به نام خدا

دانشگاه اصفهان
معاونت پژوهش و فناوری

پیشنهاده رساله پایان‌نامه 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 پیشنهاده در ایرانداک:

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

هدفمندسازی پایان‌نامه 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:

بنیادی **توسعه‌ای** **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 حتما 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 عدیده ای 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 ـ 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 ـ۱ ـ سیستم‌های پرسش و پاسخ استخراجی

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 معمولا از یک مدل زبانی مبتنی بر کدگذار استفاده میکنند

ب ـ۲ـ سیستم‌های پرسش و پاسخ تولید کننده

با توسعه و گسترش هوش مصنوعی تولید کننده 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 معمولا از مدل های زبانی دارای کدگشا استفاده میشود.

در این پایان نامه نیز هدف ایجاد یک سیستم پرسش و پاسخ تولید کننده به وسیله توسعه یک مدل زبانی پزشکی فارسی میباشد همچنین با توجه به این موضوع که در محاوره های پزشکی ممکن است از اصطلاحات انگلیسی استفاده شود مدل پایه بایستی خود یک مدل چند زبانه باشد.

ب ـ شیوه سنجش کیفیت پاسخ‌های مدل های 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 مدل های زبانی را مطرح خواهیم کرد و مزایای و معایب هر یک را بررسی خواهیم کرد.

ب ـ۱ـ معیار امتیاز برت

در این معیار، کیفیت پاسخ‌های تولید شده توسط یک مدل زبانی BERT 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 امتیاز ها برای تمامی پاسخ‌ها در مجموعه داده، این امتیازات جمع‌آوری و میانگین‌گیری می‌شوند. میانگین امتیاز ها 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 پرسش های چهار 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 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
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 گشته اند. 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 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 و حقیقت داشتن به صورت زیر بدست می آید.

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

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

در اینجا ri پاسخ مدل به سوال i ام است، Ai لیست تمامی توضیحات مربوط به سوال i ام را شامل می‌شود،
A′i لیست توضیحات الزامی است و a یک توضیح خاص است. تابع شاخص 1[cond] در صورتی که شرط برقرار باشد، مقدار 1 را برمی‌گرداند و در غیر این صورت مقدار 0 را ارائه می‌دهد. نمرات کامل بودن و واقعیت به طور میانگین بر روی تمامی سوالات موجود در مجموعه داده محاسبه می‌شوند.
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 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 مؤثری 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، دقت 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 کند.

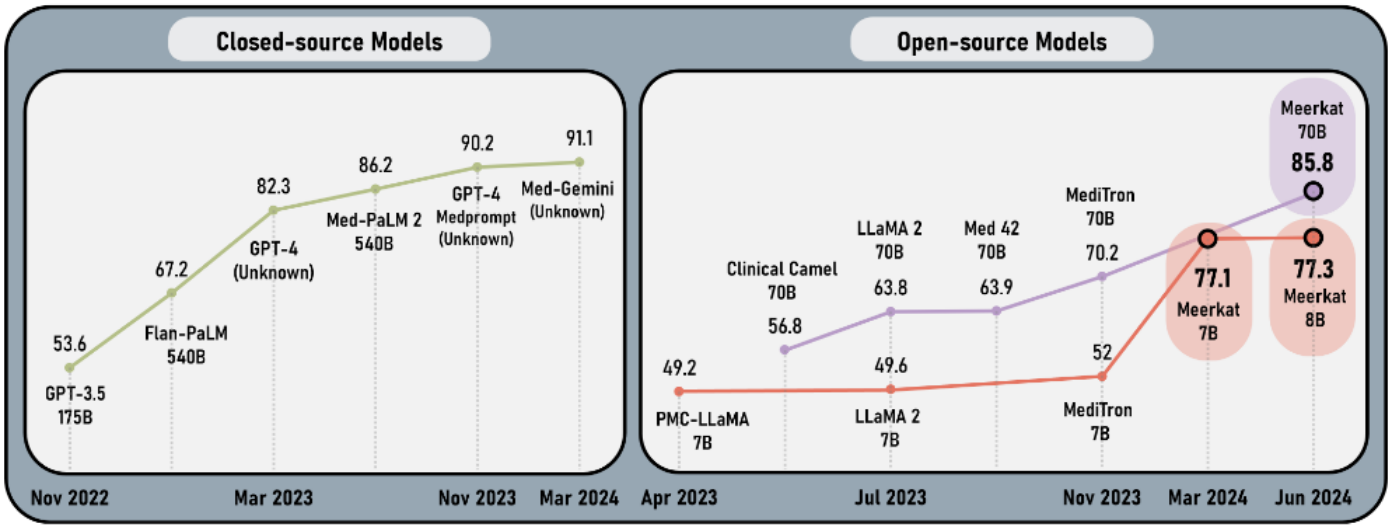
ج ـ مقایسه با جایگزین های خط لوله ای

یکی از گزینه های جایگزین برای توسعه یک مدل زبانی پزشکی فارسی، استفاده از جایگزین های خط لوله ای است. یعنی به جای توسعه یک مدل زبانی فارسی پزشکی از یک مدل زبانی انگلیسی پزشکی در کنار یک مدل مترجم استفاده گردد.

ج ـ۱ـ مشکلات این جایگزین های خط لوله ای
یکی از مشکلات عمده در سیستم‌های خط لوله‌ای هوش مصنوعی، 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 می‌تواند به شدت زمان‌بر باشد و به خصوص وقتی زمان بارگذاری و کنار گذاری پارامترها به آن اضافه شود، کارایی سیستم به طور قابل توجهی کاهش می‌یابد. در این شرایط، کاربر ممکن است با تأخیر قابل توجهی در دریافت نتایج مواجه شود که می‌تواند بر تصمیم‌گیری‌ها تأثیر منفی بگذارد و در مواقع بحرانی، به از دست رفتن فرصت‌ها یا منابع منجر شود.
یکی دیگر از مشکلات عمده در این نوع سیستم‌ها، انتشار خطا است. در فرآیندهای چند مرحله‌ای، خطاهای کوچک در خروجی یک مدل می‌توانند به سرعت در مراحل بعدی گسترش یابند و به نتایج نادرست یا غیرقابل اعتماد منجر شوند. به عنوان مثال، اگر مدل اول خروجی نادرستی تولید کند، این خروجی نادرست به مدل دوم منتقل می‌شود و این مدل نیز بر اساس داده‌های نادرست عمل خواهد کرد. در نتیجه، این خطاها می‌توانند به صورت تصاعدی افزایش یابند و منجر به انحرافات جدی در تصمیم‌گیری‌های نهایی شوند.

ج ـ۲ـ مقایسه عملکرد با مدل توسعه داده شده
در این پایان نامه، ما قصد داریم مدل زبانی ارائه شده خود را با این جایگزین های خط لوله ای نیز مقایسه کرده تا نقاط قوت و ضعف هر دو رویکرد را شناسایی کنیم و به درک بهتری از عملکرد و کارایی مدل خود دست یابیم. البته بایستی در نظر داشت که چه مدل زبانی انگلیسی پزشکی و چه مدل مترجم بایستی قابلیت اجرا روی دستگاه های خانگی قابل اجرا باشند تا بتوان مقایسه منصفانه ای بین این دو گزینه داشت

د ـ پیشینه پژوهش در زبان‌های مختلف
در این قسمت به بررسی پیشینه پژوهش در این زمینه در دو زبان انگلیسی و فارسی خواهیم پرداخت.
د ـ۱ ـ زبان انگلیسی
در حوزه زبان انگلیسی، تحقیقات گسترده‌ای در زمینه مدل‌های زبانی بزرگ پزشکی صورت گرفته است. این تحقیقات به دلیل اهمیت بالای حوزه پزشکی و نیاز به پردازش و تحلیل داده‌های پیچیده و متنوع در این زمینه، به سرعت در حال گسترش است. مدل‌های زبانی بزرگ، به ویژه آن‌هایی که با داده‌های پزشکی آموزش دیده‌اند، می‌توانند به طور مؤثری در تسهیل فرآیندهای تشخیص، درمان و مدیریت بیماری‌ها کمک کنند.
یکی از این پژوهش‌ها، ارائه مدل Med-Gemini بوده است. این مدل که به صورت متن بسته طراحی شده است، قابلیت منحصر به فردی دارد که به آن اجازه می‌دهد 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 MRIکمک کند و در عین حال، اطلاعات مربوط به تاریخچه پزشکی بیماران را از متون پزشکی استخراج کند.



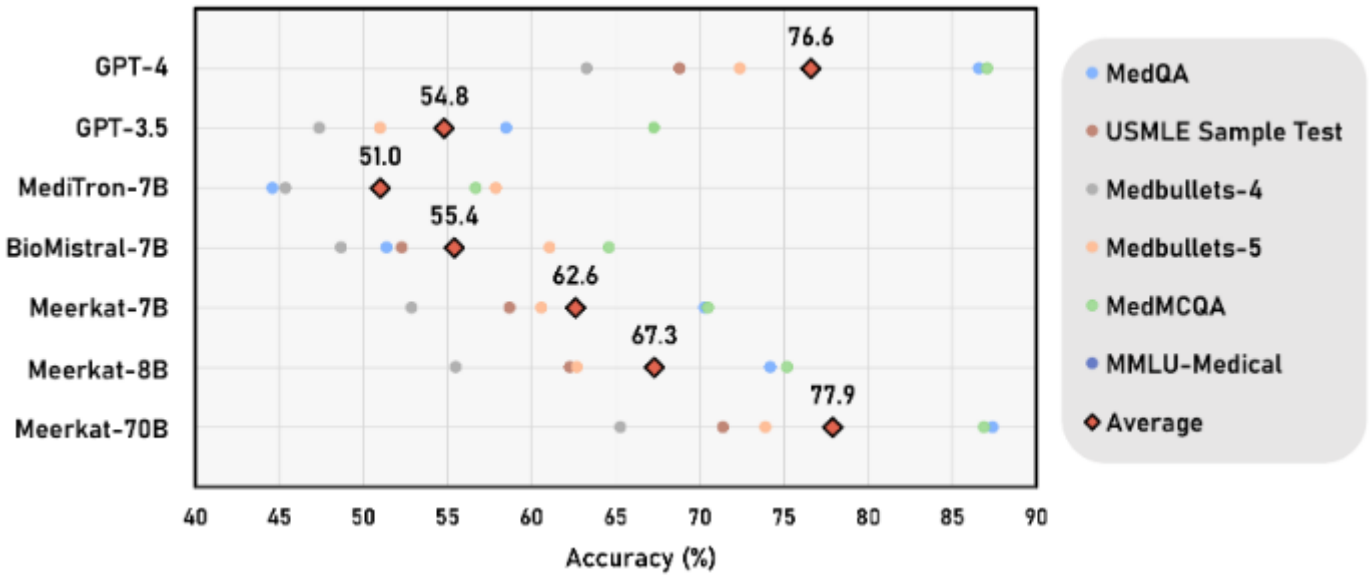
شکل ۱ : نتایج مدل های زبانی متن بسته و متن باز انگلیسی روی مجموعه داده MedQA

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. در این پژوهش، برای توسعه این مدل از یک مدل زبانی بزرگ بهره‌برداری شده است. این مدل بزرگ به منظور تولید داده‌های زنجیره افکار مورد استفاده قرار گرفته و سپس یک مدل زبانی کوچک‌تر بر اساس این داده‌ها تنظیم 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 و حقیقت داشتن این مدل را در مقایسه با مدل‌های دیگر مشاهده می‌کنید.



شکل ۲ : نتایج مدل های زبانی مختلف روی چندین مجموعه داده

Model	Size	Completeness	Factuality
GPT-4	Unknown	81.0	92.5
GPT-3.5	175B	71.4	92.0
ChatDoctor	7B	63.0	89.1
Mistral-Instruct	7B	62.4	88.1
Med-Alpaca	13B	6.8	-
PMC-LLaMA	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 مدل های زبانی مختلف روی مجموعه داده K-QA

د ـ۲ ـ زبان فارسی

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

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

در حوزه پاسخگویی به سؤالات پزشکی، در زبان فارسی تحقیقات به دلیل کمبود مستندات و وب‌سایت‌های 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: پردازش پرسش، بازیابی مستندات و استخراج پاسخ. این سیستم با استفاده از ابزارهای پردازش زبان سفارشی و الگوریتم‌های تشخیص شباهت، توانست دقت ۸۳.۶ درصدی در پاسخ به سؤالات مربوط به بیماری‌ها و داروها کسب کند.

تقی زاده و همکاران مدل زبانی SINA-BERT را معرفی کردند 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 پیش‌آموزش 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 دسته‌بندی 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 مدل‌های مبتنی بر برت قبلی که در زبان فارسی موجود بودند، نشان داد.

ه ـ اهمیت زنجیره افکار در تولید پاسخ

زنجیره افکار در تولید پاسخ‌های پزشکی در مدل‌های زبانی، بهبود دقت و صحت پاسخ‌ها را امکان‌پذیر می‌سازد. این رویکرد به مدل‌ها کمک می‌کند تا مراحل استدلال را به وضوح دنبال کرده و تحلیل‌های عمیق‌تری از سؤالات پیچیده پزشکی ارائه دهند. با استفاده از زنجیره افکار، مدل‌ها می‌توانند به عنوان ابزارهای آموزشی مؤثر عمل کنند و به دانشجویان پزشکی در یادگیری فرآیندهای استدلالی کمک کنند. همچنین، این رویکرد شفافیت را افزایش می‌دهد و اعتماد به مدل‌ها را تقویت می‌کند. زنجیره افکار به مدل‌ها این امکان را می‌دهد که به سؤالات چندمرحله‌ای پاسخ دهند و از بروز تعصب‌ها و خطاهای منطقی جلوگیری کنند. به طور کلی، زنجیره افکار به بهبود کیفیت پاسخ‌ها و ایجاد سیستم‌های پزشکی هوشمندتر و قابل اعتمادتر کمک می‌کند.

ه ـ۱ ـ تولید زنجیره افکار توسط یک مدل زبانی

همانطور که پیش‌تر اشاره شد، مدل میرکت برای آموزش از داده‌های زنجیره افکار تولید شده توسط GPT-4 استفاده می‌کند. این مدل به طور خاص طراحی شده تا از قدرت استدلال و تحلیل عمیق اطلاعات بهره‌برداری کند و به شناسایی الگوهای پیچیده و روابط بین داده‌ها بپردازد. با استفاده از زنجیره افکار، Meerkat می‌تواند در زمینه‌های مختلف، از جمله پزشکی، پاسخ‌های دقیق‌تری به سؤالات چندمرحله‌ای ارائه دهد و در تحلیل‌های پیچیده دقت بیشتری داشته باشد. این ویژگی به مدل کمک می‌کند تا به عنوان یک ابزار آموزشی عمل کند و فرآیندهای استدلالی را به کاربران آموزش دهد. در نهایت، استفاده از داده‌های زنجیره افکار به بهبود عملکرد Meerkat و توسعه هوش مصنوعی و یادگیری ماشین کمک می‌کند.

ه ـ۲ ـ بهسازی زنجیره افکار توسط یادگیری تقویتی

درست بودن زنجیره افکار تولید شده توسط یک مدل زبانی دارای اهمیت زیادی است، زیرا مدل‌های زبانی بزرگ 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

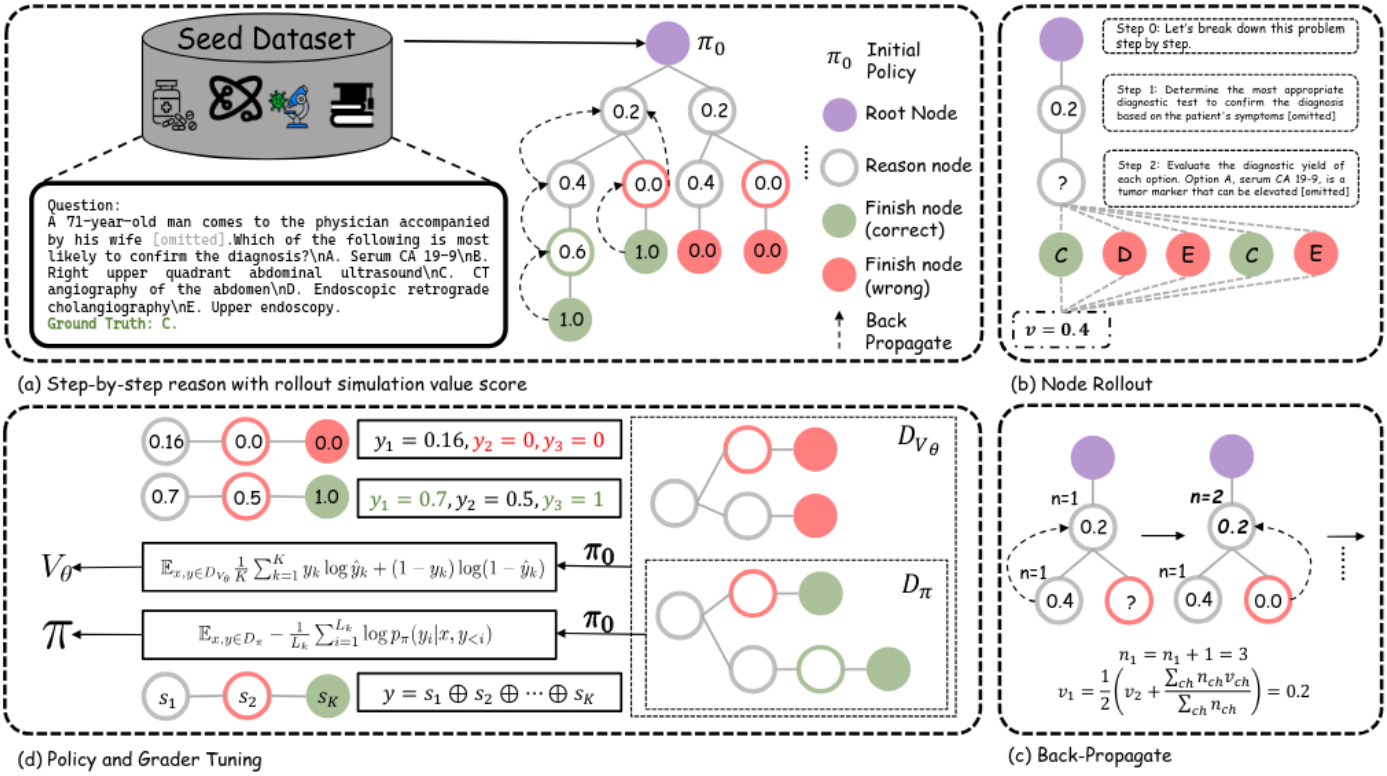
اعتماد به سیستم‌های هوش مصنوعی را در محیطهای پزشکی کاهش دهد. بنابراین، نیاز داریم تا به گونه‌ای صحت و درستی این افکار را تأیید کنیم و از ایجاد خطاهای جدی جلوگیری کنیم.

در این راستا، جیانگ و همکاران مدل زبان پزشکی جدیدی به نام MedS3 را معرفی کردند که با هدف بهبود پردازش زبان طبیعی در حوزه پزشکی طراحی شده است. این مدل از یادگیری تقویتی و جستجوی درخت مونت کارلو بهره می‌برد تا 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 MCTSبه آن اجازه می‌دهد تا به طور مؤثری زنجیره‌های استدلال را بررسی و ارزیابی کند.

مدل MedS3 به منظور رفع مشکلات مدل‌های قبلی که معمولاً به روش‌های پیش‌آموزش و تنظیم دقیق نظارت‌شده وابسته بودند، ایجاد شده است. این مشکلات شامل کارایی پایین داده و محدودیت‌های عملی در کاربردهای بالینی می‌شود. به عنوان مثال، مدل‌های قبلی ممکن است در شرایطی که داده‌های آموزشی محدود هستند، عملکرد مناسبی نداشته باشند و همچنین نگرانی‌های مربوط به حریم خصوصی داده‌ها و چالش‌های پیاده‌سازی در محیطهای بالینی، مانع از استفاده مؤثر از این مدل‌ها می‌شود. در واقع، این چالش‌ها می‌توانند به عدم اعتماد پزشکان و متخصصان به این مدل‌ها منجر شوند و در نتیجه، از پذیرش آن‌ها در محیطهای بالینی جلوگیری کنند.

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 سیاست چندین پاسخ تولید می‌کند و سپس مدل پاداش بهترین پاسخ را بر اساس نمره پاداش انتخاب می‌کند. این فرآیند به بهبود دقت و کیفیت پاسخ‌ها کمک می‌کند و از تولید زنجیره‌های نادرست جلوگیری می‌کند. به این ترتیب، MedS3 می‌تواند به عنوان یک ابزار مؤثر در پردازش زبان طبیعی پزشکی در محیطهای بالینی عمل کند و به بهبود تصمیم‌گیری‌های بالینی کمک کند .



شکل ۴ استفاده مدل meds3 از جستجو درخت مونته کارلو

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:

- جمع آوری پیکره متنی پزشکی و مجموعه داده پرسش و پاسخ پزشکی
- استفاده از دادگان زنجیره افکار به چه میزان در تولید پاسخ‌های صحیح پزشکی اهمیت دارد؟
- مدل های زبانی کوچک در حوزه پزشکی به چه میزان میتوانند مفید باشند؟

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.

- جمع آوری 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 پیاده سازی

پس از شناسایی چالش‌های موجود در مرحله گسترش پیشینه تحقیق، در این مرحله ضروری است که یک معماری کلی برای پژوهش 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 بپردازیم. همچنین، باید به نیازهای خاص حوزه تحقیق توجه شود تا بتوانیم راه‌حل‌های مناسبی ارائه دهیم. با طراحی یک ساختار منسجم و هدفمند، می‌توانیم به بهبود کیفیت پژوهش و دستیابی به نتایج معتبرتر و کاربردی‌تر کمک کنیم.

- نگارش پایان نامه

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

جدول زمانی پژوهش:

مراحل پژوهش / زمان مورد نیاز بر حسب ماه	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲
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., Park, J., Reykhart, O. and Fetherston, T., 2025. Small language models learn enhanced reasoning skills from medical textbooks. NPJ digital medicine, 8(1), p.240.

[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 $^ 3$: Towards Medical Small Language Models with Self-Evolved Slow Thinking. arXiv preprint arXiv:2501.12051.

وزارت علوم، تحقیقات و فناوری
معاونت پژوهش و فناوری

به نام خدا

منشور اخلاق پژوهش

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

تمام تلاش خود را براي کشف حقیقت و فقط حقیقت به کار خواهیم بست و از هر گونه جعل و تحریف در فعالیت‌هاي علمي پرهیز مي‌کنیم.
حقوق پژوهشگران، پژوهیدگان (انسان، حیوان، گیاه و اشیا)، سازمان‌ها و سایر صاحبان حقوق را به رسمیت مي‌شناسیم و در حفظ آن مي‌کوشیم.
به مالکیت مادي و معنوي آثار پژوهشی ارج مینهیم، براي انجام پژوهشی اصیل اهتمام ورزیده از سرقت علمي و ارجاع نامناسب اجتناب میکنیم.
ضمن پایبندي به انصاف و اجتناب از هر گونه تبعیض و تعصب، در کلیة 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 موازي همپوشان و چندگانه (تکه‌ای) پرهیز میکنیم.
اصل محرمانه بودن و رازداري را محور تمام فعالیت‌های پژوهشی خود قرار میدهیم.
در همه فعالیت‌های پژوهشی به منافع ملی توجه کرده و براي تحقق آن میکوشیم.
خویش را ملزم به رعایت کلیه هنجارهاي علمي 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 درجامعه اهتمام مي‌ورزیم.

امضاء دانشجو **امضاء استاد راهنما**