دانشنامه علمی چاوش طب آذین · chavoshteb.ir · به‌روزرسانی ۲۷ آبان ۱۴۰۵

مدل‌های زبانی بزرگ (LLM) در رادیولوژی دهان و فک: عملکرد، خطاها و چالش‌ها

مدل‌های زبانی بزرگ در مرور ۱۹ مطالعه رادیولوژی دهان و فک، دقت تشخیصی ۳۷٪ تا ۹۲٫۵٪ و دقت ۳۳٫۳٪ تا ۸۶٫۱٪ در پرسش‌های آزمون داشتند. چت‌بات‌های چندوجهی در پرسش‌های متنی بیش از ۸۰٪ اما در پرسش‌های تصویری کمتر از ۳۰٪ دقت داشتند و توهم، بیش از ۸۷٪ خطاها را در شناسایی ایمپلنت تشکیل داد.

نکات کلیدی این مقاله

- مرور سیستماتیک ۱۹ مطالعه، دقت تشخیصی مدل‌های زبانی بزرگ را ۳۷٪ تا ۹۲٫۵٪ و دقت تولید خودکار گزارش را ۷۰٫۴٪ تا ۸۱٫۳٪ گزارش کرد.[۱]
- در ۲۰۰ پرسش آزمون تخصصی رادیولوژی دهان و فک، ChatGPT-5.2 به دقت ۹۱٫۹٪ و ۹۵٫۷٪ در دو دوره رسید.[۴]
- چت‌بات‌های چندوجهی در پرسش‌های متنی بیش از ۸۰٪ و در پرسش‌های تصویری کمتر از ۳۰٪ دقت داشتند.[۲]
- در شناسایی ایمپلنت روی رادیوگرافی، دقت تشخیص برند ۴٫۱۷٪ تا ۳۲٫۷۸٪ بود و توهم بیش از ۸۷٪ خطاها را تشکیل داد.[۵]
- با سرنخ گمراه‌کننده مقتدرانه، دقت تشخیصی از ۵۳٫۵٪ تا ۶۵٫۰٪ به ۲۸٫۵٪ تا ۴۴٫۰٪ کاهش یافت.[۳]

مدل‌های زبانی بزرگ به‌سرعت در آموزش، پاسخ به پرسش‌های تشخیصی و تولید گزارش رادیولوژی دهان و فک آزموده می‌شوند. مطالعات اخیر نشان می‌دهند که عملکرد این مدل‌ها به نوع وظیفه، متنی یا تصویری بودن ورودی و حتی نحوه طرح پرسش وابسته است.[۱،۲،۳] این مقاله شواهد عملکرد، الگوی خطا و ملاحظات کاربرد ایمن را برای رادیولوژیست مرور می‌کند.

فهرست مطالب

1. مدل‌های زبانی بزرگ در رادیولوژی دهان و فک

2. عملکرد در وظایف متنی و آزمون‌ها
3. شکاف عملکرد متن و تصویر
4. توهم و چاپلوسی در پاسخ‌های تشخیصی
5. الزامات کاربرد ایمن مدل‌های زبانی بزرگ
6. جدول جمع‌بندی نکات
7. پرسش‌های متداول
8. جمع‌بندی
9. مطالعه مرتبط در دانشنامه
10. منابع

مدل‌های زبانی بزرگ در رادیولوژی دهان و فک

Liu و همکاران در مرور سیستماتیک، از ۱۶۲۱ عنوان اولیه ۱۹ مطالعه را وارد کردند که مدل‌های زبانی بزرگ را برای وظایف رادیولوژی دهان و فک توسعه داده، تنظیم کرده یا ارزیابی کرده بودند. این مطالعات سه حوزه را پوشش می‌دادند: تولید و پاسخ پرسش‌های آزمون و آموزشی (۸ مطالعه)، تشخیص و پیشنهاد درمان (۷ مطالعه) و تولید گزارش و ارتباط با بیمار (۴ مطالعه).[1]

خطر سوگیری با معیارهای اقتباس‌شده از راهنمای گزارش‌دهی TRIPOD-LLM و با تمرکز بر شفافیت منشأ داده، استحکام روش ارزیابی و اعتبار استاندارد مرجع سنجیده شد.[1]

عملکرد در وظایف متنی و آزمون‌ها

عملکرد مدل‌های زبانی بزرگ در وظایف رادیولوژی دهان و فک

وظیفه	تعداد مطالعات	بازه عملکرد
تشخیص بیماری‌های دندانی	۷ (تشخیص و درمان)	دقت ۳۷٪ تا ۹۲٫۵٪
تشخیص افتراقی و طرح درمان	۷ (تشخیص و درمان)	امتیاز متخصص ۳٫۶ تا ۴٫۷ از ۵

وظیفه	تعداد مطالعات	بازه عملکرد
پرسش‌های آزمون و بورد	۸ (آزمون و آموزش)	پاسخ صحیح ۳۳٫۳٪ تا ۸۶٫۱٪
تولید خودکار گزارش	۴ (گزارش و ارتباط)	دقت ۷۰٫۴٪ تا ۸۱٫۳٪

منبع: Liu و همکاران، Dentomaxillofac Radiol 2025.

Icen و Temur همه ۲۰۰ پرسش رادیولوژی دهان و فک آزمون تخصصی دندانپزشکی (DSE) از ۲۰۱۲ تا ۲۰۲۵ را به شش سامانه مبتنی بر مدل زبانی بزرگ دادند. ChatGPT-5.2 در دوره‌های اول و دوم به دقت ۹۱٫۹٪ و ۹۵٫۷٪ رسید و ChatGPT-4.0 با ۷۹٫۸٪ تا ۸۲٫۸٪ در رتبه بعد قرار گرفت. تفاوت میان مدل‌ها معنادار بود.[۴]

این نتایج به پرسش‌های چندگزینه‌ای و متنی مربوط است و نباید به تفسیر تصویر بالینی تعمیم داده شود؛ شواهد بخش بعد همین محدودیت را نشان می‌دهد.[۲]

شکاف عملکرد متن و تصویر

Jeong و همکاران ۹۰ پرسش متنی و تصویری رادیولوژی دهان را به ChatGPT-4o و Gemini 2.0 Flash دادند و دقت، ثبات پاسخ در ده تکرار و توهم را بررسی کردند. هر دو مدل در پرسش‌های متنی بیش از ۸۰٪ دقت داشتند، اما در وظایف تصویری دقت کمتر از ۳۰٪ بود، پاسخ‌ها تغییرپذیری زیادی داشتند و توهم با اطلاعات نادرست به‌کرات دیده شد.[۲]

> «این یافته‌ها نشان می‌دهد که چت‌بات‌های هوش مصنوعی هنوز برای استفاده قابل اعتماد در رادیولوژی دهان مناسب نیستند.»
>
> “These findings suggest that AI chatbots are not yet suitable for reliable use in oral radiology.”
>
> — Jeong و همکاران، Dentomaxillofac Radiol 2026[۲]

شناسایی ایمپلنت روی رادیوگرافی

Hlaing و همکاران ۱۲۰ تصویر استاندارد (۶۰ پری‌اپیکال و ۶۰ پانورامیک) از شش سیستم ایمپلنت را هر کدام شش بار به چهار مدل چندوجهی دادند. ویژگی‌های ساختاری (سطح ایمپلنت و نوع اتصال) با دقت ۵۱٫۶۷٪ تا ۸۲٫۵٪ شناسایی شد، اما تشخیص برند و سازنده ۴٫۱۷٪ تا ۳۲٫۷۸٪ باقی ماند. توهم، یعنی پاسخ قطعی نادرست، در همه مدل‌ها بیش از ۸۷٪ خطاها را تشکیل داد.[۵]

دانشنامه علمی چاوش طب آذین

بیشترین دقت گزارش‌شده LLMها در چهار نوع وظیفه

درصد پاسخ صحیح (بالاترین مقدار گزارش‌شده در هر مطالعه)



منبع: Icen و ۲۰۲۶ Temur؛ Geduk و همکاران ۲۰۲۶؛ Hlaing و همکاران ۲۰۲۶.

بالاترین دقت گزارش‌شده مدل‌های زبانی بزرگ در چهار نوع وظیفه؛ نمودار تحریریه بر پایه مطالعات Geduk، Icen و Hlaing (۲۰۲۶).

توهم و چاپلوسی در پاسخ‌های تشخیصی

Geduk و همکاران پنج مدل را در ۲۰۰ سناریوی تشخیصی (۱۰۰ تصویری و ۱۰۰ متنی) و سه شرایط پرسش، یعنی خنثی، سرنخ گمراه‌کننده و سرنخ مقتدرانه، آزمودند و ۳۰۰۰ پاسخ را تحلیل کردند. دقت تشخیصی در شرایط خنثی ۵۳٫۵٪ تا ۶۵٫۰٪ بود و در شرایط سرنخ مقتدرانه به ۲۸٫۵٪ تا ۴۴٫۰٪ کاهش یافت. احتمال چاپلوسی در سرنخ مقتدرانه ۲٫۹۲ برابر سرنخ گمراه‌کننده بود.[۳]

> «جدایی میان اطمینان خوداظهاری و دقت نشان می‌دهد که امتیاز اطمینان خوداظهاری به‌تنهایی شاخص قابل اعتمادی برای عملکرد تشخیصی نیست.»
>
> “The dissociation between self-reported confidence and accuracy indicates that self-reported confidence scores alone are unreliable indicators of diagnostic performance.”
>
> — Geduk و همکاران، Diagnostics 2026[۳]

الگوی خطای مدل‌های زبانی بزرگ در مطالعات اخیر

مطالعه	طراحی	یافته کلیدی
Jeong ۲۰۲۶	۹۰ پرسش متنی و تصویری	تصویری کمتر از ۳۰٪، توهم مکرر
Hlaing ۲۰۲۶	۱۲۰ رادیوگرافی، ۶ سیستم ایمپلنت	برند ۴٫۱۷٪ تا ۳۲٫۷۸٪؛ توهم بیش از ۸۷٪ خطاها
Geduk ۲۰۲۶	۲۰۰ سناریو، ۳۰۰۰ پاسخ	نسبت شانس چاپلوسی ۲٫۹۲ با سرنخ مقتدرانه

منبع: مطالعات فهرست‌شده در همین مقاله.

الزامات کاربرد ایمن مدل‌های زبانی بزرگ

جایگاه در گردش کار رادیولوژیست

Liu و همکاران نتیجه گرفتند که دقت، کامل بودن و ثبات مدل‌ها همچنان متغیر است و پیش از ادغام قابل اعتماد در گردش کار بالینی و آموزشی، توسعه، اعتبارسنجی و استانداردسازی بیشتری لازم است.[۱] Hlaing و همکاران نیز خروجی این مدل‌ها را تنها ابزار تکمیلی نیازمند تأیید بالینی دانستند.[۵]

نکات عملی

برداشت تحریریه از این شواهد:[۲، ۵، ۳]

- تفسیر تصویر را به چت‌بات عمومی نسپارید؛ دقت تصویری کمتر از ۳۰٪ گزارش شده است.[۲]
- پرسش را بدون پیش‌فرض تشخیصی طرح کنید؛ سرنخ مقتدرانه گمراه‌کننده دقت را به‌شدت کاهش داد.[۳]
- به اطمینان اعلام‌شده مدل تکیه نکنید؛ اطمینان در پاسخ‌های درست و نادرست تفاوت معنادار نداشت.[۳]
- پاسخ قطعی درباره برند ایمپلنت یا تشخیص نهایی را بدون تأیید مستقل نپذیرید.[۵]
- داده قابل شناسایی بیمار را در سامانه‌های عمومی وارد نکنید و مسئولیت گزارش نهایی با رادیولوژیست است (برداشت تحریریه).

جدول جمع‌بندی نکات

نکات کاربردی مدل‌های زبانی بزرگ در رادیولوژی دهان و فک

محور	نکته کلیدی
دقت تشخیصی	۳۷٪ تا ۹۲٫۵٪ در مرور سیستماتیک
پرسش متنی در برابر تصویری	بیش از ۸۰٪ در برابر کمتر از ۳۰٪
توهم	بیش از ۸۷٪ خطاها در شناسایی ایمپلنت
سرنخ گمراه‌کننده	افت دقت تا ۲۸٫۵٪ تا ۴۴٫۰٪
اطمینان مدل	شاخص قابل اعتماد دقت نیست
جایگاه	ابزار کمکی با تأیید متخصص

منبع: جمع‌بندی تحریریه بر پایه منابع فهرست‌شده در همین مقاله.

پرسش‌های متداول

مدل‌های زبانی بزرگ در رادیولوژی دهان و فک برای چه وظایفی آزموده شده‌اند؟

آزمون و آموزش، تشخیص و پیشنهاد درمان، و تولید گزارش و ارتباط با بیمار.[۱]

دقت این مدل‌ها در پرسش‌های آزمون تخصصی چقدر است؟

در مرور سیستماتیک ۳۳٫۳٪ تا ۸۶٫۱٪ و در یک مطالعه جدید برای ChatGPT-5.2 تا ۹۵٫۷٪.[۱، ۴]

آیا LLMها می‌توانند تصاویر رادیوگرافی را تفسیر کنند؟

در یک ارزیابی، دقت چت‌بات‌های چندوجهی در پرسش‌های تصویری کمتر از ۳۰٪ بود.[۲]

توهم در این مدل‌ها چه شکلی دارد؟

پاسخ قطعی اما نادرست؛ در شناسایی ایمپلنت بیش از ۸۷٪ خطاها از این نوع بود.[۵]

چاپلوسی مدل زبانی یعنی چه؟

پیروی مدل از سرنخ نادرست کاربر؛ با سرنخ مقتدرانه دقت به ۲۸٫۵٪ تا ۴۴٫۰٪ کاهش یافت.[۳]

آیا اطمینان اعلام‌شده مدل قابل اعتماد است؟

خیر؛ در شرایط سرنخ مقتدرانه، اطمینان خوداظهاری میان پاسخ‌های درست و نادرست تفاوت معنادار نداشت.[۳]

جمع‌بندی

مدل‌های زبانی بزرگ در پرسش‌های متنی رادیولوژی دهان و فک پیشرفت سریعی نشان داده‌اند، اما در تفسیر تصویر، شناسایی جزئیات ایمپلنت و برابر سرنخ‌های گمراه‌کننده خطای زیادی دارند. توهم و چاپلوسی، همراه با اطمینان غیرقابل اتکا، مهم‌ترین خطرهای کاربرد بالینی‌اند. تا اعتبارسنجی و استانداردسازی بیشتر، این مدل‌ها ابزار کمکی آموزشی و نگارشی با بازبینی کامل متخصص هستند.

مطالعه مرتبط در دانشنامه

- هوش مصنوعی در رادیولوژی دندان و فک
- اعتبارسنجی هوش مصنوعی پوسیدگی
- برنامه‌ریزی ایمپلنت با هوش مصنوعی
- سگمنتیشن دندان در CBCT با هوش مصنوعی
- یادگیری عمیق در CBCT اندودانتیک
- مسئولیت تفسیر کل حجم CBCT

منابع

شماره‌های بالانویس در متن به همین فهرست پیوند دارند؛ هر منبع با شناسه DOI یا نشانی رسمی ناشر قابل بررسی است.

1. Liu Z, Nalley A, Hao J, Ai QYH, Yeung AWK, Tanaka R, et al. The performance of large language models in dentomaxillofacial radiology: a systematic review. Dentomaxillofac Radiol. 2025;54(8):613-631. DOI PubMed PMC

2. Jeong H, Jeon KJ, Lee C, Choi YJ, Jo GD, Han SS. Leveraging multimodal large language model chatbots in oral radiology: a comprehensive evaluation using questions from a Korean dental university. Dentomaxillofac Radiol. 2026;55(3):276-286. DOI PubMed

3. Geduk G, Çapar İ, Hasırcı UC. Sycophancy in Large Language Models in Oral and Maxillofacial Radiology: The Impact of Misleading Context on Image- and Text-Based Diagnostic Scenarios.

Diagnostics (Basel). 2026;16(20):3296. DOI PubMed PMC

4. Icen M, Temur KT. Comparative diagnostic accuracy of multiple large language models in oral and maxillofacial radiology specialty examinations: a 13-year analysis of performance and topic trends. BMC Oral Health. 2026;26(1):822. DOI PubMed PMC

5. Hlaing NHMM, Park K, Wan Q, Lee SJ, Gallucci GO, Lee JH. Accuracy and hallucination behaviors of artificial intelligence-based large language models for dental implant identification on radiographs. J Dent. 2026;173:106754. DOI PubMed

---

درباره نویسنده: تیم تحریریه علمی چاوش طب آذین

این مقاله را تیم تحریریه علمی چاوش طب آذین بر پایه منابع داوری‌شده و بیانیه‌های رسمی فهرست‌شده تهیه کرده است؛ ادعای بازبینی بالینی مستقل ندارد و جایگزین قضاوت متخصص نیست. آخرین به‌روزرسانی: ۲۷ آبان ۱۴۰۵.