفاده بهره‌ورانه خواهیم کرد.

از انتشار غیراخلاقی نتایج پژوهش نظیر انتشار موازی همپوشان و چندگانه (تکه‌ای) پرهیز می‌کنیم.

اصل محرمانه بودن و رازداری را محور تمام فعالیت‌های پژوهشی خود قرار می‌دهیم.

در همه فعالیت‌های پژوهشی به منافع ملی توجه کرده و برای تحقق آن میکوشیم.

خویش را ملزم به رعایت کلیه هنجارهای علمی رشته خود، قوانین و مقررات، سیاست‌های حرفه‌ای،

سازمانی، دولتی و راهبردهای ملی در همه مراحل پژوهش میدانیم.

رعایت اصول اخلاق در پژوهش را اقدامی فرهنگی میدانیم و به منظور بالندگی این فرهنگ، به ترویج و

اشاعه آن در جامعه اهتمام می‌ورزیم.

