Persian Language International Teaching Studies
Vol. 11, Issue 21, 2026; pp (65-92)
10.22034/maz.2025.22520.1177

مطالعات آموزش بین المللی زبان فارسی
دوره ۱۱، شمارهٔ ۲۱، پاییز و زمستان ۱۴۰۴، صفحات (۶۵-۹۲)
نشریه علمی / مقالهٔ پژوهشی

A Differential Analysis of Artificial Intelligence Capabilities in Assessing the Three Primary Skills of Persian Language Based on the Saadi Foundation's Reference Standard Relative to the Speaking Domain

Mojtaba Ebrahimi [۱]
Instructor of Persian Language Teaching affiliated with Al-Mustafa International University of Qom, Iran.

Mohammad Hosein Taleiy [۲]
PhD in Persian Language and Literature; Instructor of Persian Language Teaching affiliated with Al-Mustafa International University, Qom, Iran.

Abstract

The application of artificial intelligence (AI) in automated language assessment, while promising objectivity and scalability, faces the fundamental challenge of clashing with native, competency-based standards-particularly in low-resource languages such as Persian. Adopting a differential approach, this study conducts a comparative analysis of the theoretical capabilities of AI in assessing reading, writing, and listening skills against the descriptors of the «Reference Standard for Persian Language Education» by the Saadi Foundation.

Employing a critical documentary analysis method and designing a three-dimensional analytical matrix (comprising the indicators of machine-processability, context-dependency, and inference robustness), ۳۶ skill descriptors from the standard were examined. The analysis was conducted with the participation of two analysts and underwent iterative review. The findings revealed a pattern of «asymmetric validity distribution.» AI demonstrated the highest alignment in the reading skill (due to the static nature of text). In the writing skill, the contradiction of «correct form versus invalid content» (AI hallucination) and the inability to assess discourse logic reduce validity at advanced levels. In the listening domain, a profound gap was observed between acoustic decoding and the inference of layered meaning (tone, irony).

The differential analysis indicates that the gap in the three skills is primarily epistemological and concentrated at advanced levels (C۱–C۲). In contrast, the gap in the speaking skill (serving as the comparative reference point) is interactional-situational in nature and begins at intermediate levels. Consequently, complete reliance on AI for comprehensive Persian language assessment, especially at advanced levels, carries the risk of reducing communicative competence and threatening construct validity. Accordingly, a «skill-based hierarchical hybrid assessment model» is proposed, wherein the role of AI is moderated from being the exclusive assessor at basic levels to a supplementary tool under human supervision at advanced levels.

.

Keywords: Automated Language Assessment, Artificial Intelligence, Saadi Foundation Reference Standard, Validity Differential Analysis, Hybrid Assessment Model.

۱. Email: ebrahimi۲٤۹۱@gmail.com (Corresponding Author)
۲. Email: mohammadhosein.talei@iau.ac.ir

مطالعات آموزش بین المللی زبان فارسی
دوره ۱۱، شمارهٔ ۲۱، پاییز و زمستان ۱۴۰۴، صفحات (۶۵-۹۲)
نشریه علمی / مقالهٔ پژوهشی

Persian Language International Teaching Studies
Vol. 11, Issue 21, 2026; pp (65-92)
10.22034/maz.2025.22520.1177

تفاوت‌سنجیِ قابلیت‌های هوش مصنوعی در ارزیابی مهارت‌های سه‌گانه زبان فارسی، بر اساس استاندارد مرجع بنیاد سعدی، نسبت به حوزه گفتار

مجتبی ابراهیمی[۱]
مدرس آموزش زبان فارسی وابسته به جامعه المصطفی، قم.

محمدحسین طالعی[۲]
دکتری زبان و ادبیات فارسی،مدرس آموزش زبان فارسی وابسته به جامعه المصطفی، قم.

چکیده
کاربست هوش مصنوعی در ارزیابی خودکار زبان، با وعده عینیت و مقیاس‌پذیری، با چالش بنیادین تضاد با استانداردهای بومی قابلیت‌محور، به‌ویژه در زبان‌های کم‌منابعی مانند فارسی، مواجه است. این پژوهش با رویکردی تفاوت‌سنجانه، به تحلیل تطبیقی قابلیت‌های نظری هوش مصنوعی در سنجش مهارت‌های خوانداری، نوشتاری و شنیداری، در تقابل با توصیف‌گرهای «استاندارد مرجع آموزش زبان فارسی» بنیاد سعدی می‌پردازد.
با به‌کارگیری روش تحلیل اسنادی انتقادی و طراحی یک ماتریس تحلیلی سه‌بعدی (شامل شاخص‌های ماشین‌پذیری، وابستگی به بافت و استواری استنتاج)، ۳۶ توصیف‌گر مهارتی استاندارد واکاوی شد. تحلیل با مشارکت دو تحلیلگر و بازبینی مجدد انجام پذیرفت. یافته‌ها الگویی از «توزیع نامتقارن روایی» را آشکار کرد. هوش مصنوعی در مهارت خوانداری (به‌دلیل ماهیت ایستایی متن) بیشترین انطباق را نشان داد. در مهارت نوشتاری، تناقض «فرم درست در برابر محتوای نامعتبر» (توهم هوش مصنوعی) و ناتوانی در ارزیابی منطق گفتمان، روایی سطوح عالی را کاهش می‌دهد. در حوزه شنیدار، شکاف عمیقی میان رمزگشایی صوتی و استنباط معنای لایه‌دار (لحن، کنایه) مشاهده شد.
تحلیل تفاوت‌سنجی، حاکی از آن است که شکاف در مهارت‌های سه‌گانه عمدتاً معرفت‌شناختی و در سطوح عالی (۱-۲) متمرکز است، درحالی‌که در مهارت گفتار (به عنوان نقطه مرجع مقایسه) شکاف ماهیتی تعاملی-موقعیتی دارد و از سطوح میانی آغاز می‌شود. درنتیجه، اتکای کامل به هوش مصنوعی برای سنجش جامع زبان فارسی، به‌ویژه در سطوح پیشرفته، با خطر تقلیل توانش ارتباطی و تهدید روایی سازه همراه است. بر این اساس، «مدل سنجش ترکیبی سلسله‌مراتبی مهارت‌محور» پیشنهاد می‌شود که در آن، نقش هوش مصنوعی از ارزیاب انحصاری در سطوح پایه، به ابزار کمکی تحت نظارت انسان در سطوح عالی تعدیل می‌یابد.

واژه‌های کلیدی: **ارزیابی خودکار زبان، هوش مصنوعی، استاندارد مرجع بنیاد سعدی، تفاوت‌سنجی‌روایی، مدل سنجش ترکیبی.**

تاریخ دریافت: ۱۴۰۴/۰۹/۱۳ | تاریخ بازنگری: ۱۴۰۴/۱۱/۲۷ | تاریخ پذیرش: ۱۴۰۵/۰۱/۲۹

۱. Email: ebrahimi۲۴۹۱@gmail.com (نویسندهٔ مسئول)
۲. Email: mohammadhosein.talei@iau.ac.ir

مقدمه

۱–۱. چارچوب نظری و بیان مسأله

سامانه آموزش و سنجش زبان در دهه اخیر، با جهش مدل‌های زبانی بزرگ (LLM) و گسترش سیستم‌های سنجش خودکار زبان (ALA)، در حال گذار از الگوی «ارزیابی صورتی» به «تحلیل هوشمند معنایی» است. این فناوری‌ها با وعده ارتقای عینیت، مقیاس‌پذیری و حذف سوگیری‌های انسانی، به عرصه حساس آزمون‌سازی راه یافته‌اند. با این حال، به‌کارگیری هوش مصنوعی در زبان فارسی، به عنوان یک زبان کم‌منبع با ویژگی‌های جامعه‌شناختی و فرهنگی منحصربه‌فرد، با چالشی بنیادین به نام «شکاف روایی سازه» مواجه است. مسأله اصلی زمانی آشکار می‌شود که خروجی‌های احتمال‌محور و الگومحور ماشین در تقابل با توصیف‌گرهای کیفی، بافت‌محور و قابلیت‌محور «استاندارد مرجع بنیاد سعدی» قرار می‌گیرد.

مطالعات پیشین به وضوح نشان می‌دهند که هوش مصنوعی در سنجش مهارت گفتار زبان فارسی با چالش‌های بنیادین معرفت‌شناختی روبروست. پژوهش مؤلف، ابراهیمی و طالعی، (۱۴۰۴) با ارائه چارچوبی تحلیلی، استدلال نموده که ماهیت تعاملی، فی‌البداهه و وابسته به بافتِ گفتار، باعث می‌شود هوش مصنوعی نتواند «توانش ارتباطی» را در سطوح میانی و عالی به درستی بازتاب دهد و شکافی عمیق در روایی سازه ایجاد کند.

پژوهش حاضر، در امتداد و با الهام از آن چارچوب نظری، و با هدف تعمیم و تطبیق نظام‌مند آن بر سه مهارت اصلی دیگر طراحی شده است. پرسش کلیدی این است: سنجش تطبیقی قابلیت‌های هوش مصنوعی در سه حوزه مهارت های شنیداری ، نوشتاری و خوانداری، نسبت به حوزه بحرانی گفتار، چه الگوی نامتقارنی را آشکار می‌سازد؟

این پرسش زمانی اهمیتی دوچندان می‌یابد که تفاوت ذاتی در نوع داده‌های ورودی (نشانک صوتی، متن)، فرآیندهای شناختی مورد ارزیابی (درک، تولید) و ماهیت خروجی (پاسخ گزینشی، متن نگارشی) در هر مهارت را در نظر بگیریم. استاندارد مرجع بنیاد سعدی در سطوح پیشرفته، انتظاراتی نظیر «درک طنز و کنایه»، «رعایت سیاق‌های فرهنگی»، «انسجام منطقی» و «تولید گفتمان اغنایی» را مطرح می‌کند صحرایی و همکاران، (۱۳۹۵) این در حالی است که هسته عملکرد هوش مصنوعی صرفاً بر «تطبیق الگوی آماری» استوار است و از «مدیریت معنا در بافت»، که محور استاندارد سعدی است، ناتوان است. این تضاد ماهوی، به‌ویژه در تقابل میان مهارت‌های پایدارتری(مانند خواندن با متن ثابت) و مهارت‌های پویاتر (مانند گفتار تعاملی)، ضرورت یک بررسی مقایسه ای را ایجاب می‌کند تا مرزهای دقیق صلاحیت ماشین و نقاط اجتناب‌ناپذیر دخالت ارزیاب انسانی در هر یک از مهارت‌های خواندن، نوشتن و شنیدن مشخص شود.

۱–۲. اهمیت وضرورت پژوهش

ضرورت این پژوهش از سه منظر کلیدی، قابل تبیین است:

الف) صیانت از روایی سازه و اعتبار آزمون‌های بسندگی: نتایج آزمون‌های سرنوشت‌ساز (مانند آزمون «سامفا») پایه‌ای برای تصمیم‌گیری‌های مهم آموزشی و شغلی است. اتکای انحصاری به هوش مصنوعی، بدون شناخت سهم متفاوت و الگوی نامتقارن خطاهای ماشین در هر مهارت، می‌تواند منجر به

«کم‌نمایی سازه» شود. برای نمونه، ممکن است سیستم در سنجش «درک مطلب» متون پیچیده به دلیل توهم زبانی دچار خطا شود، یا در ارزیابی «انشای تحلیلی» نتواند اصالت فکر را تشخیص دهد. این نادیده گرفتنِ تفاوت‌های کارکردی، اعتبار گواهی‌نامه‌های صادره را به طور نظام‌مند تهدید می‌کند.

ب) تأمین عدالت آموزشی و آزمونی: هوش مصنوعی می‌تواند در هر یک از مهارت‌ها، سوگیری‌های پنهانی داشته باشد. در مهارت شنیداری، این سوگیری ممکن است علیه لهجه‌های غیرمعیار یا کیفیت پایین ضبط صوت باشد. در مهارت نوشتاری، ممکن است به سبک‌های نوشتاری خاص، ساختارهای پیچیده غربی یا حتی محتوای تولیدشده توسط خود ماشین نمره بالاتری دهد. در مهارت خوانداری، امکان تفسیر نادرست سؤالات استنباطی وجود دارد. شناسایی این کانون‌های متفاوت سوگیری در هر مهارت، پیش‌شرط طراحی سیستم‌های عادلانه‌تر و تضمین فرصت برابر برای همه زبان‌آموزان است.

ج) تدوین مدل سنجش ترکیبی بهینه و مبتنی بر شواهد: مهم‌ترین خروجی کاربردی این پژوهش، ارائه نقشه‌راهی برای تقسیم کار هوشمندانه بین انسان و ماشین است. تحلیل تفاوت‌سنجی قابلیت‌ها به طراحان آزمون کمک می‌کند تا مشخص کنند در کدام مهارت و در چه سطحی می‌توان با اطمینان بیشتری به ارزیابی خودکار روی آورد و در کدام حوزه‌ها، نظارت و قضاوت انسانی نقشی حیاتی و غیرقابل حذف دارد. این مدل، هم از مزایای فناوری بهره می‌برد و هم از خطر تقلیل‌گرایی و تهدید روایی سازه جلوگیری می‌کند.

۱-۳. پرسش‌ها و اهداف پژوهش

پژوهش حاضر در پی پاسخ به این پرسش اصلی است:

هوش مصنوعی بر اساس شاخص های ماشین پذیری وابستگی به بافت و استواری استنتاج، چه الگوی نامتقارنی از میزان پوشش توصیف گرهای مهارت های خوانداری، نوشتاری و شنیداری در استاندارد بنیاد سعدی را در نسبت با مهارت گفتار نشان می دهد؟

این پرسش به سه پرسش فرعی تفکیک می‌شود:

۱. میزان انطباق قابلیت‌های هوش مصنوعی با توصیف‌گرهای هر یک از سه مهارت (خوانداری، نوشتاری، شنیداری)

در سطوح مختلف استاندارد سعدی چگونه است؟

۲. الگوی «توزیع نامتقارن روایی» میان این سه مهارت و در مقایسه با مهارت گفتار چه ویژگی‌هایی دارد؟

۳. بر اساس این الگو، چه مدل سنجش ترکیبی‌ای برای آزمون‌های بسندگی زبان فارسی قابل طراحی است؟

۱-۴. اهداف پژوهش

- تبیین تفاوت‌های عملکردی (کمی و کیفی) هوش مصنوعی در ارزیابی مهارت‌های نوشتاری، خوانداری و شنیداری زبان فارسی، در تقابل با توصیف‌گرهای دقیق «استاندارد مرجع بنیاد سعدی».
- انجام تحلیل مقایسه‌ای (تفاوت‌سنجی) و تعیین میزان و ماهیت شکاف روایی در هر یک از این سه مهارت، در نسبت‌سنجی با یافته‌های پیشین در حوزه مهارت گفتاری.
- شناسایی و دسته‌بندی بن‌بست‌های فنی (نظیر محدودیت داده) و معرفت‌شناختی (نظیر ناتوانی در درک طنز) ماشین در مواجهه با مؤلفه‌های فرهنگی، استعاری و تعاملیِ این

مهارت‌هـا در سـطوح میانـی و پیشـرفته.

• (هـدف عملیاتـی) ارائـه چارچوبـی بـرای «مـدل سـنجش ترکیبـی سلسـله‌مراتبی مهارت‌محـور» کـه بـر اسـاس یافته‌هـای تفاوت‌سـنجی، حوزه‌هـای اقتـدار ماشـین و ضـرورت مداخلـه انسـانی را بـرای هـر مهـارت بـه طـور جداگانـه مشـخص نمایـد.

۲. مبانی نظری و پیشینه پژوهش

ارزیابـی خـودکار زبـان بـه عنـوان زیرشـاخه‌ای از پـردازش زبـان طبیعـی[۱] (NLP) ، بـا ظهـور «مدل‌هـای زبانـی بـزرگ»[۲] (LLM) و معماری‌هـای مبتنـی بـر ترنسـفورمر، تحولـی شـگرف را تجربـه کـرده اسـت. ایـن تحـول، وعـده گـذار از تحلیل‌هـای صـوری بـه پردازش‌هـای پیچیـده معنایـی و افزایـش عینیـت و کارایـی در سـنجش زبـان را داده اسـت. واس و همـکاران، (۲۰۲۳). بـا ایـن حـال، بحـث پیرامـون روایـی، تعریـف سـازه و انطبـاق بـا اسـتانداردهای بومـی، در کانـون چالش‌هـای نظـری و عملـی ایـن حـوزه قـرار دارد. مـرور نظام‌منـد پژوهش‌هـای پیشـین در دو حیطـه کلان، ضـروری اسـت: الـف) نقشـه‌ی دانشـی و روندهـای پژوهشـی در حـوزه سـنجش زبـان و اعتباریابـی، و ب) مطالعـات تجربـی و نظـری پیرامـون به‌کارگیـری هـوش مصنوعـی در ارزیابـی مهارت‌هـای زبانـی بـه تفکیـک.

۲–۱. نقشه‌ی دانشی و روندهای پژوهشی در سنجش زبان

یـک مـرور گسـترده دانش‌نگاشـتی توسـط آریادوسـت و همـکاران (۲۰۲۱) بـر روی ۴۷۳۶ مقالـه در مجـلات هسـته سـنجش زبـان و مجـلات عمومـی زبان‌شناسـی کاربـردی و اکتسـاب زبـان دوم انجـام شـد. یافته‌هـای کلیـدی ایـن پژوهـش نشـان داد کـه تمرکـز مجـلات هسـته عمدتـاً بـر سـنجش درک مطلـب خوانـداری و شـنیداری (اولویـت اول)، و سـپس بـر جنبه‌هایـی از عملکـرد گفتـاری و نوشـتاری ماننـد ارزیابـان و اعتباریابـی اسـت. در مقابـل، پژوهش‌هـای سـنجش در مجـلات عمومـی، اولویـت اصلـی را بـر سـنجش واژگان، مهـارت گفتـاری، انشاءنویسـی، دسـتور و خوانـدن قـرار داده‌انـد. نکتـه حائـز اهمیـت، تأکیـد مجـلات عمومـی بـر سـازه‌های عاطفـی، حافظـه، دانـش صریـح در مقابـل ضمنـی، و آگاهـی زبانـی اسـت کـه در مجـلات هسـته غایـب بودنـد. ایـن مطالعـه همچنیـن یـک شـکاف روش‌شـناختی حیاتـی را آشـکار سـاخت: اکثـر پژوهش‌هـا پیـش از بـه کارگیـری ابزارهـای سـنجش، اعتباریابـی اسـتنتاج‌محور روی آنهـا انجـام نداده‌انـد آریادوسـت و همـکاران، (۲۰۲۱). ایـن یافتـه بـر ضـرورت توجّـه جدی‌تـر بـه مبانـی روایـی در پژوهش‌هـای کاربـردی، از جملـه پژوهش‌هـای مربـوط بـه سـنجش هوشـمند، تأکیـد می‌کنـد.

۲–۲. کاربست هوش مصنوعی در سنجش مهارت‌های دریافتی

مطالعـات متعـددی بـه اثربخشـی هـوش مصنوعـی در تقویـت و سـنجش مهارت‌هـای دریافتـی (شـنیداری و خوانـداری) پرداخته‌انـد. فخرریانـا و هردینـا، (۲۰۲٤) در پژوهشـی نشـان دادنـد کـه پلتفرم‌هـای هـوش مصنوعـی ماننـد(Atomic Learning) می‌تواننـد بـرای طراحـی آزمون‌هـای درک مطلـب شـنیداری و خوانـداری متناسـب بـا نیازهـای خـاص دانش‌آمـوزان بـه کار رونـد. بـه طـور کلّـی، هنـز، (۲۰۲٦) در یـک فصـل کتـاب تحلیلـی اسـتدلال می‌کنـد کـه هـوش مصنوعـی بـا ارائـه محتـوای شخصی‌سازی‌شـده، بازخـورد تشـخیصی بلادرنـگ و شبیه‌سـازی زمینه‌هـای ارتباطـی اصیـل

۱. Natural Language Processing
۲. Large Language Models

از طریـق متن‌هـای تولیـدی و منابـع چندوجهـی، مهارت‌هـای دریافتـی را تقویـت می‌کنـد. مطالعـات کاربـردی در اکـوادور نیـز ایـن ادعاهـا را پشـتیبانی می‌کننـد. پژوهـش لاسـکانو پـرز و آلتامیرانـو کارواجـال (۲۰۲۵) و آکوسـتا-ریورا و بـدون واکا، (۲۰۲۵) هـر دو بـه بهبـود معنـادار مهارت‌هـای دریافتنـی زبـان انگلیسـی در بیـن دانش‌آمـوزان بـا اسـتفاده از ابزارهـای هـوش مصنوعـی ماننـد(Duolingo و ChatGPT) اشـاره داشـته‌اند. بـه طـور مشـابه، وایـه تـورس (۲۰۲۵) گـزارش داد کـه دانشـجویان مهندسـی، مهـارت خوانـدن و شـنیدن خـود را از طریـق اسـتفاده از ایـن ابزارهـا بهبـود بخشـیده‌اند. ایـن مجموعـه یافته‌هـا حاکـی از پتانسـیل بـالای هـوش مصنوعـی در پـردازش و سـنجش داده‌هـای متنـی و صوتـی سـاختاریافته اسـت. بـا ایـن حـال، ایـن مطالعـات عمومـاً بـر بهبـود کلـی مهارت‌هـا یـا رضایـت کاربـر متمرکـز بـوده و کمتـر بـه تحلیـل نظام‌منـد شـکاف‌های روایـی ایـن سیسـتم‌ها در مواجهـه بـا مؤلفه‌هـای پیچیـده‌ای ماننـد اسـتنباط، درک معنـای ضمنـی یـا سـوگیری فرهنگـی پرداخته‌انـد.

۳-۲. کاربست هوش مصنوعی در سنجش مهارت‌های تولیدی

در حـوزه مهارت‌هـای تولیـدی (گفتـار و نوشـتار)، پژوهش‌هـا نتایـج پیچیده‌تـری را نشـان می‌دهنـد و بـر چالش‌هـای پیـشِ رو تأکیـد دارنـد.

مهـارت گفتـاری: در پژوهش‌هـای انسـان‌محور بـرای ارزیابـی گفتـار در زبـان فارسـی، اسـتاجی و همـکاران (۱۴۰۳) نشـان دادنـد کـه مؤلفه‌هایـی ماننـد تکـرار و خطـای شـروع، قـادر بـه تمایـز سـطوح بسـندگی نیسـتند. ایـن یافته‌هـا معیارهـای تشـخیصی انسـانی را روشـن می‌سـازد و زمینه‌ای بـرای مقایسـه بـا دقـت تشـخیص سیسـتم‌های خـودکار فراهـم می‌کنـد. از سـوی دیگـر، مطالعـات متعـددی بهبـود در روانـی، دقـت، تلفـظ و تمایـل بـه ارتبـاط زبان‌آمـوزان را بـا اسـتفاده از ارزیابی‌هـای گفتاری مبتنـی بر هوش مصنوعـی گـزارش کرده‌انـد. گـراب، (۲۰۲۵) ؛ لوپـز-مینوتـا و همـکاران، (۲۰۲۵) دخیـل و همـکاران، (۲۰۲۵). ایـن ابزارهـا بازخـورد فـوری و شـخصی و فرصت‌هـای تمریـن بی‌قضاوت فراهـم می‌کننـد. سـاپوترا و ویدیاسـتوتی، (۲۰۲۴). شـواهد مشـابهی از اثربخشـی ایـن ابزارهـا در بافـت زبـان فارسـی نیـز گـزارش شـده اسـت. صبـوری و حاج‌ملـک، (۱۴۰۲) در یـک مطالعـه نیمه‌تجربـی نشـان دادنـد کـه فنـاوری بازشناسـی گفتـار می‌توانـد بـا دقـت بالایی پیشـرفت مهـارت تلفـظ زبان‌آمـوزان را ردیابی کنـد. به طور مکمل، قاسـمی و برومنـد، (۱۴۰۳) بـر اسـاس داده‌هـای پیمایشـی تأکیـد کردنـد کـه شخصی‌سـازی یادگیـری از طریـق ابزارهـای هـوش مصنوعـی، نظیـر(ChatGPT و Duolingo) منجـر بـه افزایـش خودکارآمـدی زبان‌آمـوزان در مهـارت مکالمـه می‌شـود. ایـن یافته‌هـا همسـو بـا ادبیـات جهانی، بـر نقـش حمایتی هوش مصنوعـی در تمریـن و سـنجش مؤلفه‌هـای صـوری و روان‌شـناختی گفتـار تأکیـد دارنـد.

بـا ایـن حـال، محدودیت‌هـای جـدی نیـز شناسـایی شـده‌اند. سـاپوترا و ویدیاسـتوتی، (۲۰۲۴) و گـراب، (۲۰۲۵) هـر دو بـه ناتوانـی هـوش مصنوعـی در درک هـوش هیجانـی، آگاهـی فرهنگـی و ظرافت‌هـای ارتبـاط انسـانی، و همچنیـن نگرانی‌هـا دربـاره سـوگیری الگوریتمـی اشـاره کرده‌انـد. مطالعـه دخیـل و همـکاران، (۲۰۲۵) نیـز اگرچـه بهبـود در دسـتور، واژگان، آهنـگ کلام و روانـی را نشـان داد، امـا تفـاوت معنـاداری در تلفـظ بیـن گـروه کنتـرل و آزمایـش نیافـت کـه نشـان‌دهنده حساسـیت محـدود هـوش مصنوعـی بـه واج‌شناسـی دقیـق زبـان دوم اسـت.

مهارت نوشتاری: چالش اصلی در این حوزه، تعریف سازه و مسائل اعتبار در صورت دسترسی آزمون‌دهندگان به ابزارهای کمکی هوشمند است. واس و همکاران، (۲۰۲۳) در یک مناظره نظری-

عملی به مسائل پیچیده‌ای مانند طراحی روبریک، عدالت، انصاف، سوگیری و مالکیت معنوی در صورت استفاده از فناوری‌های کمکی از جمله هوش مصنوعی تولیدی در آزمون‌سازی پرداختند. این بحث، تناقض اساسی بین «تولید فرم‌های درست» و «ارزیابی محتوای اصیل» را برجسته می‌سازد. این مباحث نظری، شکاف عمیق بین ارزیابی صوری (فرم) توسط ماشین و ارزیابی محتوایی-اصیلی (معنا) توسط انسان را برجسته می‌سازد، که هسته چالش پژوهش حاضر است.

۲-۴. یافته‌های کلیدی در مهارت گفتار: نقطه عزیمت پژوهش حاضر

فراتر از گزارش‌های کاربردی از اثربخشی یا عدم اثربخشی ابزارهای خاص، ادبیات نظری و تجربی از محدودیت‌های ذاتی و سیستماتیک هوش مصنوعی در بازتولید قضاوت کیفی انسانی در سنجش زبان پرده برمی‌دارد. این محدودیت‌ها ریشه در معماری بنیادین مدل‌های کنونی داشته و درک آن‌ها برای طراحی هرگونه مدل سنجش ترکیبی ضروری است.

۱-۴-۲. چارچوب نظری:

جدایی فرم از معنا و مسأله روایی در قلب چالش انطباق هوش مصنوعی با استانداردهای قابلیت‌محور، تضاد ماهوی بین پردازش فرم و درک معنا قرار دارد. از نگاه نظری، مدل‌های زبانی بزرگ، صرفاً بر اساس توزیع‌های آماری نشانه‌های زبانی آموزش دیده‌اند و فاقد هرگونه دسترسی به قصد ارتباطی یا تجربه تجسد یافته در جهان هستند. بنابراین، این سیستم‌ها در بهترین حالت می‌توانند الگوهای صوری صحیح را بازتولید کنند، اما قادر به فهم معنای گزاره در بافت فرهنگی-اجتماعی خاص نیستند بندار و کلر، (۲۰۲۰). این ناتوانی معرفت‌شناختی به طور عینی در پدیده «توهم» تجلی می‌یابد، که در آن مدل متنی از نظر زبانی روان اما از نظر محتوایی نادرست یا بی‌اساس تولید می‌کند؛ پدیده‌ای که به عنوان یک نقص ساختاری در معماری مبتنی بر یادگیری عمیق شناخته شده است جی و همکاران، (۲۰۲۳). با این حال، شواهد جدیدی نیز از قابلیت‌های رو به رشد این مدل‌ها در درک معانی پیچیده حکایت دارد. برای نمونه، یی و همکاران (۲۰۲۵) نشان دادند که مدل‌های زبانی بزرگ در چارچوبی مبتنی بر یادگیری زمینه‌ای، می‌توانند طنز را - که نمونه‌ای بارز از تفاوت معنی ظاهری و نیت گوینده است - با دقت بالا تشخیص داده و برای آن استدلالی شبه‌انسانی ارائه کنند. این یافته حاکی از آن است که اگرچه مدل‌ها فاقد تجربه زیسته هستند، اما از طریق الگوهای آماری پیچیده می‌توانند تا حدی بر شکاف فرم و معنا فائق آیند. ارزیابی مبتنی بر چنین سیستمی، مستعد آن است که «فرم» غنی را با «محتوا» معتبر اشتباه گرفته و روایی سازه را تهدید کند. از این منظر، اعتبارسنجی هر گونه به کارگیری فناوری در سنجش، مستلزم استدلالی است که نشان دهد تفسیر و استفاده مورد نظر از نمرات، توسط شواهد کافی پشتیبانی می‌شود. کان، (۲۰۱۳) ، چاپل و واس، (۲۰۲۱). این نگرانی‌های نظری در مورد روایی سازه، در مطالعات کاربردی در بافت‌های آموزشی غیرانگلیسی نیز بازتاب یافته است. برای نمونه، رشیدی (۱۴۰۴) در یک مطالعه تحلیلی-توصیفی با مرور نظام‌مند منابع فارسی به این نتیجه رسید که در «ارزشیابی توصیفی خودکار»، روایی سازه در غیاب بازبینی انسانی به شدت آسیب‌پذیر است. این یافته به طور مشخص نشان می‌دهد که چگونه حذف قضاوت کیفی انسانی می‌تواند سازه مورد نظر در ارزیابی را تهدید کند و لزوم یک رویکرد ترکیبی (انسان-ماشین) را برای حفظ اعتبار ارزیابی‌ها پررنگ می‌سازد.

۲-۴-۲. شواهد تجربی از شکاف‌های کلیدی:

لهجه، تعامل و پیچیدگی وظیفه این چالش‌ها در چارچوب گسترده‌تری به نام «تحلیل عملگرایی» در پردازش زبان طبیعی قرار می‌گیرد که هدف آن درک معانی ضمنی، نیت گوینده و نشانه‌های متنی و فرامتنی است گومده، (۲۰۲۴). پیشرفت‌های اخیر مبتنی بر مدل‌های ترنسفورمر، اگرچه بهبودهایی در درک زمینه ایجاد کرده‌اند، اما هنوز فاصله زیادی با درک عملگرایانه انسانی دارند.شواهد تجربی این شکاف را در حوزه‌های عینی تأیید می‌کنند. نخست، در حوزه پردازش صدا، مطالعات نشان می‌دهند سوگیری علیه گونه‌های غیرمعیار ذاتی طراحی سیستم‌هاست. پژوهش‌ها حاکی از آن است که تفاوت لهجه به تنهایی می‌تواند سبب سوگیری ادراکی در تشخیص هویت صدا شود، گویی سیستم‌ها (و انسان‌ها) فرض می‌گیرند افراد عمدتاً با یک لهجه ثابت صحبت می‌کنند. سانتوس و همکاران، (۲۰۲۵). این مسأله در زبان‌های چندلهجه‌ای و فاقد نگارش استاندارد به اوج می‌رسد، جایی که بهترین سیستم‌های بازشناسی گفتار نیز ممکن است نرخ خطای کلمه‌ای در حدود ۳۰٪ را گزارش کنند نیگماتولینا و همکاران، (۲۰۲۰). جالب آنکه، این چالش لزوماً ناشی از «نویز» یا «ناپایداری» بیشتر گفتار زبان‌آموزان نیست، بلکه اغلب به دلیل انحراف میانگین آماری تولیدات آن‌ها از الگوی گفتار بومی معیار است لی و جگر، (۲۰۲۰). دوم، در حوزه سنجش تعامل، هسته «توانش ارتباطی» در استانداردهایی مانند سعدی، یعنی مدیریت پویای گفتمان، به طور بنیادین از دسترس سیستم‌های کنونی خارج است. پژوهش‌ها به صراحت نشان می‌دهند که ماهیت الگوریتمی و متن‌محور این سیستم‌ها، آن‌ها را از درک و ارزیابی منابع تعاملی زنجیره‌ای، زمانی و تجسم‌یافته که کنش متقابل انسانی را تشکیل می‌دهند، ناتوان می‌سازد هوت، (۲۰۲۰)، راد، (۲۰۲۵). این ناتوانی به صورت عملی در شکاف عملکردی سیستم‌های ارزیابی خودکار بین تکالیف «کنترل‌شده» (مانند خواندن با صدای بلند) و تکالیف «آزاد و خودجوش» (مانند ارائه یا گفت‌وگو) آشکار می‌شود، به طوری که دقت و روایی سیستم در مورد دوم به شدت کاهش می‌یابد. لیو و همکاران، (۲۰۲۵)، زو و همکاران، (۲۰۲٤).

۳-۴-۲. خلأ موجود و ضرورت پژوهش حاضر:

رویکرد تفاوت‌سنجی مرور حاضر نشان می‌دهد اگرچه مطالعات متعددی به محدودیت‌های هوش مصنوعی در حیطه‌های خاص (به ویژه گفتار) اشاره کرده‌اند، و پژوهش‌های داخلی نیز عمدتاً بر جنبه‌های صوری یا انگیزشی تمرکز داشته‌اند صبوری و حاج‌ملک،(۱۴۰۲)؛ قاسمی و برومند، ۱۴۰۳) یا به کلیت چالش‌ها اذعان نموده‌اند (رشیدی، ۱۴۰۴؛) اما خلأی بنیادین در ادبیات وجود دارد. هیچ پژوهش نظام‌مندی با رویکرد «تفاوت‌سنجی» انجام نشده است که میزان و ماهیت «شکاف روایی» هوش مصنوعی را نه در یک مهارت، بلکه در تمامی مهارت‌های اصلی زبانی (خواندن، نوشتن، شنیدن) به صورت تطبیقی و در تقابل با یک استاندارد معیار بومی و قابلیت‌محور (مانند استاندارد بنیاد سعدی) تحلیل کند. پرسش اصلی این است که آیا الگوی شکاف شناسایی‌شده در مهارت گفتار، با همان شدت و کیفیت در مهارت‌های دیگر تکرار می‌شود یا خیر؟ پژوهش حاضر با هدف پر کردن این خلأ و ارائه الگویی از «توزیع نامتقارن روایی» هوش مصنوعی در پهنه مهارت‌های زبان فارسی طراحی شده است.

۲-۴-۴. پیوند چارچوب نظری با مطالعه حاضر

مرور نظری و تجربی فوق نشان می‌دهد که محدودیت‌های هوش مصنوعی در سنجش گفتار، ریشه در تضادی ماهوی بین پردازش آماری فرم و درک تعاملی معنا دارد. پژوهش پایه‌ای مؤلف ابراهیمی و طالعی، (۱۴۰۴) این چارچوب نظری را در بافت زبان فارسی و در تقابل با استاندارد بنیاد سعدی گسترش داده و گفتار را به عنوان حوزه‌ای با حداکثر شکاف روایی معرفی کرده است. مطالعه کنونی، این چارچوب تحلیلی را به عنوان سنگ بنای نظری خود می‌پذیرد و در صدد است تا با طراحی یک ابزار تحلیلی نوین و نظام‌مند (ماتریس تطبیقی سه‌بعدی که در بخش روش شرح خواهد آمد)، دامنه این واکاوی را به سه مهارت اصلی دیگر (شنیداری، خوانداری، نوشتاری) تعمیم دهد. بنابراین، هدف، نه بازتولید یا آزمون آن یافته‌ها، بلکه انجام یک تحلیل تفاوت‌سنجیِ ساختاریافته برای ترسیم نقشه کامل صلاحیت هوش مصنوعی در پهنه مهارت‌های زبان فارسیِ است.

۲-۵. شکاف‌های پژوهشی و ضرورت مطالعه حاضر

با وجود حجم رو به رشد پژوهش‌های کاربردی، چند شکاف معرفت‌شناختی و روش‌شناختی عمده در ادبیات موجود چشمگیر است:

۱. تمرکز بر زبان انگلیسی و غفلت از زبان‌های کم‌منابع: اکثر قریب به اتفاق مطالعات تجربی ذکرشده (مانند گراب، ۲۰۲۵؛ دخیل و همکاران، ۲۰۲۵؛ آکوستا-ریورا و بدون واکا، ۲۰۲۵) بر بستر زبان انگلیسی انجام شده‌اند. ویژگی‌های زبان‌های کم‌منابع مانند فارسی — با پیچیدگی‌های صرفی-نحوی، نظام نوشتاری خاص و بافت فرهنگی منحصربه‌فرد — در این پژوهش‌ها نادیده گرفته شده‌اند.

۲. عدم انطباق با استانداردهای بومی قابلیت‌محور: مطالعات پیشین عمدتاً به بهبود نمره یا رضایت‌سنجی پرداخته‌اند و کمتر به بررسی انطباق سیستمیک خروجی‌های هوش مصنوعی با توصیف‌گرهای کیفی و قابلیت‌محور یک استاندارد مرجع بومی (مانند استاندارد بنیاد سعدی) پرداخته‌اند. پژوهش استاجی و همکاران (۱۴۰۳) اگرچه در حوزه سنجش گفتار فارسی انجام شده، اما بر مؤلفه‌های عینی روانی گفتار متمرکز است و به چالش ارزیابی مؤلفه‌های پیچیده‌تر فرهنگی-معنایی نپرداخته است.

۳. تفاوت‌سنجی نظام‌مند بین مهارت‌ها: ادبیات موجود فاقدمطالعه‌ای است که به صورت تطبیقی و افتراقی، «شکاف روایی» هوش مصنوعی را در مهارت‌های چهارگانه زبانی، با اتکا به چارچوب نظری منسجم (مانند روایی مبتنی بر استدلال) و در تقابل با یک استاندارد معیار، بررسی کند. مرور آریادوست و همکاران (۲۰۲۱) گرچه حیطه‌های پژوهشی را ترسیم می‌کند، اما به تحلیل توانایی‌های فنی هوش مصنوعی در هر حیطه نمی‌پردازد.

٤. چالش‌های پردازش معنای ضمنی و کاربردشناسی: مطالعات محدودی در حیطه پردازش زبان طبیعی به چالش‌های بنیادین مدل‌ها در درک طنز، کنایه و معانی ضمنی پرداخته‌اند یی و همکاران،(۲۰۲۵؛ گومده، (۲۰۲٤). با این حال، پیوند این محدودیت‌های فنی با الزامات سنجش زبان در سطوح پیشرفته (C۱ و C۲) در یک استاندارد مرجع، در پژوهش‌های حوزه سنجش زبان مغفول مانده است.

۲-۶. جمع‌بندی و جایگاه پژوهش حاضر

بنابراین، پژوهش حاضر در پی پر کردن شکاف‌های شناسایی‌شده فوق است. این تحقیق با تمرکز بر زبان فارسی به عنوان یک زبان کم‌منابع، و با اتکا به استاندارد مرجع بنیاد سعدی به عنوان معیار بومی قابلیت‌محور، به انجام یک تحلیل تفاوت‌سنجی نظام‌مند از قابلیت‌های هوش مصنوعی در ارزیابی سه مهارت خوانداری، نوشتاری و شنیداری می‌پردازد. این مطالعه با وام‌گیری از چارچوب نظری «روایی مبتنی بر استدلال» و با آگاهی از محدودیت‌های فنی مدل‌ها در پردازش معنای ضمنی و کاربردشناسی، درصدد است تا الگویی از «توزیع نامتقارن روایی» هوش مصنوعی در مهارت‌های مختلف را ترسیم و ماهیت «شکاف معرفت‌شناختی» آن را در سطوح عالی زبان تبیین نماید. از این رو، یافته‌های این پژوهش می‌تواند به غنای ادبیات نظری سنجش زبان در بافت فناوری‌های نوین و ارائه راهکارهای عملی برای طراحی «مدل‌های سنجش ترکیبی» معتبر در آزمون‌های سرنوشت‌ساز زبان فارسی کمک شایانی کند.

۳. روش‌شناسی پژوهش

۳-۱. طرح پژوهش و رویکرد کلی

پژوهش حاضر از نظر هدف، کاربردی-توسعه‌ای و از نظر ماهیت و روش، یک تحلیل اسنادیِ انتقادی با رویکرد تطبیقی است. هدف بنیادی پژوهش، آزمودن تجربی یک سامانه نرم‌افزاری خاص نیست، که یافته‌های آن محدود به آن نرم‌افزار و تاریخ‌مصرف دارد، بلکه نقد معرفت‌شناختیِ امکانِ سنجش کیفی زبان توسط ماشین است. این نقد با تحلیل تقابلی ویژگی‌های ذاتی معماری هوش مصنوعی مبتنی بر الگوی آماری (نظیر مدل‌های زبانی بزرگ) در برابر الزامات قابلیت‌محورِ استاندارد مرجع بومی انجام می‌پذیرد. از این رو، تمرکز بر محدودیت‌های ساختاری و پایدار است، نه خطاهای گذرای یک پیاده‌سازی خاص. در این راستا، برای تبیین اعتبار سنجش هوشمند، از چارچوب نظری «روایی مبتنی بر استدلال کین، (۲۰۱۳) بهره گرفته شده است. این پژوهش در تداوم مطالعات پیشین نویسنده در حوزه سنجش گفتار، به دنبال ترسیم الگوی «تفاوت‌سنجی» صلاحیت ماشین در سایر مهارت‌های زبانی است.

۳-۲. جامعه تحلیلی و واحد پژوهش

تحلیل حاضر بر دو ستون اصلی استوار است:

۱. اسناد معیار: متن کامل «استاندارد مرجع آموزش زبان فارسی در جهان» بنیاد سعدی، (۱۳۹۵). واحد تحلیل در این بخش، ۳۶ توصیف‌گر عملکردی مربوط به مهارت‌های شنیداری، خوانداری و نوشتاری بودند که از جداول صفحات ۳۳ تا ۵۲ این سند استخراج شدند.

توصیف‌گرهای مهارت شنیداری				
توصیف‌گر	**صفحه**	**شماره جدول**	**سطح**	**کد**
تشخیص واج‌ها و آواهای زبان فارسی، درک محدود واژه‌های منفرد یا عبارات پربسامد در بافت	**۳۳**	**۲–۶**	نوآموز (N)	LIS-۰۱
درک واژه‌ها و عبارات پربسامد وابسته به بافت در مورد خود، خانواده و محیط اطراف	**۳۵**	**۲–۷**	مقدماتی (A۱)	LIS-۰۲
درک اطلاعات گفتاری به طول یک جمله در بافت ابتدایی فردی و اجتماعی	**۳۵**	**۲–۷**	مقدماتی (A۲)	LIS-۰۳
درک اطلاعات از گفتار جمله‌محور در بافت ابتدایی فردی و اجتماعی	**۳۸**	**۲–۸**	پیش‌میانی (B۱٫۱)	LIS-۰٤
درک گفتار ساده جمله‌محور در بافت‌های متنوع ابتدایی فردی و اجتماعی	**۳۸**	**۲–۸**	پیش‌میانی (B۱٫۲)	LIS-۰۵
درک گفتار ساده به طول جمله در بافت ابتدایی شخصی و اجتماعی	**۴۱**	**۲–۹**	میانی (B۲٫۱)	LIS-۰٦
درک متن کوتاه روایتی و توصیفی معمولی با ساختار زیربنایی روشن	**۴۱**	**۲–۹**	میانی (B۲٫۲)	LIS-۰۷
درک متون روایی و توصیفی معمولی (توصیف اشخاص، اشیا، مکان‌ها)	**۴۴**	**۲–۱۰**	فوق‌میانی (B۳٫۱)	LIS-۰۸
درک متون شنیداری روایت‌گونه و توصیفی بلند و منابع واقعی پیچیده	**۴۴**	**۲–۱۰**	فوق‌میانی (B۳٫۲)	LIS-۰۹
درک متون مختلف حتی پیچیده و بلند با موضوعات آشنا و ناآشنا	**۴۷**	**۲–۱۱**	پیشرفته (C۱٫۱)	LIS-۱۰
درک گفتارهای بلند، اصطلاحات خاص، ساختارهای پیچیده و ارجاعات فرهنگی	**۴۷**	**۲–۱۱**	پیشرفته (C۱٫۲)	LIS-۱۱
درک اطلاعات ضمنی و استنباطی، لحن، دیدگاه‌ها، کنایه، لطیفه، دوپهلوگویی و لهجه‌های غیرمعیار	**۵۰**	**۲–۱۲**	ماهر (C۲)	LIS-۱۲

توصیف‌گرهای مهارت خوانداری				
توصیف‌گر	**صفحه**	**شماره جدول**	**سطح**	**کد**
شناخت حروف الفبا، توانایی خواندن برخی واژه‌های ساده و جملات ساده منفرد	**۳۳**	**۲–۶**	نوآموز (N)	REA-۰۱
تشخیص حروف و نشانه‌های الفبایی، خواندن نسبتاً روان واژه‌ها و جملات بسیار ساده	**۳۵**	**۲–۷**	مقدماتی (A۱)	REA-۰۲
تشخیص واژه‌های کلیدی، عبارات کلیشه‌ای، درک پیام‌های قابل پیش‌بینی	**۳۵**	**۲–۷**	مقدماتی (A۲)	REA-۰۳
درک برخی معانی در متون سادهٔ پیوسته مرتبط با نیازهای فردی و اجتماعی	**۳۸**	**۲–۸**	پیش‌میانی (B۱٫۱)	REA-۰٤
فهم متون سادهٔ غیرپیچیده حاوی اطلاعات پایه‌ای در موضوعات فردی و اجتماعی	**۳۸**	**۲–۸**	پیش‌میانی (B۱٫۲)	REA-۰۵
درک کامل متن کوتاه و غیرپیچیده با اطلاعات پایه‌ای	**۴۱**	**۲–۹**	میانی (B۲٫۱)	REA-۰٦
درک متن نسبتاً بلند روایی و توصیفی با ساختار روشن و واژه‌های پربسامد	**۴۱**	**۲–۹**	میانی (B۲٫۲)	REA-۰۷
مطالعه و درک متون روایی و توصیفی معمولی، مقالات و گزارش‌های روز	**۴۴**	**۲–۱۰**	فوق‌میانی (B۳٫۱)	REA-۰۸
درک کامل متون روایی و توصیفی عادی و حتی پیچیده واقعی	**۴۴**	**۲–۱۰**	فوق‌میانی (B۳٫۲)	REA-۰۹
درک متون بسیاری از ژانرها با موضوعات گسترده آشنا و ناآشنا	**۴۷**	**۲–۱۱**	پیشرفته (C۱٫۱)	REA-۱۰
درک متون بلند با ماهیت تخصصی، دانشگاهی یا ادبی	**۴۷**	**۲–۱۱**	پیشرفته (C۱٫۲)	REA-۱۱
درک مؤلفه‌های زیبایی‌شناختی زبان و سبک‌های ادبی، آگاهی از بینامتنیت و ارجاعات فرهنگی	**۵۰**	**۲–۱۲**	ماهر (C۲)	REA-۱۲

توصیف‌گرهای مهارت نوشتاری				
توصیف‌گر	**صفحه**	**شماره جدول**	**سطح**	**کد**
نوشتن حروف الفبا، رونویسی واژه‌ها و عبارات ساده، تولید محدود واژه‌های منفرد	۳۴	۲–۶	نوآموز (N)	WRI-۰۱
نوشتن متن کوتاه و ساده، تکمیل فرم‌های اطلاعات شخصی	۳۷	۲–۷	مقدماتی (A۱)	WRI-۰۲
رفع نیازهای نوشتاری کاربردی (فهرست‌برداری، پیام‌های کوتاه، کارت‌پستال)	۳۷	۲–۷	مقدماتی (A۲)	WRI-۰۳
تولید جملات خبری، طراحی پرسش، نوشتن یادداشت و نامه ساده	۴۰	۲–۸	پیش‌میانی (B۱٫۱)	WRI-۰٤
رفع نیازهای کاربردی نوشتاری، نوشتن انشاء درباره اولویت‌های شخصی	۴۰	۲–۸	پیش‌میانی (B۱٫۲)	WRI-۰۵
نوشتن متن ساده و منسجم درباره موضوع روزمره، خلاصه‌نویسی ساده	۴۳	۲–۹	میانی (B۲٫۱)	WRI-۰٦
رفع نیازهای نوشتاری شغلی و دانشگاهی مقدماتی، روایت و توصیف در قالب‌های اصلی	۴۳	۲–۹	میانی (B۲٫۲)	WRI-۰۷
رفع نیازهای نوشتاری شغلی و دانشگاهی، خلاصه‌نویسی روان	۴۶	۲–۱۰	فوق‌میانی (B۳٫۱)	WRI-۰۸
نوشتن متن با جزئیات، نوشتن نامه رسمی و غیررسمی با آداب مناسب	۴۶	۲–۱۰	فوق‌میانی (B۳٫۲)	WRI-۰۹
تولید انواع مکاتبات رسمی و غیررسمی، خلاصه‌نویسی، گزارش‌ها و مقالات تحقیقی	۴۹	۲–۱۱	پیشرفته (C۱٫۱)	WRI-۱۰
تولید متن خوش‌ساخت و گویا، دفاع از عقاید با ارائه بحث قاطع و فرضیه‌سازی	۴۹	۲–۱۱	پیشرفته (C۱٫۲)	WRI-۱۱
تسلط بر دستور پیچیده حتی زمانی که تمرکز بر موضوع دیگری است، اشراف بر معانی ضمنی، کاربرد ضرب‌المثل‌ها و باهم‌آیی‌ها	۵۲	۲–۱۲	ماهر (C۲)	WRI-۱۲

۲. مبانی نظری معماری هوش مصنوعی: ادبیات نظری و فنی مرتبط با محدودیت‌های ذاتی مدل‌های زبانی بزرگ (LLM) و سیستم‌های پردازش زبان طبیعی، با تمرکز بر معماری مبتنی بر ترنسفورمر و چالش‌های آن در مواجهه با معانی ضمنی، بافت فرهنگی و ارزیابی محتوا (مانند بندار و کلر، ۲۰۲۰؛ جی و همکاران، ۲۰۲۳).

توجیه رویکرد تحلیلی: هدف این پژوهش، آزمون تجربی یک سامانه خاص نیست، بلکه نقد معرفت‌شناختی امکان سنجش کیفی توسط ماشین است. تمرکز بر محدودیت‌های ذاتی معماری، که تا زمان یک تحول الگوی پایدار خواهند ماند، به پژوهش امکان می‌دهد فراتر از ارزیابی‌های زودگذر نرم‌افزاری، به تحلیل مرزهای نظری صلاحیت هوش مصنوعی بپردازد.

۳-۳. چارچوب تحلیلی: ماتریس تفاوت‌سنجی سه‌بعدی

بـرای عملیاتی‌سـازی رویکـرد تفاوت‌سـنجی و تضمیـن نظام‌منـد بـودن تحلیـل، یـک ماتریـس تحلیلـی سـاختاریافته طراحـی و بـه کار گرفتـه شـد. فرآینـد تحلیـل در چهـار گام کلیـدی انجـام پذیرفـت:

گام ۱: استخراج و چارچوب‌بندی انتظارات. تمامی توصیف‌گرهای مهارتی استاندارد سعدی در سه حیطه هدف، استخراج و دسته‌بندی شدند.

گام ۲: طراحی و اعمال شاخص‌های کیفی. هر توصیف‌گر بر اساس سه شاخص کلیدی زیر مورد تحلیل کیفی قرار گرفت:

الف) شاخص ماشین‌پذیری: میزانی که ویژگی مورد نظر را می‌توان به داده‌های کمی، قطعی و عینی (مانند تعداد کلمات، فراوانی ساختارها) تبدیل کرد.

ب) شاخص وابستگی به بافت: درجه نیازمندی تفسیر صحیح آن ویژگی به دانش فرهنگی، پیش‌زمینه تاریخی، دانش جهان مشترک یا موقعیت ارتباطی خاص.

ج) شاخص استواری استنتاج: میزان قابلیت اعتماد به نتیجه‌گیری ماشین در مورد لایه‌های معنایی پنهان مانند طنز، کنایه، قصد گوینده یا استحکام منطقی استدلال.

گام ۳: تحلیل تقابلی. برای هر توصیف‌گر، توانایی ذاتی معماری هوش مصنوعی در برآورده کردن انتظار استاندارد، با توجه به سه شاخص فوق، مورد قضاوت تحلیلی قرار گرفت. این قضاوت بر اساس مطالعه عمیق مبانی نظری محدودیت‌های هوش مصنوعی و انطباق آن با ماهیت هر توصیف‌گر بود.

گام ٤: ترکیب‌بندی و تفاوت‌سنجی. خروجی تحلیل گام سوم برای هر مهارت، جمع‌بندی و الگوی حاکم بر آن استخراج شد. سپس الگوهای حاصل از سه مهارت شنیداری، خوانداری و نوشتاری با یکدیگر و با الگوی از پیش شناخته‌شده مهارت گفتار (به عنوان نقطه مرجع) به صورت تطبیقی مقایسه شد تا «توزیع نامتقارن روایی» ترسیم گردد.

۳-۳-۱. تعریف عملیاتی شاخص‌های سه‌گانه

بـرای تضمیـن شـفافیت فرآینـد تحلیـل و امـکان بازتولیـد پژوهـش، هـر یـک از سـه شـاخص کلیـدی بـر اسـاس طیـف لیکـرت، (۱۹۳۲) ۵ درجـه‌ای بـه صـورت زیـر عملیاتـی شـدند. ایـن تعاریـف مبنـای قضـاوت تحلیلـی در مـورد تمامـی ۳۶ توصیف‌گـر قـرار گرفـت:

الف) شاخص ماشین‌پذیری: میزان قابلیت تبدیل ویژگی مورد نظر به داده‌های کمی، قطعی و عینی.

نمره ۱ (کاملاً ناممکن): ویژگی کاملاً کیفی و وابسته به قضاوت انسانی است؛ مانند «زیبایی‌شناسی متن»، «لذت ادبی» یا «اصالت فک».

نمره ۲ (تا حد زیادی ناممکن): ویژگی عمدتاً کیفی است و تنها بخش کوچکی از آن قابل کمّی‌سازی است.

نمره ۳ (تا حدی ممکن): بخش‌هایی از ویژگی قابل کمّی‌سازی است اما جنبه‌های کیفی مهمی نیز دارد.

نمره ٤ (تا حد زیادی ممکن): ویژگی عمدتاً قابل کمّی‌سازی است، اما نیاز به تنظیم پارامترهای الگوریتمی دارد.

نمره ۵ (کاملاً ممکن): ویژگی به شاخص‌های کمی و عینی قابل تبدیل است؛ مانند «تعداد لغات صحیح»، «فراوانی ساختارهای دستوری» یا «املا».

ب) شاخص وابستگی به بافت: درجه نیازمندی تفسیر صحیح ویژگی به دانش فرهنگی، پیش‌زمینه تاریخی، دانش جهانِ مشترک یا موقعیت ارتباطی خاص.

نمره ۱ (کاملاً مستقل از بافت): ویژگی با دانش زبانی صرف (نحو و واژگان) قابل ارزیابی است.

نمره ۲ (وابستگی اندک): ویژگی نیازمند دانش محدود و عمومی از جهان است.

نمره ۳ (وابستگی متوسط): ویژگی نیازمند درک موقعیت ارتباطی و نقش‌های اجتماعی است.

نمره ٤ (وابستگی زیاد): ویژگی نیازمند دانش فرهنگی و اجتماعی خاص جامعه هدف است.

نمره ٥ (کاملاً وابسته به بافت): ویژگی بدون دانش عمیق فرهنگی، تاریخی، بینامتنی یا موقعیتی قابل ارزیابی نیست؛ مانند «درک کنایه»، «طنز»، «تلمیحات تاریخی» یا «ادب موقعیتی».

ج) شاخص استواری استنتاج: میزان قابلیت اعتماد به نتیجه‌گیری ماشین در مورد لایه‌های معنایی پنهان (منطبق با شواهد تجربی از ادبیات پژوهش).

نمره ۱ (کاملاً ناسازگار): استنتاج ماشین در این زمینه به صورت سیستمی خطا دارد و شواهد تجربی حاکی از عملکرد تصادفی یا معکوس است.

نمره ۲ (ناسازگار): دقت ماشین به‌طور معناداری پایین‌تر از ارزیاب انسانی است.

نمره ۳ (تا حدودی سازگار): ماشین در شرایط کنترل‌شده عملکرد قابل قبولی دارد، اما در شرایط طبیعی خطا می‌کند.

نمره ٤ (سازگار): دقت ماشین به ارزیاب انسانی نزدیک است، اما همچنان در موارد پیچیده **خطا دارد.**

نمره ٥ (کاملاً سازگار): استنتاج ماشین با قضاوت انسانی همخوانی بالایی دارد و شواهد تجربی این همخوانی را تأیید می‌کنند.

تبصره: قضاوت در مورد شاخص استواری استنتاج بر اساس مرور نظام‌مند شواهد تجربی موجود در ادبیات مانند یی و همکاران،(۲۰۲۵) در مورد تشخیص کنایه، یا جی و همکاران،(۲۰۲۳)در مورد توهم، انجام شده است.

۳-۳-۲. عملیاتی‌سازی کمی برای مقایسه مهارت‌ها و تعیین «درجه روایی کلی»

برای امکان مقایسه‌پذیری کمی بین مهارت‌ها و ارائه تصویری خلاصه از الگوی توزیع نامتقارن در جدول ۲، یک فرآیند کمّی‌سازی شفاف مبتنی بر خروجی‌های کیفی ماتریس طراحی شد. تأکید می‌شود که اعداد حاصل، داده‌های خام تجربی نیستند، بلکه بازنمایی کمّی‌شده از تحلیل کیفی برای مقایسه و خلاصه‌سازی هستند. این فرآیند در چهار گام زیر انجام پذیرفت:

گام اول: کدگذاری کیفی به کمّی. قضاوت‌های تحلیلی حاصل از ماتریس برای هر توصیف‌گر، بر اساس سه شاخص، در یک مقیاس عددی ۱ تا ۵ کدگذاری شد:

ماشین‌پذیری (M): ۱ (کاملاً ناممکن) ← ۵ (کاملاً ممکن)

وابستگی به بافت (C): ۱ (کاملاً مستقل) ← ۵ (کاملاً وابسته)

استواری استنتاج (I): ۱ (کاملاً ناسازگار) ← ۵ (کاملاً سازگار)

گام دوم: محاسبه شاخص ترکیبی شکاف روایی (CRGI). از آنجا که جهت تأثیر شاخص‌ها بر «شکاف» متفاوت بود، برای یکپارچه‌سازی از فرمول زیر استفاده شد:

$$CRGI = [(٦ - M) + C + (٦ - I)] / ٣$$

منطق فرمول: در شاخص‌های M و I، نمره بالاتر به معنای شکاف کمتر است، لذا با محاسبه مکمل (امتیاز – ٦)، جهت آن معکوس شد تا همسو با «شکاف بیشتر» باشد. در شاخص C، نمره بالاتر از ابتدا به معنای شکاف بیشتر است. حاصل، عددی بین ۱ تا ۵ است که CRGI بالاتر نشان‌دهنده شکاف روایی گسترده‌تر (یعنی قابلیت اعتماد کمتر ارزیابی هوش مصنوعی) است.

گام سوم: تبدیل به «درجه روایی کلی» برای جدول نهایی. برای آنکه در جدول ۲ مقادیر به شکل مستقیم‌تری تفسیر شوند (عدد بالاتر = روایی بیشتر)، شاخص CRGI معکوس گردید: درجه روایی اولیه = ۶ – CRGI

گام چهارم: گردسازی نهایی. مقادیر حاصل برای نمایش در ستون «درجه روایی کلی» جدول ۲، به یک مقیاس صحیح‌شده ۱ تا ۴ گرد شدند (۱: کمترین روایی، ۴: بیشترین روایی). این رویه کمّی‌سازی ثانویه، یافته‌های کیفی پژوهش را بر پایه‌ای شفاف، قابل بازبینی و قابل مقایسه استوار می‌سازد.

در جدول۳-۱، نمونه‌ای از تحلیل شش توصیف‌گر کلیدی از سه مهارت اصلی ارائه شده است. این نمونه‌ها چگونگی اعمال شاخص‌های سه‌گانه و محاسبه CRGI را نشان می‌دهند

جدول **۳-۱**. نمونه‌ای از تحلیل توصیف‌گرهای کلیدی بر اساس شاخص‌های سه‌گانه

مهارت	سطح	توصیف‌گر (شماره صفحه)	M	C	I	CRGI	استدلال تحلیلی
خوانداری	C۱	درک ایده‌های اصلی متون مباحثه‌ای	۴	۳	۴	**۲.۳۳**	استخراج ایده اصلی با الگوریتم‌های خلاصه‌سازی ممکن است؛ وابستگی به بافت متوسط
خوانداری	C۲	درک مؤلفه‌های زیبایی‌شناختی و زبان تاریخی	۲	۵	۲	**۴.۳۳**	زیبایی‌شناسی نیازمند تجربه انسانی است؛ ماشین صرفاً الگوی آماری می‌بیند
نوشتاری	B۲	تولید متن خوش‌ساخت با ساختار دستوری پیچیده	۵	۲	۴	**۲.۰۰**	ارزیابی دستور و ساختار برای ماشین کاملاً ممکن است
نوشتاری	C۱	دفاع از عقاید با استدلال منطقی	۲	۴	۲	**۴.۳۳**	ماشین قادر به تشخیص «توهم محتوایی» و ارزیابی منطق استدلال نیست
شنیداری	A۲	بازشناسی واج‌ها و کلمات پایه	۵	۱	۵	**۱.۳۳**	تبدیل گفتار به متن در سطح پایه با دقت بالا انجام می‌شود
شنیداری	C۱	درک لحن، کنایه و اطلاعات ضمنی	۲	۵	۲	**۴.۳۳**	شکاف در لایه تفسیر معنایی؛ لحن به‌عنوان داده مستقل پردازش می‌شود

۳-۳-۳. نمونه کدگذاری یک توصیف‌گر

برای شفاف‌سازی فرآیند تحلیل، در اینجا یک نمونه کامل از کدگذاری ارائه می‌شود:

توصیف‌گر: سطح ماهر (C2)، مهارت شنیداری، جدول ۱۲-۲، صفحه ۵۰: «توانایی درک اطلاعات ضمنی، استنباطی، لحن و دیدگاه‌ها در مکالمات تند و سریع»

تحلیل:

ماشین‌پذیری (نمره ۲): بخش «لحن و دیدگاه‌ها» به‌سختی قابل تبدیل به داده‌های کمی است. اگرچه بازشناسی کلمات ممکن است، اما تشخیص لحن عاطفی نیازمند مدل‌های پیچیده‌ای است که همچنان خطای بالایی دارند.

وابستگی به بافت (نمره ۵): درک «اطلاعات ضمنی و استنباطی» کاملاً وابسته به دانش فرهنگی و پیش‌زمینه مشترک است که ماشین فاقد آن است.

استواری استنتاج (نمره ۲): تحقیقات نشان می‌دهد مدل‌های زبانی در تشخیص کنایه و لحن عاطفی دقت پایینی دارند .

$$CRGI = [(٤,٣٣ = ٣/١٣ = ٣ / [٤ + ٥ + ٤] = ٣ / [(٢-٦) + ٥ + (٢-٦)$$

عدد ۳۳/۴ نشان‌دهنده شکاف روایی بسیار بالا در این توصیف‌گر است

۳-۴. چارچوب مقایسه‌ای و نقطه مرجع تحلیل

به منظور عملیاتی‌سازی رویکرد تفاوت‌سنجی، این مطالعه نیازمند یک نقطه مرجع مقایسه‌ایِ ثابت بود. چارچوب نظری و یافته‌های مطالعه پیشین مؤلف در حوزه «سنجش گفتار زبان فارسی» ابراهیمی و طالعی، (۱۴۰۴) که بر پایه استاندارد بنیادسعدی استوار است، به عنوان این نقطه مرجعِ از پیش تثبیت‌شده انتخاب گردید. در آن پژوهش، مهارت گفتار به دلیل مواجهه با ماهیت پویا، تعاملی و چندلایه، به عنوان حوزه‌ای با حداکثر شکاف روایی شناسایی و تحلیل شده بود. در مطالعه حاضر، «شکاف روایی» شناسایی‌شده در هر یک از مهارت‌های سه‌گانه با استفاده از ماتریس تحلیلی نوین، از نظر عمق (شدت)، دامنه (وسعت) و ماهیت (فنی در مقابل معرفت‌شناختی)، در مقایسه با آن نقطه مرجع نظری سنجیده می‌شود. این رویکرد امکان ترسیم الگوی توزیع نامتقارن روایی هوش مصنوعی در سراسر طیف مهارت‌های زبانی را فراهم می‌آورد.

۳-۵. اعتبار (روایی) تحلیلی

برای افزایش اعتبار تحلیلی پژوهش، از راهبردهای زیر استفاده شد:

مثلث‌سازی منابع: هر توصیف‌گر از سه منظر متفاوت تحلیل شد: الف) دقت متن استاندارد سعدی، ب) مبانی نظری محدودیت‌های هوش مصنوعی، ج) شواهد تجربی قیاسی از مطالعات زبان‌های مشابه.

ثبات درونی قضاوت: تمامی تحلیل‌های کیفی اولیه پس از فاصله زمانی دو هفته، مجدداً و بدون مراجعه به نتایج اولیه بازبینی شدند. سازگاری بالای قضاوت‌ها در دو مرحله، حاکی از ثبات نظام تحلیلی بود.

شفافیت و قابلیت پیگیری: با ارائه واضح ماتریس تحلیلی، شاخص‌ها و فرآیند استدلال (همانند

نمونه ارائه‌شده در بخش ۴ برای یک توصیفگر)، امکان بازبینی، نقد و بازتولید استدلال‌های پژوهش فراهم آمده است.

ارجاع به چارچوب تثبیت‌شده: استفاده از یافته‌های مقاله پایه در حوزه گفتار به عنوان نقطه ثابت مقایسه، انسجام و اعتبار تحلیل تفاوت‌سنجانه را افزایش می‌دهد.

برای اطمینان از اعتبار و پایایی تحلیل کیفی، فرآیند کدگذاری و ارزیابی توصیف‌گرها بر اساس ماتریس تحلیلی، با استفاده از روش توافق بین کدگذاران انجام شد. بدین منظور، دو تحلیلگر با تخصص در حوزه سنجش زبان و فناوری آموزشی، توصیف‌گرها را بر اساس سه شاخص ماشین‌پذیری، وابستگی به بافت و استواری استنتاج مورد قضاوت قرار دادند. پس از مرحله اول، تحلیل‌ها با فاصله دو هفته مجدداً توسط هر دو تحلیلگر بازبینی شد و میزان سازگاری درونی (ثبات قضاوت در طول زمان) و توافق بین تحلیلی‌ها (ثبات بین دو کدگذار) محاسبه و در سطح مطلوبی تأیید گردید.

۳-۶. پایایی تحلیل کیفی

برای اطمینان از پایایی قضاوت‌های تحلیلی، از روش توافق بین کدگذاران، استفاده شد. دو تحلیلگر با تخصص در سنجش زبان و فناوری آموزشی، به‌طور مستقل تمامی ۳۶ توصیف‌گر را بر اساس سه شاخص اصلی کدگذاری کردند. پایایی به دو روش محاسبه شد:

۱. **توافق درون‌کدگذاری:** هر تحلیلگر پس از دو هفته، کدگذاری ۱۰ توصیف‌گر (حدود ۲۸٪ از کل) را تکرار کرد. درصد سازگاری درون‌فردی برای تحلیلگر اول ۹۲٪ و برای تحلیلگر دوم ۸۹٪ به‌دست آمد.

۲. **توافق بین‌کدگذاران:** ضریب همبستگی درون‌طبقه‌ای (ICC) برای شاخص‌های سه‌گانه محاسبه شد. میانگین ICC برای شاخص ماشین‌پذیری ۰/۸۵، برای وابستگی به بافت ۰/۸۱ و برای استواری استنتاج ۰/۷۹ به‌دست آمد، که بر اساس معیار چیچتی، (۱۹۹۴) در سطح «خوب» تا «عالی» قرار می‌گیرد. موارد اختلاف از طریق بحث و توافق دوطرفه حل‌وفصل شدند.

۴. تحلیل تطبیقی و یافته‌ها

۴-۱. چارچوب تحلیلی و مرور نقطه مرجع

این بخش به واکاوی تطبیقی توصیف‌گرهای مهارتی استاندارد مرجع بنیاد سعدی در سه حوزه شنیداری، خوانداری و نوشتاری می‌پردازد. همانطور که در بخش روش اشاره شد، تحلیل حاضر با اتکا به چارچوب نظری مطالعه پیشین در مورد گفتار به عنوان نقطه مرجع، به تفاوت‌سنجی سه مهارت دیگر می‌پردازد. یافته‌های جدول ۲ (که در ادامه ارائه می‌شود) نشان می‌دهد که الگوی شکاف روایی در مهارت‌های سه‌گانه، از نظر ماهیت و دامنه با الگوی حاکم بر مهارت گفتار تفاوت بنیادین دارد. در ادامه، این تفاوت‌ها در پرتو توصیف‌گرهای استاندارد واکاوی می‌شوند

۴-۲. ارائه یافته کلیدی: الگوی توزیع نامتقارن (جدول ۲)

حاصل تحلیل ماتریسی و محاسبه شاخص ترکیبی شکاف روایی (CRGI)، الگویی است که در جدول ۲ تحت عنوان «توزیع نامتقارن روایی هوش مصنوعی در مهارت‌های چهارگانه زبان فارسی» خلاصه شده است.

جدول شماره ۲

توزیع نامتقارن روایی هوش مصنوعی در مهارت‌های چهارگانه زبان فارسی				
مهارت / مولفه	**گفتار(نقطه مرجع)**	**شنیدار**	**خواندار**	**نوشتار**
طبیعت داده	تولید صوتیِ پویا، تعاملی، وابسته به بافت	دریافت صوتیِ پویا	دریافت متنیِ ایستا	تولید متنیِ پایدار
حداکثر انطباق (در چه سطوحی؟)	سطوح پایین (A) (تلفظ، واژگان پایه)	سطوح پایین و میانی (A-B) (رمزگشایی صوتی، درک گزاره‌های روشن)	اکثر سطوح (A-C۱) (درک صریح متن، استخراج ایده اصلی)	سطوح پایین و میانی (A-B) (دقت صوری: دستور، املا، نگارش)
نوع شکاف غالب	معرفت‌شناختی-تعاملی (راهبردهای جبرانی، مدیریت گفتمان)	معرفت‌شناختی-حسی (درک لحن، کنایه، لهجه‌های غیرمعیار)	معرفت‌شناختی-فرهنگی (بینامتنیت، زیبایی‌شناسی، زبان تاریخی)	معرفت‌شناختی-منطقی (ارزیابی صحت محتوا، استدلال، منطق گفتمان بومی)
بحرانی‌ترین سطح	سطوح میانی به بالا (C۲-B۲)	سطوح پیشرفته (C۲-C۱)	سطح ماهر (C۲)	سطوح پیشرفته (C۱-C۲)
ماهیت محدودیت	ذاتی و تعاملی (نیاز به هوش موقعیتی و فرهنگی)	ذاتی-حسی و فنی-منابعی (وابسته به داده‌های گفتاری متنوع)	ذاتی-فرهنگی (نیاز به دانش پیشینه و تجسد فرهنگی)	ذاتی-ارزیابی (ناتوانی در تمایز فرم از محتوای معتبر)
درجه روایی کلی (از کم به زیاد)	۱ (کمترین)	۲	٤ (بیشترین)	۳
صلاحیت جایگزینی کامل با انسان	خیر (به‌ویژه در سطوح B۲+)	خیر (در سطوح C۱+)	آری (با احتیاط) (به جز در مؤلفه‌های ادبی-فرهنگی سطح C۲)	خیر (در ارزیابی محتوا و استدلال در سطوح C۱+)

همانطور که در جدول ۲ مشاهده می‌شود، مهارت خوانداری با کسب بالاترین «درجه روایی کلی» (عدد ۴) در یک سر طیف، و مهارت گفتار با پایین‌ترین درجه (عدد ۱) در سر دیگر قرار دارد. مهارت‌های شنیداری و نوشتاری به ترتیب با درجات ۲ و ۳، موقعیت میانی را اشغال می‌کنند. این رتبه‌بندی کمّی، به وضوح توزیع نامتقارن قابلیت اعتماد هوش مصنوعی را نشان می دهد.

۴-۳. واکاوی تفصیلی مهارت‌ها

۴-۳-۱. مهارت شنیداری : شکاف میان بازشناسی آواشناختی و استنباط عملیاتی

تحلیل جداول ۶-۲ تا ۸-۲ (صفحات ۳۳ تا ۳۸) نشان می‌دهد که در سطوح نوآموز و مقدماتی، هوش مصنوعی به دلیل تمرکز استاندارد بر «بازشناسی واج‌ها، کلمات پایه و عبارات کوتاه»، دارای بالاترین سطح انطباق است. در این لایه، الگوریتم‌های «تبدیل گفتار به متن«

با اتکا بر مدل‌های آواشناختی، عملکردی دقیق دارند. اما با صعود به سطوح عالی، چالش معرفت‌شناختی آغاز می‌شود.

علاوه بر این، یکی از جدی‌ترین بن‌بست‌های هوش مصنوعی در انطباق با سطح ماهر (جدول ۱۲-۲، صفحه ۵۰)، درک «گونه‌های غیرمعیار و لهجه‌های خاص» است. در حالی که استاندارد بنیاد سعدی در سطوح عالی، فارسی‌آموز را ملزم به درک مکالمات تند، محاوره‌ای و دارای «زوائد زبانی»می‌کند، مدل‌های زبانی هوش مصنوعی عمدتاً بر پیکره‌های رسمی (کتابی) آموزش دیده‌اند. این امر منجر به یک «خطای سیستماتیک» در سنجش می‌شود؛ جایی که ماشین، شکستگی‌های طبیعی فعل و واژه در زبان فارسی گفتاری را به عنوان «غلط شنیداری» یا «نویز» تلقی کرده و بر خلاف صراحت استاندارد در سطح ماهر، توانایی استنباطِ واقعی زبان‌آموز در محیط‌های بومی را دست‌کم می‌گیرد.

طبق جدول ۱۲-۲ (صفحه ۵۰)، زبان‌آموز سطح ماهر باید قادر به درک «اطلاعات ضمنی، استنباطی، لحن و دیدگاه‌ها» در مکالمات تند و سریع باشد. هوش مصنوعی در لایه «رمزگشای» موفق است، اما در لایه «تفسیرِ ارزش عاطفی لحن» دچار نقص فنی است. در زبان فارسی، معنای بسیاری از گزاره‌ها از طریق «تکیه» و «آهنگ» متغییر می‌کند (مانند تفاوت پرسش معکوس یا طنز با جملات خبری مشابه). ماشین به دلیل معماری مبتنی بر احتمالات متنی، «نوای گفتار» را به عنوان داده‌ای مستقل از معنا پردازش می‌کند. این امر منجر به «کم‌نمایی سازه» در سنجش درک مطلب شنیداری می‌شود، زیرا ماشین قادر نیست میان «معنای لغوی» و «ارزش کاربردشناختی» که از طریق لحن منتقل می‌شود، پیوند برقرار کند.

- تفاوت‌سنجی نسبت به گفتار: برخلاف مهارت گفتار که در آن ماشین با چالش مضاعفِ پردازشِ «سیگنال‌های صوتیِ نویزدار و تولیدات آواشناختی لرزانِ زبان‌آموز» روبروست، در مهارت شنیداری ورودی معمولاً از منابع استاندارد و باکیفیت است. لذا هرچند هوش مصنوعی در هر دو مهارت در لایه «استنباط معنای پنهان» ناتوان است، اما در شنیدار به دلیل حذف متغیرِ نویزِ لهجه، روایی و انطباق به مراتب بالاتری نسبت به گفتار نشان می‌دهد.

۲-۳-۴. مهارت خوانداری : روایی حداکثری در لایه متن و بن‌بستِ بینامتنیت

طبق تحلیل اسنادی این پژوهش، مهارت خواندن سازگارترین مهارت با منطق هوش مصنوعی است. تحلیل جدول ۱۰-۲ (صفحه ۴۴) نشان می‌دهد که توصیف‌گرهایی نظیر «درک ایده‌های اصلی متون مباحثه‌ای» کاملاً با قابلیت‌های «خلاصه‌سازی استخراجی» و «درک مطلب ماشین»در مدل‌های ترنسفورمر همخوانی دارند.

اما در سطح ماهر (صفحه ۵۰)، استاندارد سعدی بر درک «مولفه‌های زیبایی‌شناختی» و «زبان تاریخی و مذهبی» تأکید دارد. هوش مصنوعی متن را به صورت «توزیع آماری واژگان» می‌بیند و فاقد «تجسد فرهنگی» است. در نتیجه، درک «لذت ادبی» یا «تلمیحات بینامتنی» که ریشه در تاریخ و حافظه جمعی ایرانیان دارد، برای ماشین صرفاً یک «احتمال کلامی» است نه یک فهم عمیق. بنابراین، سنجش این لایه توسط ماشین، فاقد روایی ماهوی است و آزمون را به یک تستِ واژگانیِ پیشرفته تقلیل می‌دهد.

• تفاوت‌سنجی نسبت به گفتار: برخلاف مهارت گفتار که ماهیتی «پویا، لحظه‌ای و غیرقابل بازگشت» دارد، مهارت خوانداری دارای ماهیتی «ایستا، متنی و پایدار» است. در اینجا ماشین با هیچ نویز ورودی (مانند لکنت یا غلط تلفظی) مواجه نیست و بر داده‌های متنیِ پاک‌سازی شده تمرکز دارد. به همین دلیل، هوش مصنوعی در سنجش خواندن، انطباق بسیار بالاتری با استاندارد سعدی نسبت به مهارت گفتار دارد و شکاف روایی در اینجا در کمترین سطح ممکن قرار می‌گیرد.

۳-۳-۴. مهارت نوشتاری: تناقض انسجام صوری و توهم محتوایی

در مهارت نوشتاری، هوش مصنوعی با پیچیده‌ترین تناقض خود یعنی «فرمِ غنی در برابرِ محتوایِ نامعتبر» روبروست. تحلیل جدول ۱۱-۲ (صفحه ۴۹) بر تولید متن «خوش‌ساخت، گویا و با ساختار دستوری پیچیده» تأکید دارد. هوش مصنوعی در تصحیح صوری (نحو، املا و علائم نگارشی) بی‌رقیب است، اما در سنجش «صحت محتوایی» و «استدلال منطقی» دچار پدیده «توهم هوش مصنوعی »است.

طبق صفحه ۴۹ استاندارد، زبان‌آموز باید بتواند «از عقاید خود دفاع کند». اگر داوطلب، استدلالی سست یا غلط را با فارسیِ سره و ساختارهای دستوری پیچیده بیان کند، ماشین به دلیل «فریبِ فرم»، نمره بالایی به او می‌دهد، در حالی که ارزیاب انسانی متوجه «پوچی محتوا» می‌شود. برای نمونه، فرض کنید زبان‌آموزی در سطح پیشرفته موضوع انشای «تأثیر شبکه‌های اجتماعی بر انسجام خانواده» را دریافت کند. وی متنی روان با ساختارهای پیچیده (مانند جملات شرطی مرکب، اسم مصدرهای متعدد) و واژگانی فاخر تولید می‌کند، اما محتوای استدلالش مبتنی بر یک «توهم» آماری رایج در مدل‌های زبانی است، مثلاً ادعای نادرستِ «بر اساس آمار منتشر شده در سال ۱۴۰۲، ۹۵٪ از خانواده‌های تهرانی به دلیل اعتیاد به شبکه‌های اجتماعی از هم پاشیده‌اند» که منبعی ندارد. یک ارزیاب انسانی فوراً نامعتبر بودن این آمار و سست بودن استدلال متکی بر آن را تشخیص می‌دهد. اما یک سیستم ارزیابی خودکار مبتنی بر(LLM)به دلیل تمرکز بر شاخص‌های صوری (روانی، دستور، تنوع واژگان) و ناتوانی ذاتی در راستی‌آزمایی محتوایی، احتمالاً به چنین متنی نمره بالایی خواهد داد. این مثال عینی، خطر جدی «جایگزینی اعتبار صوری با اعتبار محتوایی» را در سنجش خودکار نوشتار نشان می‌دهد. همچنین، ماشین ممکن است منطق گفتمان فارسی (مانند تواضع کلامی یا اطناب‌های ادبی) را که در صفحه ۵۱ استاندارد بر آن تأکید شده، به دلیل آموزش بر پایه منطق خطی و مستقیم زبان‌های غربی، به عنوان «عدم انسجام» جریمه کند و روایی آزمون را مخدوش سازد.

۴-۳-۳-۱ سوگیری سبکی و گفتمانی در نوشتار

فراتر از چالش محتوا، هوش مصنوعی ممکن است در مواجهه با «سبک‌های نگارشی ریشه‌دار در فرهنگ‌های مختلف» دچار سوگیری شود. استاندارد سعدی در سطح ماهر (صفحه ۵۲) بر «تناسب سبک با موضوع و موقعیت» تأکید دارد. با این حال، مدل‌های زبانی عمدتاً بر پیکره‌های متنی با منطق گفتمانی مستقیم و کم‌ابهام (عمدتاً انگلیسی) آموزش دیده‌اند. بنابراین، ممکن است «سبک‌های تعاملی، غیرمستقیم و مبتنی بر ایهام» که در سنت ادبی

و مکاتبات رسمی فارسی کاربرد دارند، از سوی الگوریتم به عنوان «عدم وضوح»، «انحراف از موضوع» یا حتی «ضعف در انسجام» تفسیر و جریمه شوند. این سوگیری پنهان، نه تنها عدالت آموزشی را به خطر می‌اندازد، بلکه باعث تقلیل غنای سبکی زبان فارسی در فرآیند آموزش و سنجش به یک الگوی یکسان و ماشین‌پسند می‌گردد.

- تفاوت‌سنجی نسبت به گفتار: در نوشتار، چالش‌های فنیِ بازشناسیِ گفتار (ASR) حذف شده و ماشین زمان کافی برای تحلیل لایه‌های نحوی و ساختاری را دارد؛ لذا در سنجش «دقت صوری»، نوشتار نمره‌ای بالاتر از گفتار می‌گیرد. اما در لایه «تولید معنا»، هر دو مهارت در سطوح عالی با ریزش شدید روایی مواجه‌اند؛ با این تفاوت که در گفتار، «روانی کلام» عامل فریب ماشین است و در نوشتار، «پیچیدگی دستوری و بلاغت صوری.
- در لایه‌ی دستوری، هوش مصنوعی با چالشی روبروست که استاندارد سعدی در تبیین سطح ماهر (صفحه ۵۲) به خوبی بر آن اشاره کرده است. «نظارت بر تولیدات خود حتی زمانی که تمرکز بر موضوع دیگری است». ماشین در ارزیابی نوشتار، صرفاً به محصول نهایی نگاه می‌کند و نمی‌تواند میان «لغزش» ناشی از سرعت و «خطای ساختاری» ناشی از عدم دانش، تمایز قائل شود. بر اساس استاندارد سعدی، یک نویسنده سطح ماهر ممکن است به دلیل تمرکز بر پیچیدگی محتوا، دچار لغزش‌های جزئی شود که از ارزش مهارتی او نمی‌کاهد، اما هوش مصنوعی با رویکردی «صفر و یکی» و «مکانیکی» این لغزش‌ها را هم‌تراز با عدم تسلط قلمداد کرده و روایی سنجش را در سطوح پیشرفته مخدوش می‌کند.

۴-۴. بازخوانی یافته‌های سنجش گفتار (به عنوان نقطه مرجع مقایسه‌ای)

برای تکمیل فرآیند تفاوت‌سنجی، ضروری است وضعیت مهارت گفتاری به عنوان «بحرانی‌ترین نقطه شکاف روایی» مورد استناد قرار گیرد. همانطور که در پژوهش‌های پیشین تبیین گردید، گفتار به دلیل گره خوردن با «توانش تعاملی» و لزوم مدیریت «راهبردهای جبرانی» (صفحه ۴۲ استاندارد)، سخت‌ترین آزمون برای ماشین است. ماشین در این حوزه با ناتوانی در تشخیص «راهبردهای ارتباطی» (مانند دگرگویی کلام برای جبران فراموشی واژه) روبروست و این هوشمندیِ زبان‌آموز را به اشتباه به عنوان «عدم تسلط» جریمه می‌کند. همچنین، سوگیری الگوریتمی علیه لهجه‌های غیرمعیار و ناتوانی در درک «تاب‌آوری زبانی» در تعاملات زنده، گفتار را در سطوح عالی« میانی به بالا» به طور کامل از قلمرو صلاحیتِ روایی ماشین خارج می‌کند. این بن‌بستِ تعاملی، به عنوان یک «نقطه مرجع»، نشان‌دهنده حداکثرِ انحراف از استاندارد مرجع بنیاد سعدی است که در مقایسه با آن، مهارت‌های سه‌گانه فوق، پایداری بیشتری از خود نشان می‌دهند.

۵. بحث و نتیجه‌گیری نهایی

برای درک الگوی «توزیع نامتقارن روایی» که در این مطالعه شناسایی شد، لازم است ابتدا وضعیت مهارت گفتار، که به عنوان خط مبنا و نقطه مرجع مقایسه عمل می‌کند، مرور گردد. همانطور که در پژوهش پایه‌ای مؤلف (ابراهیمی و طالعی، ۱۴۰۴) به تفصیل نشان داده شد، در مهارت گفتار یک رابطه معکوس و وابسته به سطح میان پیچیدگی انتظارات استاندارد سعدی و قابلیت‌های ماشین حکمفرماست. یافته‌های کلیدی آن پژوهش که چارچوب مقایسه حاضر را شکل می‌دهد، در جدول ۱ خلاصه شده است.

جدول شماره ۱

سطح مهارت	مولفه کلیدی استاندارد سعدی	وضعیت عملکرد هوش مصنوعی	سطح چالش روایی سازه
نوآموز/مقدماتی (N-A)	صحت تلفظ، واژگان پایه، جملات کلیشه‌ای	بسیار کارآمد: دقیق‌تر و سریع‌تر از انسان در تشخیص الگو	ناچیز
میانی (B)	راهبردهای جبرانی، تعامل، نیاز به مخاطب تنظیم‌کننده	دچارکژفهمی: تفسیر استراتژی ارتباطی به عنوان خطا؛ ناتوان در ایفای نقش تسهیل‌گر	متوسط
پیشرفته/ماهر (C)	طنز، کنایه، اقناع، ادب، تناسب اجتماعی، مدیریت گفتمان زنده	ناتوان: فقدان شعور فرهنگی و عاطفی، کوری نسبت به هنجارهای موقعیتی، ناتوانی در مدیریت گفتگوی پویا	بسیار بالا

تحلیل تفاوت‌سنجی حاضر، که در ادامه و با گسترش آن چارچوب پایه انجام شده است، نشان می‌دهد که الگوی کلی وابستگی به سطح (رابطه معکوس پیچیدگی و روایی) که در مهارت گفتار مشاهده شد، در سه مهارت شنیداری، خوانداری و نوشتاری نیز صادق است. با این حال، نقطه آغاز چالش، عمق شکاف و ماهیت محدودیت‌ها در هر مهارت، به دلیل تفاوت در ماهیت داده‌ورزی (صوتی/متنی، دریافتی/تولیدی) به طور قابل توجهی متفاوت است.

همانطور که در جدول ۲ (بخش ۴-۲) مشاهده می‌شود، مهارت خوانداری: بیشترین انطباق را دارد. چالش‌های اصلی (مانند درک بینامتنیت و زیبایی‌شناسی) تنها در سطوح عالی و کاملاً از نوع معرفت‌شناختی ظاهر می‌شوند. در سطوح پایه و میانی، شکاف روایی نزدیک به صفر است (حتی بهتر از وضعیت گفتار در جدول بالا).

مهارت نوشتاری: در سطوح پایه و میانی در سنجش صوری عالی عمل می‌کند، اما با تناقض ذاتی تولید فرم/توهم محتوا روبروست. شکاف روایی آن در سطوح عالی شدید است، اما ماهیت آن با گفتار متفاوت است: نه ناتوانی در تعامل، بلکه ناتوانی در ارزیابی صحت منطقی و انسجام گفتمانیِ مبتنی بر فرهنگ.

مهارت شنیداری: شکافی دوگانه دارد. در لایه رمزگشایی صوتی (خصوصاً در سطوح پایه) عملکرد خوبی دارد، اما در لایه تفسیر معنایی لایه‌دار (لحن، کنایه) در سطوح میانی به بالا دچار شکاف معرفت‌شناختی می‌شود. با این حال، به دلیل حذف متغیر نویز تولید گفتار زبان‌آموز، روایی کلی آن به مراتب بالاتر از مهارت گفتار است.

در نهایت، مهارت گفتار (نقطه مرجع جدول فوق) به دلیل درگیر بودن همزمان با چالش‌های رمزگشایی صوتیِ نویزدار، تولید تعاملی و نیاز به درک بافت موقعیتی زنده، همچنان در دورترین فاصله از صلاحیت کامل ماشین قرار دارد. این در حالی است که مهارت خوانداری در نزدیک‌ترین فاصله، و مهارت‌های نوشتاری و شنیداری در میانه این طیف قرار می‌گیرند. برای درک شهودی این الگو، می‌توان به خلاصه تطبیقی ارائه‌شده در جدول ۲ مراجعه کرد که به وضوح نشان می‌دهد چگونه ماهیت هر مهارت، نوع و شدت شکاف هوش مصنوعی را تعیین می‌کند.

۵-۱. تلفیق یافته‌ها و بحران روایی در سطوح عالی

تحلیل تفاوت‌سنجی نشان می‌دهد که انطباق هوش مصنوعی با استاندارد سعدی، تابعی خطی نبوده و به شدت تحت تأثیر «ماهیت مهارت» و «سطح پیچیدگی زبانی» قرار دارد. شکاف روایی از مهارت خوانداری (با بیشترین انطباق) به سمت مهارت‌های نوشتاری، شنیداری و در نهایت گفتاری (با کمترین انطباق) به طور نامتقارن افزایش می‌یابد. در سطوح عالی (C۱-C۲)، هسته چالش، یک تضاد معرفت‌شناختی بنیادین است: در حالی که استاندارد بر درک و تولید «معنا در بافت اجتماعی» تأکید دارد، معماری هوش مصنوعی صرفاً قادر به پردازش و ارزیابی «فرم» و الگوی آماری زبان است. این ناتوانی ذاتی در مواجهه با مؤلفه‌هایی مانند طنز، کنایه، بینامتنیت و ادب موقعیتی، منجر به «کم‌نمایی سازه» شده و روایی آزمون‌های بسندگی را به شکلی سیستماتیک تهدید می‌کند.

۵-۲. چالش‌های فنی: توهم و سوگیری

این شکاف معرفت‌شناختی در عمل به صورت چالش‌های فنی جدی خود را نشان می‌دهد. نخست، پدیده «توهم هوش مصنوعی» در مهارت نوشتاری می‌تواند منجر به نمره‌دهی کاذب به متونی با ساختار صوری صحیح اما محتوای نامعتبر یا بی‌اساس گردد. دوم، سوگیری الگوریتمی ذاتی علیه گونه‌های غیرمعیار و لهجه‌های متنوع در مهارت‌های شنیداری و گفتاری، با انتظار استاندارد از درک و تولید زبان در بافت‌های واقعی و متنوع در تناقض است و عدالت آزمون را مخدوش می‌سازد.

۵-۳. تقلیلِ توانشِ زبانی به توانشِ متنی:

نگاهی به جداول تحلیلی استاندارد سعدی (به‌ویژه ستون‌های مجزای مهارت دستوری و واژگانی در صفحات ۴۳ و ۴۶) نشان می‌دهد که این استاندارد، «دانش زبانی» را نه صرفاً در خدمت تولید متن، بلکه به عنوان یک ستون مستقل از شخصیت زبانیِ فرد می‌بینند. هوش مصنوعی با رویکردی عملکردگرا، دستور و واژه را فقط در بافت می‌سنجد. این «روش‌شناسیِ تقلیل‌گرایانه» باعث می‌شود که دانشِ بالقوه‌ی زبان‌آموز و «عمقِ واژگانی» او که در استاندارد سعدی (سطح فوق‌میانی به بالا) بر آن تأکید شده، نادیده گرفته شود و سنجش هوشمند فقط به «سطحِ لغزانِ متن» محدود بماند.»

۵-۴. محدودیت‌های پژوهش

در کنار یافته‌های حاضر، شایان ذکر است که این مطالعه با چند محدودیت روش‌شناختی همراه بوده است که تفسیر نتایج را مقید می‌سازد. نخست، تمرکز تحلیل بر محدودیت‌های ذاتی و نظری معماری هوش مصنوعی مبتنی بر الگوی آماری بوده است، نه ارزیابی تجربی عملکرد یک سامانه خاص. بنابراین، یافته‌ها بیشتر مرزهای امکان نظری را ترسیم می‌کنند و ممکن است با پیشرفت‌های فنی آینده یا معماری‌های الگوی جدید نیاز به بازنگری داشته باشند. دوم، فرآیند تحلیل کیفی و کدگذاری توصیف‌گرهای استاندارد سعدی، با به کارگیری ماتریس ساختاریافته و روش سه‌بعدی، تا حدی متکی بر تفسیر پژوهشگران از این توصیف‌گرها و تطابق آن با مبانی نظری هوش مصنوعی است. جهت کاهش این ذهنیت، قضاوت‌های تحلیلی در دو مرحله مجزا بازبینی شدند، اما رویکرد

کیفی حاضر، لزوماً عینیت یک آزمون تجربی کمّی را ندارد. سوم، این مطالعه عمدتاً بر «ارزیابی» متمرکز بود و کاربرد هوش مصنوعی در نقش مکمل آموزشی (مانند ارائه بازخورد تمرینی) را که ممکن است روایی سازه در آن مطرح نباشد، به طور تفصیلی مورد بحث قرار نداد. پذیرش این محدودیت‌ها، زمینه را برای مطالعات آینده که به آزمون تجربی مدل پیشنهادی یا تحلیل دقیق‌تر هر مؤلفه در محیط‌های واقعی می‌پردازند، فراهم می‌سازد.

۵-۵. پیشنهادهای راهبردی و مدل اجرایی

بر اساس نتایج این «تفاوت‌سنجی»، مدل «سنجش ترکیبی سلسله‌مراتبی» به عنوان راهکار بهینه پیشنهاد می‌گردد:

۱. اتوماسیون در سطوح پایه (A۱-A۲): در این سطوح که زبان جنبه مکانیکی دارد، هوش مصنوعی می‌تواند به عنوان ارزیاب اولیه با حداقل نیاز به نظارت انسانی باشد.

۲. نظارت انسانی در سطوح میانی (B۱-B۲): در این لایه، ماشین نقش غربالگر اولیه را ایفا کرده و موارد دارای ناسازگاری در نمرات فرعی، برای بازبینی به انسان ارجاع داده شوند.

۳. مرجعیت انسانی در سطوح عالی (C۱-C۲): در سنجش مهارت‌های تولیدی (نوشتار و گفتار) در سطوح عالی، به دلیل بن‌بست‌های معرفت‌شناختی شناسایی شده، ارزیاب انسانی باید مرجع نهایی باشد و هوش مصنوعی صرفاً به عنوان ابزار کمکی برای سنجشِ «دقت صوری» به کار گرفته شود.

به طور خلاصه، این پژوهش در تداوم و گسترش یافته‌های مطالعه پیشین در حوزه گفتار ابراهیمی، طالعی، (۱۴۰۴) نقشه کاملی از صلاحیت هوش مصنوعی در ارزیابی مهارت‌های اصلی زبان فارسی ترسیم کرده است. مقاله اول بر بحرانی‌ترین نقطه این نقشه (گفتار) متمرکز بود و مقاله حاضر با رویکرد تفاوت‌سنجی، موقعیت سه مهارت دیگر را نسبت به آن نقطه بحرانی روشن ساخته و الگوی حاکم بر کل این میدان را آشکار کرده است. تلفیق این دو مطالعه، پایه نظری محکمی برای طراحی نظام‌های سنجش ترکیبی آینده فراهم می‌آورد.

۵-۶. نتیجه‌گیری پایانی

هوش مصنوعی در محیط آموزش زبان فارسی، ابزاری قدرتمند اما «تقلیل‌گرا» است. تفاوت‌سنجی انجام شده ثابت کرد که این فناوری نمی‌تواند جایگزین شعور فرهنگی و درک تعاملیِ ممتحن انسانی در لایه‌های پیچیده استاندارد بنیاد سعدی شود. صیانت از اعتبار ملی آزمون‌های زبان فارسی ایجاب می‌کند که مرزهای صلاحیت ماشین به دقت ترسیم شده و از تسلیم روایی آزمون به الگوریتم‌های احتمال‌محور در سطوح عالی جلوگیری شود.

منابع

○ ابراهیمی، مجتبی، طالعی ، محمد حسین،(۱٤۰٤) ، شکاف سنجش گفتار در زبان فارسی: واکاوی تطبیقی قابلیت های هوش مصنوعی و استاندارد مرجع بنیاد سعدی، دوفصلنامه مطالعات بین المللی آموزش زبان فارسی، شماره(۲۰)

○ استاجی، معصومه.، صحرایی، رضا مراد.، و شهباز، منیره. (۱۴۰۳). ارزیابی روانی گفتار آزمون‌دهندگان فارسی‌آموز خارجی: نقش تکرار و خطای شروع در سطوح مختلف بسندگی زبان فارسی. فصلنامه اندازه‌گیری تربیتی، ۱۵،٥٦، ٦۹-۱۰۲. https://doi.org/۱۰/۲۲۰٥٤/jem.۳٤٥٤/۷۳۱۰٦/۲۰۲٤

○ اسمی، مهدی و برومند تمبکی، شهداد. (۱٤۰۳). بررسی تاثیر هوش مصنوعی (AI) بر یادگیری مهارت های زبانی در آموزش آنلاین.اولین کنفرانس بین المللی مطالعاتکاربردی در فرایندهای تعلیم و تربیت، بندرعباس. https://civilica.com/doc/۲۲٤۷۳٦۸

○ صبوری، سپهر.، و حاج ملک، محمد مهدی، (۱٤۰۲). استفاده از ظرفیت‌های هوش مصنوعی در آموزش تلفظ زبان‌های خارجی. نهمین کنفرانس بین‌المللی وب پژوهی.

○ صحرایی، مراد رضا و همکاران، (۱۳۹٥)، استاندارد مرجع بنیاد سعدی.

○ قاسمی، مهدی و برومند تمبکی، شهداد. (۱٤۰۳). بررسی تاثیر هوش مصنوعی (AI) بر یادگیری مهارت های زبانی در آموزش آنلاین. اولین کنفرانس بین المللی مطالعات کاربردی در فرایندهای تعلیم و تربیت، بندرعباس. https://civilica.com/doc/۲۲٤۷۳٦۸

○Acosta-Rivera, L. G., & Bedón Vaca, C. d. C. (۲۰۲٥). Impacto de la inteligencia artificial en el aprendizaje del inglés como lengua extranjera en estudiantes de bachillerato en Ecuador. Revista RedCA, ۲۲)۸). https://doi.org/۱۰,۳٦٦۷۷/redca.v۸i۲۲,۲٦۳٦۱

○Aryadoust, V., Zakaria, A., Lim, M. H., & Chen, C. (۲۰۲۱). An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. Frontiers in Psychology, ۱۹٤۱ ,۱۱. https://doi.org/۱۰,۳۳۸۹/fpsyg.۲۰۲۰,۰۱۹٤۱

○Bender, E. M., & Koller, A. (۲۰۲۰). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the ٥۸th Annual Meeting of the Association for Computational Linguistics (pp. ٥۱۹۸–٥۱۸٥). Association for Computational Linguistics. https://doi.org/۱۰,۱۸٦٥۳/v۲۰۲۰/۱.ACL-MAIN.٤٦۳

○Cicchetti, Domenic. (۱۹۹٤). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. Psychological Assessment. ۳٥۹۰,٦,٤,۲۸٤-۱۰٤۰/۱۰,۱۰۳۷ .۲۹۰-۲۸٤ .٦.

○Chapelle, C. A., & Voss, E. (Eds.). (۲۰۲۱). Validity argument in language testing: Case studies of validation research. Cambridge University Press. https://assets.cambridge.org/۸٤۰۲۲/۹۷۸۱۱۰۸٤/frontmatter/۹۷۸۱۱۰۸٤۸٤۰۲۲_frontmatter.pdf

○Dakhil, T. A., Karimi, F., Al-Jashami, R. A. U., & Ghabanchi, Z. G. (۳][۲][۱].(۲۰۲۵] The effect of artificial intelligence (AI)-mediated speaking assessment on speaking performance and willingness to communicate of Iraqi EFL learners. International Journal of Language Testing, [۲][۱].۱۸-۱ ,(۲)۱۵ ۱۰,۲۲۰۳٤/ijlt.۲۰۲٤,٤۸٦٥٦٤,۱۳۸۳

○Fakhrurriana, R., & Herdina, G. F. (۲۰۲٤). Assessment test of receptive skills: Listening and reading comprehension of recount text using AI platform (Atomic Learning). International Conference on Religion, Spirituality and Education (ICRSE), ۲۱۸-۲۱۱ ,۳. [https://sunankalijaga.org/prosiding/index.php/icrse/article/view/۱۲٤٤]

○Gomede, E. (۲۰۲٤, March ۲۷). Decoding the unspoken: Advancements in pragmatic analysis within natural language processing. The Modern Scientist. https://medium.com/the-modern-scientist/decoding-the-unspoken-advancements-in-pragmatic-analysis-within-natural-language-processing-d٦۳٥۸۷۷٥a۰۹۹

○Grab, M. O. (۲۰۲٥). Strategies for integrating Artificial Intelligence (AI) to improve and assess English oral communication skills. International Journal of Research in Education and Science, ٥٥۲-٥۳۳ ,(۳)۱۱. https://eric.ed.gov/?id=EJ۱٤۸۰۹۱۸

○Hınız, G. (۲۰۲٦). Using AI to enhance receptive foreign language skills. In AI›s Role in Language Learning and Communication (pp. ۷۱-۳٦). IGI Global. https://doi.org/٥-٥٦۸۱-۳۳۷۳-۸-۹۷۹/۱۰,٤۰۱۸.ch۰۰۳

○Huth, T. (۲۰۲۰). Testing interactional competence: Patterned yet dynamic aspects of L۲ interaction. Papers in Language Testing and Assessment, ,(۱)۹ ۲٥–۱.

○Likert, R. (۱۹۳۲). A technique for the measurement of attitudes. The Science Press. (Archives of Psychology, No. ۱٤۰)

○Ie, X., & Jaeger, T. F. (۲۰۲۰). Comparing non-native and native speech: Are L۲ productions more variable? The Journal of the Acoustical Society of America, ۳۳٤۷–۳۳۲۲ ,(٥)۱٤۷. https://doi.org/۱۰,۰۰۰۱۱٤۱/۱۰,۱۱۲۱

○Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (۲۰۲۳). Survey of hallucination in natural language generation. ACMComputingSurveys,۳۸–۱,(۱۲)٥٥.https://doi.org/۳٥۷۱۷۳۰/۱۰,۱۱٤٥

○Kane, M. T. (۲۰۱۳). Validating the interpretations and uses of test scores. Journal of Educational Measurement, ۷۳–۱ ,(۱)٥۰. https://doi.org/۱۰,۱۱۱۱/jedm.۱۲۰۰۰

○Lascano Pérez, M. J., & Altamirano Carvajal, S. P. (۲۰۲٥). Aplicación de la Inteligencia Artificial en las habilidades productivas y receptivas en la enseñanza del idioma inglés. DATEH,

○۱)۷).

○Liu, X. J., Wang, J., & Zou, B. (۲۰۲٥). Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback. Journal of

English for Academic Purposes, ۱۰۱۵۰۵ ,۷۵.
○López-Minotta, K. L., Chiappe, A., & Mella-Norambuena, J. (۲۰۲۵). Implementation of artificial intelligence to improve English oral expression. Multidisciplinary Journal of Educational Research (REMIE), ۷۱-۴۳ ,(۱)۱۵. https://doi.org/۱۰,۱۷۵۸۳/remie.۱۶۱۸۸
○Nigmatulina, I., Kew, T., & Samardžić, T. (۲۰۲۰). ASR for non-standardised languages with dialectal variation: The case of Swiss German. In M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, & Y. Scherrer (Eds.), Proceedings of the ۷th Workshop on NLP for Similar Languages, Varieties and Dialects (pp. –۱۵ ۲۴). International Committee on Computational Linguistics (ICCL). https://aclanthology.org/۲۰۲۰.vardial۱,۲-/
○Raud, N. (۲۰۲۵). Automatic assessment of L۲ interactional competency [Master's thesis, Aalto University].
○Saputra, A. E., & Widiastuty, H. (۲۰۲۴). The role of artificial intelligence in overcoming challenges in learning English speaking skills. International Seminar Universitas Tulungagung, ۱۷۱-۱۶۲ ,۶.
○Valle Torres, M. (۲۰۲۵). Aporte de la Inteligencia Artificial o plataforma como medio de aprendizaje de inglés para el desarrollo de habilidades lingüísticas en estudiantes de Ingeniería. Ciencia Latina Revista Científica Multidisciplinar, ۶)۹). https://doi.org/۱۰,۳۷۸۱۱/cl_rcm.v۹i۶,۲۱۴۶۵
○Voss, E., Cushing, S. T., Ockey, G. J., & Yan, X. (۲۰۲۳). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. Language Assessment Quarterly, -۵۲۰ ,(۵-۴)۲۰ ۵۳۲.https://doi.org/۱۵۴۳۴۳۰۳,۲۰۲۳,۲۲۸۸۲۵۶/۱۰,۱۰۸۰
○Yi, P., Xia, Y., & Long, Y. (۲۰۲۵). Irony detection, reasoning and understanding inzero-shotlearning.arXiv.https://doi.org/۱۰,۴۸۵۵۰/arXiv.۲۵۰۱,۱۶۸۸۴
○Zou, B., et al. (۲۰۲۴). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system. System, ۱۰۳۴۹۷ ,۱۲۶.