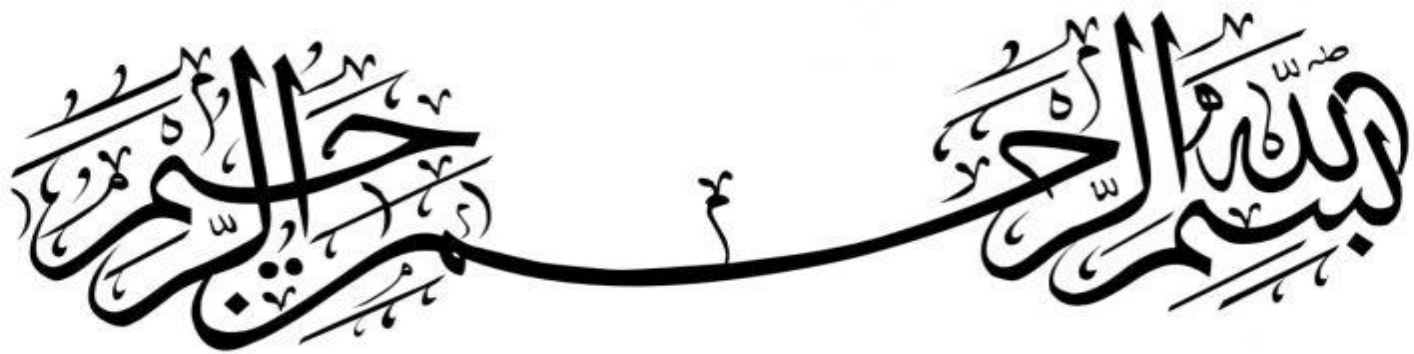

بسم الله الرحمن الرحيم

مرکز آموزش عالی علوم پزشکی وارستگان

Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study

بررسی نابرابری زبانی در عملکرد ChatGPT برای طبقه‌بندی بیماری‌های عروقی شبکیه: یک مطالعه مقطعی

مبینا فیاض

ژورنال کلاب

۱۴۰۴/۲/۲۹

مشخصات ژورنال

Indexing:
ISI, Scopus , PubMed, DOAJ

Name : Journal of Medical Internet Research

Impact Factor : 5.8

Category : Q1

مرکز آموزش عالی
علوم پزشکی
وارستگان

فهرست مطالب

اختصارات

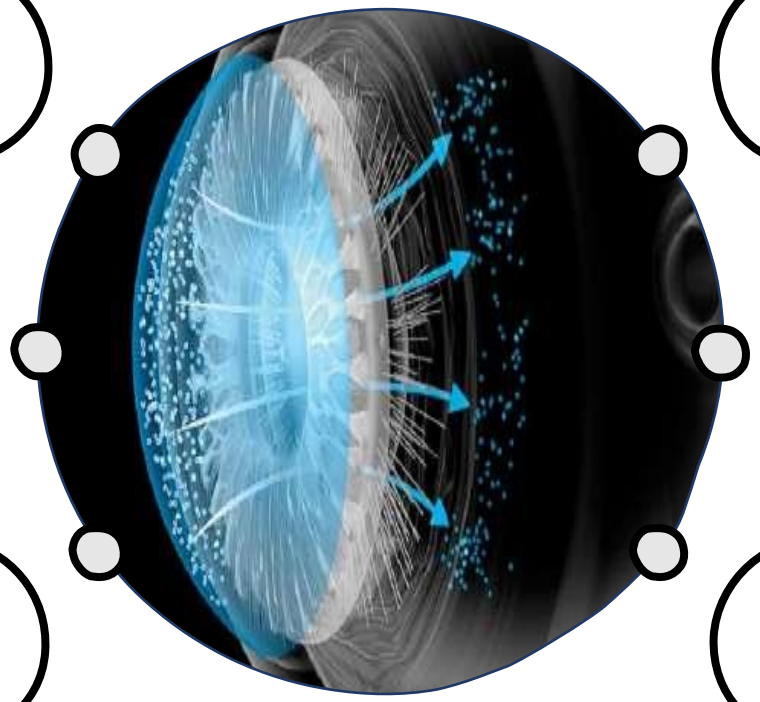
- FFA :Fundus Fluorescein Angiography…….. آنژیوگرافی فلورسئین فوندوس
- LLMs :Large Language Models…….…….. مدل های زبانی بزرگ
- API: Application Programming Interface…….. رابط برنامه نویسی کاربردی
- AI: Artificial Intelligence …….…….…… هوش مصنوعی

مرکز آموزش عالی
علوم پزشکی
وارستگان

۱

مقدمه و بیان مسئله

مقدمه و بیان مسئله



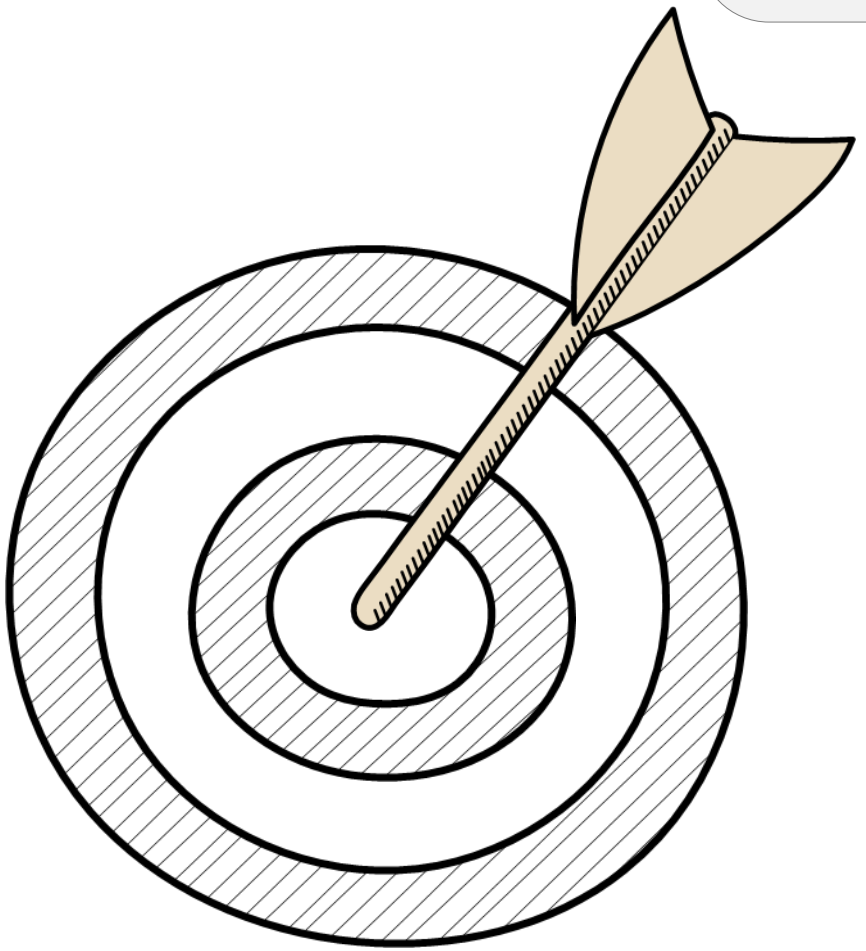
1. 2.2 میلیارد نفر در جهان دچار اختلالات بینایی هستند (آمار 2019)
2. بیماری‌های عروقی شبکیه (مثل رتینوپاتی دیابتی) از دلایل اصلی نابینایی‌اند
3. ابزار کلیدی تشخیص: آنژیوگرافی – FFA فلوروسئین شبکیه
4. تفسیر تصاویر FFA نیاز به تجربه پزشکی بالا دارد
5. مدل‌های زبانی بزرگ
6. (ChatGPT)

مرکز آموزش عالی
علوم پزشکی
وارستگان

۲

هدف مطالعه

هدف مطالعه



هدف اصلی:

ارزیابی عملکرد و استدلال ChatGPT در تشخیص بیماری‌های عروقی شبکیه از روی گزارش‌های FFA به زبان چینی

اهداف فرعی:

- **مقایسه دقت ChatGPT با پزشکان و کارورزان**
- **تحلیل تفاوت عملکرد در زبان چینی و انگلیسی**
- **ارزیابی تأثیر نوع درخواست**

مرکز آموزش عالی
علوم پزشکی
وارستگان

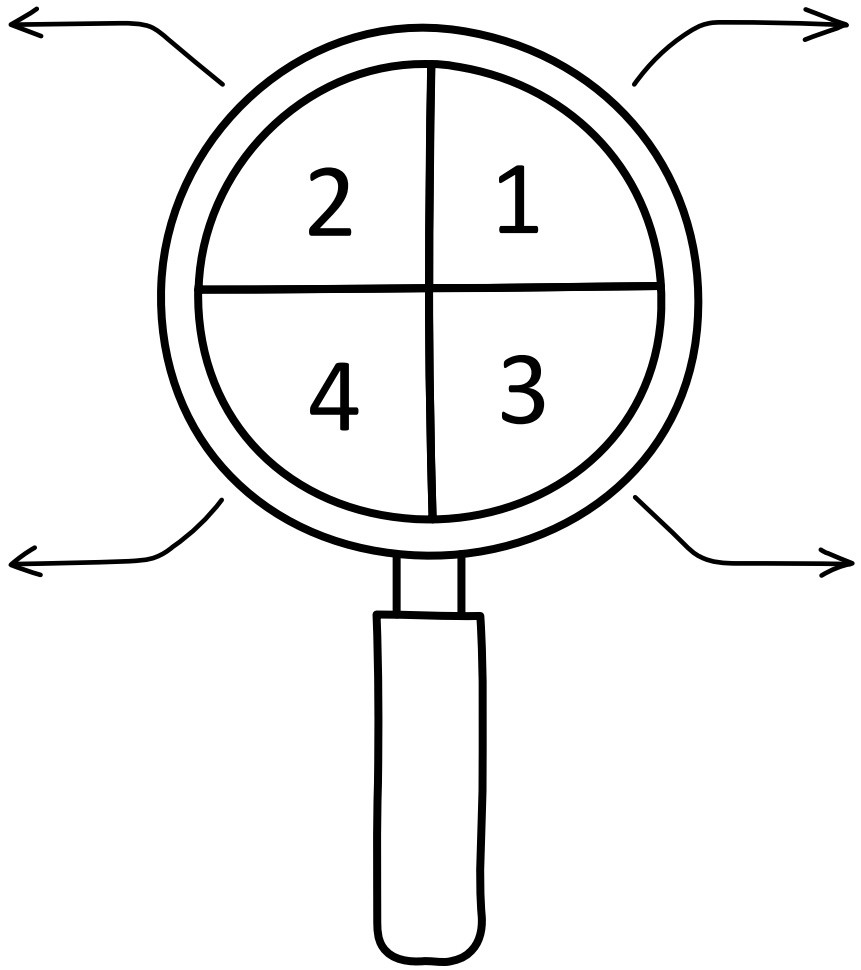
۳

روش اجرا

روش اجرا

داده‌گذاری و جمع‌آوری اطلاعات:
1226 گزارش FFA از 728 بیمار جمع آوری شد.

تشخیص بیماری‌های شبکیه با استفاده از ChatGPT
پرامپت مستقیم و استدلالی به دو زبان چینی و انگلیسی



ملاحظات اخلاقی
- تصویب اخلاقی
- حفظ حریم خصوصی

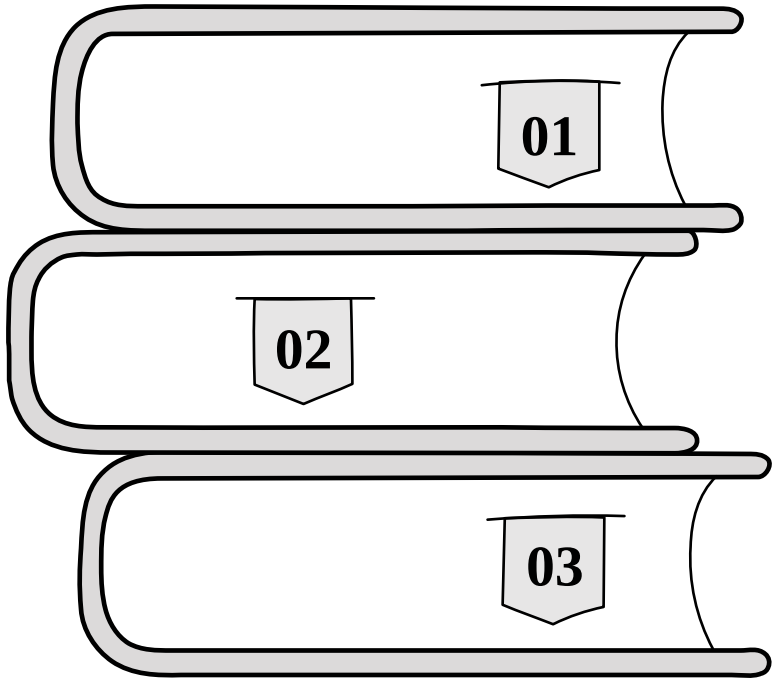
معیارهای ارزیابی
عملکرد تشخیصی
توانایی استنتاج
حذف اطلاعات
هالوکینیشن
اطلاعات نادرست
ناهماهنگی

منو

معیارهای ارزیابی

عملکرد تشخیصی

ارزیابی دقت تشخیصی ChatGPT از سه معیار دقت (Precision)، یادآوری (Recall)و امتیاز F1 استفاده شده است.



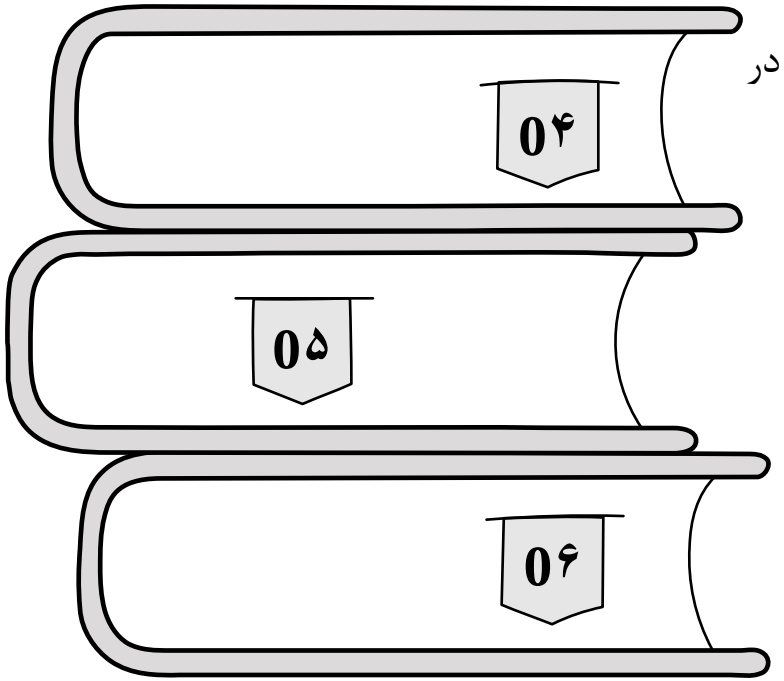
توانایی استنتاج

- تعداد مراحل استدلال
- تعداد اشتباهات استدلال
- نقص‌های موجود در فرایند استدلال

حذف اطلاعات

نادیده گرفتن یا جا انداختن نکات مهم موجود در گزارش اصلی.

معیارهای ارزیابی



هالوکینیشن
تولید اطلاعات یا نتایج پزشکی ساختگی که در گزارش اصلی وجود ندارند.

اطلاعات نادرست
ارائه دانش پزشکی اشتباه یا نقل قول ناصحیح از اطلاعات علمی.

ناهماهنگی
تضاد بین نتیجه‌گیری نهایی و مراحل استدلال ارائه‌شده توسط مدل.

مرکز آموزش عالی
علوم پزشکی
وارستگان

۴

یافته ها

یافته ها

Diagnostic Performance

Category	Direct-Chinese (%)			Direct-English (%)			Step-Chinese (%)			Step-English (%)		
	P[a]	R[b]	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Normal	100	85.47	92.17	100	88.03	93.64	98.39	52.14	68.16	97.37	94.87	96.1
DR[c]	91.55	72.52	80.93	91.05	95.12	93.04	85.07	95.4	89.94	82.13	93.58	87.48
wetAMD[d]	44.72	87.98	59.3	59.92	80.87	68.84	63.58	60.11	61.8	60	34.42	43.75
CSC[e]	4.35	2.74	3.36	33.33	1.37	2.63	34.15	19.18	24.56	50	6.85	12.05
BRVO[f]	41.32	79.37	54.35	63.33	90.47	74.51	83.61	80.95	82.26	67.95	84.13	75.18
CRVO[g]	93.1	71.05	80.6	84.85	73.68	78.87	41.27	68.42	51.49	58.33	73.68	65.12
VKH[h]	0	0	0	0	0	0	0	0	0	0	0	0
Overall	75.03	70.15	70.47	79.61	83.12	80.05	76.24	77.16	75.61	74.56	75.94	73.46

یافته ها

Inference Ability

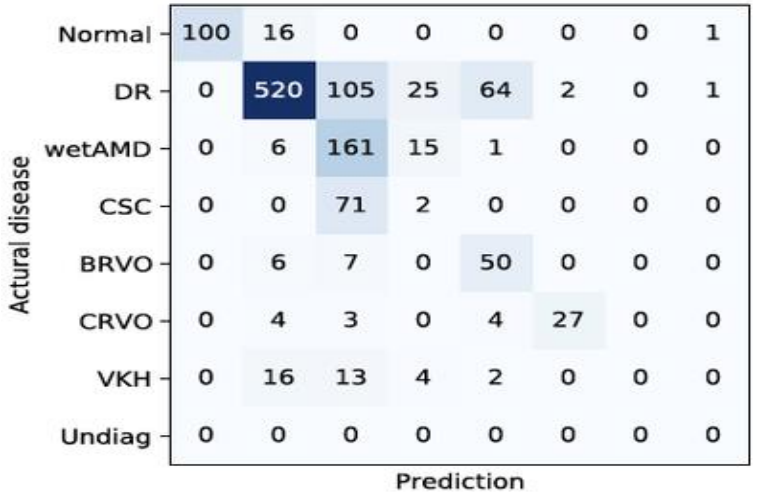
Measurement	Step-Chinese	Step-English	*P* value[a]
Reasoning steps per report, mean (SD)	1.4 (0.8)	2.6 (1.5)	<.001
Reasoning errors per report, mean (SD)	0.4 (0.5)	0.4 (0.6)	0.88
Incompleteness (%)	63.53	44.31	<.001
Omission of information (%)	0.78	0.39	0.68
Hallucinations (%)	5.88	0.59	<.001
Misinformation (%)	7.84	1.96	<.001
Inconsistency (%)	0.59	0.39	>.99

منو

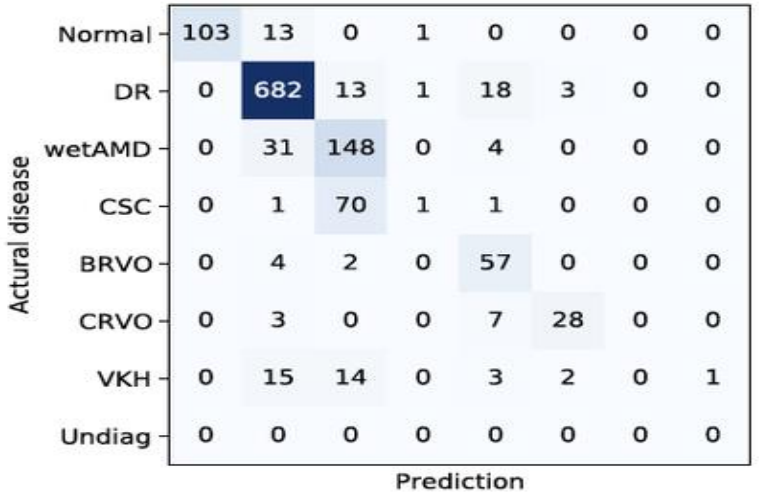
یافته ها

Confusion Matrix

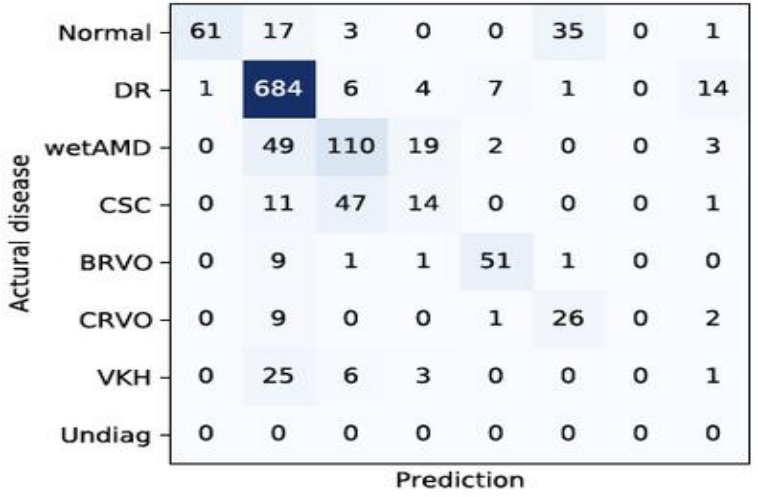
Direct-Chinese



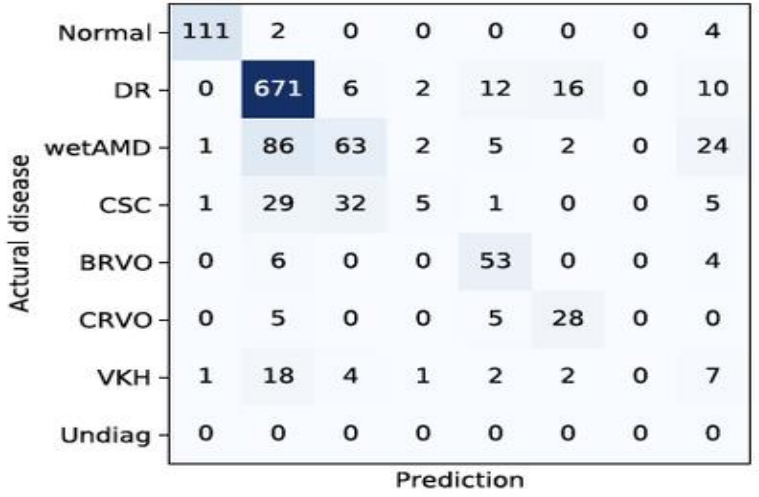
Direct-English



Step-Chinese



Step-English



مرکز آموزش عالی
علوم پزشکی
وارستگان

5

بحث و نتیجه گیری

نتایج

- ChatGPT ابزار کمکی مفید در تشخیص است،اما فقط در کنار نظارت پزشک و با دقت بالا قابل استفاده است.
- برای کاربرد گسترده‌تر، نیاز به آموزش با داده‌های واقعی، چندزبانه و بالینی دارد.

چالش‌ها :

- استدلال ناقص در زبان چینی می‌شود.
- تفاوت عملکرد بین زبان‌ها نشان‌دهنده‌ی نیاز به آموزش چندزبانه است.
- هنوز به سطح پزشکان متخصص نمی‌رسد.

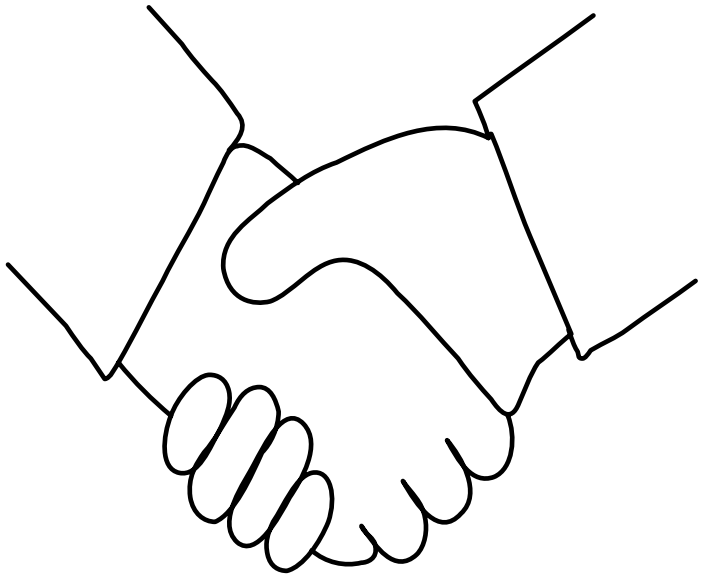
محدودیت ها

- دامنه زبانی و مکانی محدود
- عـــدم اســـتفاده از تصـــاویر پزشـــکی (FFA)
- نبود ارزیابی در محیط واقعی بالینی

مرکز آموزش عالی
علوم پزشکی
وارستگان

۶

دیدگاه من



- استفاده محدود به نسخه GPT-3.5
- نبود تحلیل چندوجهی(Multimodal)
- ضعف عملکرد در زبان غیرانگلیسی

مرکز آموزش عالی
علوم پزشکی
وارستگان

با تشکر از توجه شما