Assessing Speech Proficiency in Persian: A Comparative Analysis of Artificial Intelligence Capabilities and the Saadi Foundation Reference Standard

Abstract

The remarkable progress in Artificial Intelligence (AI), particularly in Large Language Models (LLMs), has significantly transformed the landscape of automated language proficiency assessment. While this technology has achieved success in evaluating formal and quantitative linguistic components, its adaptability to standardized frameworks that emphasize **communicative competence**, **content accuracy**, and **social interaction** remains an unresolved theoretical and technical challenge, particularly within the context of the Persian language. This study aims to delineate the epistemological and technical gaps of AI in conforming to the stringent requirements of the **Saadi Foundation Reference Standard** and to establish the boundaries of machine competence. This applied research employed a **critical documentary analysis** methodology. Functional statements and speaking skill descriptors across the seven levels of the Saadi Standard (ratified 2016) were extracted. Subsequently, these expectations were subjected to qualitative analysis through a **comparative matrix**, juxtaposing them against the inherent technical architecture and limitations of Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) algorithms in low-resource languages. The theoretical arguments were supported by analogical data on Word Error Rate (WER) in comparable L2 contexts and observational data from LLM interactions. The analysis reveals a significant **inverse correlation** between "skill-level complexity" and "machine-assessment validity." At Novice and Elementary levels, the machine demonstrates valid and substitutable performance due to the formal, quantitative, and static nature of the indicators (e.g., pronunciation accuracy, basic vocabulary). However, a deep performance gap emerges at the Intermediate (which the Saadi Standard separates into three sub-levels, unlike the CEFR) and Advanced levels. Findings indicate that the machine's inability to interpret **compensatory strategies**, its blindness to **cultural background knowledge**, its failure to detect **affective tone**, its **bias against non-standard accents**, and its deficiency in evaluating content accuracy (stemming from the phenomenon of **AI hallucination** at higher levels) pose a serious threat to **construct validity**. Exclusive reliance on machine assessment for high-stakes testing at advanced levels leads to a **negative washback effect**, reducing language education to mechanical, quantifiable patterns. Consequently, the study proposes a **Hierarchical Hybrid Assessment Model** (Machine for Form / Human for Meaning), featuring revised human-referral criteria based on sub-score inconsistency, as an optimal and scientifically-grounded solution.

Keywords: Speech Assessment, Artificial Intelligence (AI), Saadi Foundation Standard, Communicative Competence, Construct Validity, Washback Effect, Natural Language Processing (NLP), AI Hallucination.

شکاف سنجش گفتار در زبان فارسی: واکاوی تطبیقی قابلیت های هوش مصنوعی و استاندارد مرجع بنیاد سعدی

چکیده

پیشرفت‌های خیره‌کننده اخیر در هوش مصنوعی (AI) و مدل‌های زبانی بزرگ(LLMs) ، چشم‌انداز سنجش خودکار مهارت‌های زبانی را دگرگون کرده است. هرچند هوش مصنوعی در سنجش صوری زبان موفق بوده، اما انطباق آن با توانش ارتباطی و صحت محتوایی در زبان فارسی، مسئله‌ای حل‌نشده باقی مانده است. پژوهش حاضر با هدف تبیین شکاف‌های معرفت‌شناختی و فنی هوش مصنوعی در انطباق با الزامات دقیق «استاندارد مرجع بنیاد سعدی» و تعیین مرزهای صلاحیت ماشین انجام شده است.

پژوهش حاضر از نظر هدف کاربردی و از نظر ماهیت، «تحلیل اسنادی انتقادی» است. در این راستا، گزاره‌های عملکردی و توصیف‌گرهای مهارت گفتاری در ۷ سطح استاندارد بنیاد سعدی (مصوب ۱۳۹۵) استخراج گردید. سپس، طی یک «ماتریس تطبیقی»، این انتظارات در برابر معماری فنی و محدودیت‌های ذاتی الگوریتم‌های «بازشناسی گفتار» (ASR) و «پردازش زبان طبیعی» (NLP) در زبان‌های کم‌منابع قرار گرفت و مورد تحلیل کیفی واقع شد. این تحلیل، برای تقویت استدلال نظری، به داده‌های قیاسی «نرخ خطای کلمه» (WER) در زبان‌های مشابه «زبان دوم»(L2) و همچنین موارد مشاهده‌ای از تعامل با(LLM) ها اتکا نموده است.

تحلیل‌ها نشان می‌دهد میان «پیچیدگی سطح مهارت» و «روایی سنجش ماشین» رابطه معکوس معناداری وجود دارد. در سطوح نوآموز و مقدماتی، ماشین به دلیل ماهیت صوری، کمی و ایستای شاخص‌ها (نظیر صحت تلفظ و واژگان پایه)، عملکردی معتبر و جایگزین‌پذیر دارد. اما در سطوح میانی که در استاندارد سعدی برخلاف چارچوب مشترک اروپایی (CEFR) به سه زیرسطح پیش‌میانی، میانی و فوق‌میانی تفکیک شده است، و نیز در سطوح پیشرفته، هوش مصنوعی دچار شکاف عملکردی عمیق است. یافته‌ها نشان می دهد که ناتوانی ماشین در تفسیر «راهبردهای جبرانی»، کوری نسبت به «دانش پیشینه فرهنگی»، عدم تشخیص «بار عاطفی لحن»، سوگیری علیه لهجه‌های غیرمعیار و نارسایی در ارزیابی درستی محتوا، به دلیل پدیده «توهم هوش مصنوعی»در سطوح عالی، تهدیدی جدی برای «روایی سازه» محسوب می‌شود.

اتکای انحصاری به ماشین در آزمون‌های سطوح عالی، منجر به «اثر بازگشتی منفی» و تقلیل آموزش به الگوهای مکانیکی می‌شود. لذا الگوی «سنجش ترکیبی سلسله‌مراتبی» (ماشین برای سنجش فرم/ انسان برای سنجش معنا) با معیارهای اصلاح‌شده برای ارجاع انسانی (مبتنی بر ناسازگاری نمرات فرعی) به عنوان راهکار بهینه پیشنهاد می‌گردد.

واژگان کلیدی: سنجش گفتار، هوش مصنوعی، توانش ارتباطی، روایی سازه، اثر بازگشتی

۱ . مقدمه

۱-۱ . بیان مسأله

زیست‌بوم آموزش و سنجش زبان در دهه اخیر تحت تأثیر همگرایی فناوری‌های نوین اطلاعاتی، دستخوش دگردیسی شده است. فناوری‌های نوین نظیر «بازشناسی خودکار گفتار» [1] (ASR) و «پردازش زبان طبیعی» [2] (NLP) با وعده حذف محدودیت‌های انسانی (نظیر خستگی، ذهنیت‌گرایی، هزینه بالا و زمان‌بر بودن تصحیح)، نویدبخش الگوواره جدیدی در آزمون‌سازی زبان هستند. **هینگینز و همکاران (۲۰۱۱).** امروزه سامانه‌های هوشمند قادرند مهارت‌های نوشتاری و درک مطلب را با ضریب همبستگی قابل‌قبولی نسبت به مصححان انسانی ارزیابی کنند. **ژانگ و همکاران (۲۰۱۹).** با این حال، پرسش بنیادین و چالش‌برانگیز این است: آیا ماشینی که ماهیت عملکردش بر اساس «احتمالات آماری» و «تطبیق الگو» است، صلاحیت معرفت‌شناختی لازم برای سنجش پدیده‌ی پیچیده، پویا و چندلایه‌ای نظیر «توانش ارتباطی انسان» را در تمام سطوح داراست؟ **کین (۲۰۱۳)**

این مسأله در زبان فارسی به دلیل ویژگی‌های جامعه‌شناختی خاص این زبان، بسیار حادتر و پیچیده‌تر است. برخلاف زبان انگلیسی که از منابع عظیم داده‌ای برخوردار است، زبان فارسی در حوزه «گفتار زبان‌آموزان خارجی» جزو زبان‌های کم‌منابع محسوب می‌شود. این امر دقت مدل‌ها را به شدت کاهش می‌دهد **نیگماتولینا و همکاران (۲۰۲۰).** مدل‌های زبانی بزرگ [3] (LLM) ، عمدتاً بر روی داده‌های عمومی وب آموزش دیده‌اند و فاقد درک عمیق از ظرایف «گونه‌های محاوره‌ای»، «شکاف عمیق میان زبان رسمی و گفتاری» و «نظام پیچیده ادب و تعارفات» هستند. **قیومی (۲۰۱۹).** این فقر داده‌ای باعث می‌شود شکاف میان «خروجی ماشین» و «انتظارات استانداردهای بومی معتبر» عمیق‌تر شود. مسئله اصلی پژوهش حاضر، واکاوی دقیق و انتقادی این شکاف در پرتو «استاندارد مرجع بنیاد سعدی» است تا از بروز خطاهای راهبردی و تقلیل‌گرایانه در طراحی آزمون‌های ملی جلوگیری شود.

۱-۲ . اهمیت و ضرورت پژوهش

ضرورت انجام این پژوهش را می‌توان در سه محور کلیدی تبیین نمود:

الف) صیانت از روایی سازه و اعتبار استاندارد ملی:

بنیاد سعدی به عنوان متولی اصلی استانداردسازی آموزش زبان فارسی در جهان، چارچوبی هم‌تراز با استانداردهای جهانی [4] (CEFR) اما با بومی‌سازی دقیق تدوین کرده است. در آزمون‌سازی، مفهوم «روایی سازه» به این معناست که ابزار سنجش باید بتواند تمام ابعاد تعریف شده در نظریه را بسنجد. اگر ابزار سنجش (در اینجا هوش مصنوعی) نتواند مولفه‌هایی مثل «سیاق کلام»، «لحن عاطفی یا ادب» را که در سطوح عالی استاندارد سعدی تصریح شده، پردازش کند، آزمون دچار پدیده «کم‌نمایی سازه» شده و اعتبار گواهی‌نامه‌های صادره مخدوش می‌گردد.

ب) عدالت آموزشی در آزمون‌های سرنوشت‌ساز:

---

1. Automatic Speech Recognition
2. Natural Language Processing
3. **Large Language Models**
4. Common European Framework of Reference for Languages

آزمون‌های تعیین سطح و بسندگی معمولاً ماهیت سرنوشت‌ساز دارند. ناتوانی ماشین در تشخیص لهجه‌های غیرتهرانی یا عدم درک راهبردهای خلاقانه ارتباطی، می‌تواند منجر به نمره‌دهی ناعادلانه شود. ماشین ممکن است به زبان‌آموزی که صرفاً جملات صحیح را حفظ کرده نمره بالا بدهد و زبان‌آموزی را که با لهجه اما «ارتباط‌گر» و خلاق است، به دلیل انحراف از الگوی صوتی معیار، جریمه کند.

ج) پیشگیری از اثر بازگشتی منفی:

آزمون‌ها جهت‌دهنده‌ی آموزش هستند. اگر آزمون سطوح عالی به ماشین سپرده شود، کلاس‌های درس و محتوای آموزشی به سمت «نکات مورد پسند ماشین» منحرف می‌شوند. این امر باعث می‌شود مهارت‌های واقعی ارتباطی، تفکر انتقادی و تعاملات انسانی، قربانی الگوهای مکانیکی شوند و کلاس درس به کارگاه تربیت ربات‌های فارسی‌گو تبدیل شود.

۳-۱ .اهداف پژوهش

- تبیین دقیق مرزهای صلاحیت فنی هوش مصنوعی در انطباق با سطوح هفت‌گانه استاندارد بنیاد سعدی.
- شناسایی و تحلیل نقاط کور تکنولوژی در سنجش مولفه‌های «تعاملی»، «فرهنگی» «عاطفی» و ... مصرح در سند استاندارد.
- ارائه مدل اجرایی و عملیاتی «سنجش ترکیبی» برای تقسیم کار بهینه میان انسان و ماشین.

۲ .مبانی نظری و پیشینه پژوهش

۲-۱ .مبانی فنی: کالبدشکافی فرآیند سنجش خودکار(ALA)

برای درک دقیق چرایی محدودیت‌ها، لازم است مکانیزم عملکرد «سنجش خودکار زبان» تشریح شود. این فرآیند معمولاً شامل سه مرحله متوالی است که هر یک چالش‌های خاص خود را دارد:

بازشناسی گفتار: در این مرحله، سیگنال صوتی به متن تبدیل می‌شود. چالش اصلی در زبان فارسی، «تنوع لهجه» و «سروصدای محیط» است. مدل‌های سنجش خودکار معمولاً با گفتار معیار گویندگان بومی آموزش دیده‌اند و در مواجهه با «گونه‌های گفتاری زبان‌آموزان» که مملو از خطاهای تلفظی و نوای گفتار غیربومی است، دچار نرخ خطای کلمه» [5](WER) بالا می‌شوند.

برای مثال، چالش‌های ساختاری در زبان‌های دیگری که از الگوی «تنوع منطقه‌ای قوی» و «فقدان ارتوگرافی استاندارد» رنج می‌برند، محدودیت‌های ذاتی این فناوری را برجسته می‌سازد. در مطالعه‌ای بر روی سنجش خودکار زبان آلمانی سوئیسی (یک محیط چند لهجه‌ای و غیر استاندارد)، بهترین سیستم‌های آموزش‌دیده بر نوشتار استاندارد نیز به نرخ خطای کلمه بالایی معادل ۲۹.۳۹٪ دست یافتند. نیگماتولینا و همکاران ،(۲۰۲۰). این نرخ بالای خطا نشان می‌دهد که در زبان‌هایی با تنوع بومی و کمبود منابع، چالش نرخ خطای کلمه نه یک مشکل قابل رفع، بلکه یک محدودیت پایه‌ای در معماری فعلی است.

---

[5]. **Word Error Rate**

افزون بر این، پژوهش‌ها نشان می‌دهند که تفاوت لهجه، سوگیری‌های درون‌تری را در سیستم‌های پردازش صوتی ایجاد می‌کند. به عنوان مثال، در مطالعات مربوط به تمایز و تطبیق هویت صوتی[6] ، تفاوت لهجه حتی زمانی که هویت‌ها یکسان بودند، به طور خاص باعث افزایش سوگیری (حدود ۱۰٪) برای قضاوت یکسان بودن افراد مختلف شد. **سانتوس و همکاران(۲۰۲۵).** این نتایج نشان می‌دهند که سوگیری لهجه فراتر از یک خطای صرفاً واژگانی است و با این ایده که سیستم‌های صوتی (همانند انسان‌ها) «سوگیری دارند که تصور می‌کنند افراد معمولاً با یک لهجه صحبت می‌کنند، سازگار است. این سوگیری ادراکی می‌تواند به سوگیری الگوریتمی در مدل‌های ASR منجر شود..

استخراج ویژگی‌ها: در این مرحله، ویژگی‌های آکوستیک (مانند زیروبمی صدا، طول مکث‌ها، سرعت گفتار) و ویژگی‌های متنی (مانند پیچیدگی دستوری، تنوع واژگان) استخراج می‌شوند.»

نمره‌دهی: مدل‌های آماری (نظیر شبکه‌های عصبی عمیق)، این ویژگی‌ها را با داده‌های آموزشی مقایسه می‌کنند. نقد اصلی در این مرحله نهفته است: ماشین «معنی را در بافت اجتماعی نمی‌فهمد، بلکه «الگو را با الگوهای موجود تطبیق می‌دهد. افزون بر این، در معماری مدل‌های زبانی بزرگ (LLM)، پدیده «توهم هوش مصنوعی»[7] ، به عنوان یک نقص ذاتی مطرح است. توهم، تولید اطلاعاتی است که از نظر زبانی صحیح، اما از نظر محتوایی نادرست یا بی‌اساس است. بررسی‌های جامع نشان می‌دهد که علیرغم پیشرفت‌های فراوان در تولید زبان روان، چالش توهم، یک مشکل اساسی و حل‌نشده در معماری مبتنی بر یادگیری عمیق است که منجر به کاهش عملکرد سیستم‌ها در وظایف سطح بالا می‌شود. جی و همکاران، (۲۰۲۳).این نارسایی‌های فنی (نرخ خطای بالا در لهجه‌ها و پدیده توهم)، صرفاً محدودیت‌های ابزاری نیستند، بلکه طبق چارچوب نظری «کین»، شواهدی هستند که زنجیره استدلالِ روایی را در آزمون‌های بسندگی زبان فارسی تضعیف می‌کنند؛ امری که لزوم بازنگری در اعتماد مطلق به خروجی ماشین را دوچندان می‌سازد. این نقص نشان می‌دهد که ماشین در بنیادین‌ترین سطح، فاقد صلاحیت ارزیابی «صحت محتوایی» و «استدلال منطقی» در گفتار انسان است.

۲-۲ . استاندارد مرجع آموزش زبان فارسی (بنیاد سعدی)

این استاندارد که توسط صحرایی و همکاران (۱۳۹۵) تدوین شده است، جامع‌ترین سند بالادستی در حوزه آموزش زبان فارسی به غیرفارسی‌زبانان (آزفا) است. طبق جداول موجود در صفحات ۲۹ تا ۳۲ این سند، مسیر یادگیری زبان فارسی در ۷ سطح اصلی (نوآموز، مقدماتی، پیش‌میانی، میانی، فوق‌میانی، پیشرفته، ماهر) ترسیم شده است. یک وجه تمایز مهم این استاندارد با چارچوب‌هایی مانند چارچوب مشترک اروپایی، تفکیک دقیق‌تر سطح میانی به سه زیرسطح مجزا است که نشان‌دهنده حساسیت آن به مراحل یادگیری است. این تفکیک، نه صرفاً یک تقسیم‌بندی آموزشی، بلکه بیانگر یک تفاوت کیفی در سازه‌ی ارتباطی است که بر افزایش پیچیدگی وظایف از کاربردی در پایین ترین سطح میانی به انتزاعی در بالاترین سطح میانی تأکید دارد. منطق حاکم بر این استاندارد، حرکت تدریجی از «تولید فرم‌های بسته و کلیشه‌ای» در سطوح پایه به سمت «مدیریت تعامل پویا و تولید معنا» در سطوح بالاست. این سند در سطوح پیشرفته و ماهر صراحتاً بر مولفه‌هایی نظیر «درک زبان بینابینی»، «مدیریت راهبردهای ارتباطی»، «درک طنز و کنایه» و «رعایت سیاق‌های فرهنگی» تأکید دارد؛ مولفه‌هایی که ذاتاً کیفی، وابسته به بافت و انسانی هستند.

---

6. Voice Discrimination
7. AI Hallucination

۲-۳ .چارچوب نظری روایی: رویکرد استدلالی در آزمون‌سازی زبان

برای تحلیل علمی اعتبار سنجش مبتنی بر هوش مصنوعی، این پژوهش از چارچوب اعتبار مبتنی بر استدلال به عنوان چارچوب نظری اصلی بهره می‌گیرد.چاپل و واس، (۲۰۲۱). در این الگوواره، اعتبار یک آزمون نه به عنوان یک ویژگی ذاتی، بلکه به عنوان میزان پشتیبانی شواهد از ادعاهای خاصی تعریف می‌شود که درباره تفسیر و استفاده مورد نظر از نمرات آن آزمون مطرح می‌گردد. کین، (۲۰۱۳)

پژوهش حاضر، با الهام از این چارچوب، استدلالی را مبنی بر وجود نقصان در زنجیره استدلالی اعتبار، زمانی که هوش مصنوعی به عنوان ارزیاب انحصاری در سطوح میانی و عالی به کار می‌رود، ارائه می‌کند. شواهد تحلیل ما مستقیماً بر شکاف‌های معرفت‌شناختی و فنی‌ای متمرکز است که استنباط «برابری صلاحیت ارزیابی ماشین با انسان» را تضعیف می‌کنند.

۲-۴ .پیشینه پژوهش: مرور تحلیلی و انتقادی

مرور ادبیات پژوهشی در حوزه سنجش خودکار گفتار نشان‌دهنده یک دگردیسی از مدل‌های صرفاً آکوستیک به سمت مدل‌های زبانی پیچیده است. پژوهش‌های انجام شده را می‌توان در سه محور دسته‌بندی نمود:

الف) کارآمدی هوش مصنوعی در تکالیف کنترل‌شده و صوری:

در بخش مطالعاتی که بر جنبه‌های کمی و صوری زبان تمرکز دارند، زو و همکاران (۲۰۲۴). در یک مطالعه با روش آمیخته بر روی ۳۶۶ زبان‌آموز در ۵ دانشگاه، دریافتند که سامانه‌های هوشمند نظیر(EAP Talk) در بهبود مهارت‌های روخوانی و دقت دستوری بسیار موثرند، هرچند در ارائه بازخوردهای اصلاحیِ عمیق محدودیت دارند. در همین راستا، صبوری و حاج‌ملک (۱۴۰۲) در پژوهشی نیمه‌تجربی با بررسی ۹۳ واژه هدف طی ۱۸ جلسه آموزشی، نشان دادند که فناوری بازشناسی گفتار می‌تواند با دقت بالایی پیشرفت مهارت تلفظ زبان‌آموزان را ردیابی کند. همچنین، قاسمی و برومند (۱۴۰۳) با استفاده از ابزارهایی نظیر(ChatGPT و Duolingo) در یک پیمایش مبتنی بر پرسشنامه و مصاحبه با معلمان، تأکید کردند که شخصی‌سازی یادگیری از طریق هوش مصنوعی منجر به افزایش خودکارآمدی در مهارت مکالمه می‌شود. مظهرپور و سیدکلان (۱۴۰۳) نیز با روش سنتزپژوهی و تحلیل ۱۵ مقاله معتبر (با ضریب کاپای ۰.۸۸)، تأیید کردند که چت‌بات‌ها ابزارهای عملیاتی موثری در آموزش زبان هستند، مشروط بر اینکه معلمان سواد دیجیتال لازم برای هدایت این فرایند را داشته باشند.

ب) چالش‌های تعامل، بافت‌مندی و زبان‌های کم‌منابع:

در مقابل، لایه دوم پژوهش‌ها بر محدودیت‌های ساختاری ماشین در مواجهه با پیچیدگی‌های زبانی تمرکز دارند. لیو و همکاران (۲۰۲۵) در ارزیابی ابزار (EAP Talk)بر روی ۶۴ دانشجو، با مقایسه نمرات هوش مصنوعی و ۵ داور انسانی، دریافتند که میان «تکالیف کنترل‌شده» (روخوانی) و «تکالیف آزاد» (ارائه) شکاف روایی وجود دارد؛ به طوری که ماشین در سنجش گفتار خودجوش عملکرد ضعیفی دارد. راد (۲۰۲۵) در پایان‌نامه خود در دانشگاه آلتو، با نقد مدل‌های مبتنی بر ترنسفورمر، استدلال کرد که ماشین‌های فعلی به دلیل متن‌محور بودن، در سنجش «توانش تعاملی» که مستلزم ویژگی‌های واکنشی و چندوجهی است، ناتوان هستند. هوت (۲۰۲۰) نیز با استفاده از تحلیل گفتمان (CA)، هشدار داد که هوش مصنوعی نمی‌تواند ماهیت پویا و غیرمنضبط تعاملات انسانی را به قواعد صوری تقلیل دهد .

در حوزه زبان‌های کم‌منابع و لهجه‌ها، نیگماتولینا و همکاران (۲۰۲۰) در بررسی ۱۴ گویش آلمانی سوئیسی با ابزار(Kaldi) ، نشان دادند که فقدان ارتوگرافی استاندارد منجر به نرخ خطای کلمه ۲۹ درصدی می‌شود. همچنین، سانتوس و همکاران (۲۰۲۵) در یک مطالعه روان‌شناختی با استفاده از تکنیک شبیه‌سازی صدا اثبات کردند که تفاوت لهجه (مانند لهجه چینی یا لهستانی) باعث ایجاد ۱۰ درصد سوگیری منفی در تشخیص هویت و کیفیت گفتار توسط مدل‌های صوتی می‌شود. در بافت زبان فارسی، قیومی (۲۰۱۹) با استفاده از روش تعبیه معنایی بر روی مجموعه‌ای از ۲۰ واژه هدف در ۱۰۰ بافت جمله‌ای، نشان داد که اگرچه ماشین در تشخیص حس واژگان پیشرفت داشته، اما همچنان وابستگی شدیدی به حجم داده‌های نشانه‌گذاری شده دارد.

ج) چارچوب‌های روایی و اخلاقی:

در لایه نظری، کین (۲۰۱۳) با ارائه رویکرد استدلالی، تأکید کرد که اعتبار یک آزمون نه در نمره، بلکه در «تعبیر و استفاده» از آن نهفته است و هرچه ادعای آزمون (مانند سنجش سطح ماهر) بزرگتر باشد، شواهد بیشتری برای روایی لازم است. چاپل و واس (۲۰۲۱) در کتاب مرجع خود با بررسی مطالعات موردی متعدد، نشان دادند که فناوری باید در خدمت ساختار استدلالی آزمون باشد نه برعکس. از منظر عدالت الگوریتمی، کوردزاده و قاسم‌آقایی (۲۰۲۲) با تلفیق ادبیات موجود، مدلی را ارائه دادند که نشان می‌دهد سوگیری ماشین مستقیماً بر ادراک عدالت و پذیرش سیستم توسط کاربر تأثیر می‌گذارد. در نهایت، رشیدی (۱۴۰۴) در یک مطالعه تحلیلی-توصیفی با مرور نظام‌مند منابع فارسی، تأکید کرد که در ارزشیابی توصیفی خودکار، روایی سازه در غیاب بازبینی انسانی به شدت آسیب‌پذیر است.

جمع‌بندی پیشینه‌ها نشان می‌دهد که اگرچه مطالعات داخلی (مانند صبوری و حاج‌ملک، ۱۴۰۲؛ قاسمی و برومند، ۱۴۰۳) بر جنبه‌های انگیزشی و بهبود تلفظ تمرکز داشته‌اند، اما خلأ بزرگی در تبیین «عدم انطباقِ سازه» میان الگوریتم‌های هوش مصنوعی و «توصیف‌گرهای کیفیِ استاندارد مرجع بنیاد سعدی» (به‌ویژه در سطوح میانی به بالا) وجود دارد. پژوهش حاضر با هدف پر کردن این شکاف، برای نخستین بار به تقابل عملکردیِ کدهای فنی و استانداردهای بومی آزفا می‌پردازد.

۳ .روش‌شناسی پژوهش

۳-۱ .طرح پژوهش

پژوهش حاضر از نظر هدف، کاربردی-توسعه‌ای و از نظر ماهیت، «کیفی» است. روش اجرای پژوهش، «تحلیل اسنادی تطبیقی-انتقادی» انتخاب شده است. دلیل انتخاب این روش آن است که هدف پژوهش، تست کردن یک نرم‌افزار خاص (که نتایجش تاریخ‌مصرف دارد) نیست، بلکه نقد «معرفت‌شناختی» امکان جایگزینی ماشین با انسان بر اساس تحلیل ویژگی‌های ذاتی هر دو سیستم است.

۳-۲ .جامعه و نمونه آماری

جامعه تحلیلی این پژوهش شامل دو دسته سند است:

سند معیار: متن کامل «استاندارد مرجع آموزش زبان فارسی در جهان» (بنیاد سعدی، ۱۳۹۵) به عنوان ملاک و معیار مطلوب.

اسناد فنی: مستندات فنی و مقالات مروری مرتبط با معماری مدل‌های زبانی بزرگ (LLMs) و سیستم‌های بازشناسی گفتار (ASR) که نماینده نسل متداول فعلی فناوری در حوزه پردازش زبان هستند، به عنوان اسناد فنی مورد تحلیل قرار گرفتند. با توجه به این امر که زبان فارسی

در حوزه گفتار زبان‌آموزان (L2 Speech) جزو زبان‌های کم‌منابع محسوب می‌شود، سنجش نرخ خطای کلمه (WER) این مدل‌ها بر روی پیکره‌های بزرگ L2 به صورت عمومی در دسترس نیست. بنابراین، تحلیل حاضر، نه بر اساس نرخ‌های عددی ناموجود، بلکه بر مبنای نقد ماهوی معماری غالب در مدل‌های بازشناسی خودکار گفتار و «پردازش زبان طبیعی» استوار است. ادعای اصلی پژوهش مبنی بر «رابطه معکوس معنادار میان پیچیدگی مهارت و روایی ماشین» مستند بر یک اصل معرفت‌شناختی در هوش مصنوعی است: ماشین در «حل مسائل بسته و دارای پاسخ معین» (مانند صحت واج) عملکرد عالی دارد، اما در «حل مسائل باز و وابسته به بافت اجتماعی» (مانند طنز و استراتژی) ذاتاً ناتوان است. برای تقویت بنیان استدلال نظری، از شواهد قیاسی مطالعات زبان‌های کم‌منابع مشابه (نظیر سامی و هندواروپایی) در مورد نرخ خطای کلمه (WER) بر روی پیکره‌های L2 نیز استفاده شده است. این مدل‌ها که عمدتاً بر داده‌های بومی و استاندارد آموزش دیده‌اند، در مواجهه با ویژگی‌های منحصربه‌فرد گفتار زبان‌آموزان (نظیر انحرافات آکوستیکی و دستوری غیربومی) دچار محدودیت ذاتی و تئوریک می‌شوند و نقد ما بر همین مبنای نظری قرار دارد. شایان ذکر است که رویکرد این پژوهش، «ارزیابی عملکرد» یک سامانه خاص نیست، بلکه «تحلیل امکان‌سنجی معرفت‌شناختی» است. تمرکز بر محدودیت‌های ذاتی معماری این مدل‌ها (الگوی آماری، فقدان تجسد، آموزش بر داده‌های بومی) به ما اجازه می‌دهد تا به طور اصولی استدلال کنیم که حتی پیشرفته‌ترین نمونه‌های فعلی و آینده نزدیک این فناوری، تا زمانی که این معماری را تغییر ندهند، قادر به پر کردن شکاف‌های شناسایی‌شده نخواهند بود. این تحلیل، در حکم شناسایی موانع نظری پیش از اتلاف منابع برای آزمون‌های عملی پرهزینه است.

۳-۳ .ابزار و فرایند تحلیل داده‌ها

برای انجام این تحلیل، یک ماتریس تطبیقی طراحی گردید.

بّعد اول ماتریس (انتظارات): شامل گزاره‌های عملکردی و توصیف‌گرهای دقیق مهارت گفتاری است که عیناً از فصل دوم سند استاندارد مرجع بنیاد سعدی (صفحات ۳۳ تا ۵۲) استخراج شده است.

بعد دوم ماتریس (قابلیت‌ها): شامل ظرفیت‌های فنی و محدودیت‌های ذاتی مدل‌های هوش مصنوعی، با تمرکز ویژه بر چالش‌های آن‌ها در زبان‌های کم‌منابع. نقاط تقاطع این ماتریس بر اساس سه شاخص کلیدی مورد تحلیل قرار گرفت:

- **سنجش‌پذیری کمی:** (آیا ویژگی مورد نظر به داده عددی عینی و پایدار تبدیل می‌شود؟)
- **وابستگی به بافت:** (درجه نیازمندی ویژگی به دانش فرهنگی، موقعیتی و دانش جهان برای تفسیر صحیح. به عنوان مثال، برای تحلیل گزاره «استفاده از طنز» در استاندارد، از این شاخص استفاده شد.)
- **دینامیک‌های تعاملی:** (آیا سنجش این ویژگی نیاز به واکنش لحظه‌ای، چندطرفه و مبتنی بر فهم متقابل دارد؟)

این تحلیل تطبیقی، ما را به تفکیک محدودیت‌های شناسایی‌شده به دو دسته‌ی مجزا سوق داد: ۱) محدودیت‌های فنی و منابعی (قابل رفع با پیشرفت‌های آتی) ۲) محدودیت‌های ذاتی و معرفت‌شناختی (نامحتمل برای رفع بدون تحول الگوواره در معماری

۴ . یافته‌ها

۴-۱ .تحلیل سطوح نوآموز و مقدماتی: انطباق کامل و قلمرو اقتدار ماشین.

در سطوح اولیه، استاندارد بنیاد سعدی انتظاراتی را مطرح می‌کند که کاملاً با قابلیت‌های «تشخیص الگو» در هوش مصنوعی همخوانی دارد.

استناد به سند: طبق جدول شماره ۲-۶ (صفحه ۳۳ سند استاندارد)، در توصیف مهارت گفتاری سطح نوآموز آمده است: «**تولیدات گفتاری زبان‌آموز در حد واژه‌های منفرد و یا قالب‌های روزمره است... زبان‌آموز تلاش زیادی می‌کند تا واج‌ها را به درستی تولید کند.**» همچنین در جدول ۲-۷ (صفحه ۳۶) برای سطح مقدماتی ذکر شده: «**زبان‌آموز جملات و اصطلاحات ساده را برای توصیف محل زندگی به کار می‌برد... گفتار او اغلب از بسط موارد یادگرفته شده و عبارات دم‌دستی شکل می‌گیرد.**»

تحلیل تطبیقی: در این سطوح، ورودی‌ها دارای «آنتروپی پایین»[8]، «بسته» و «قابل پیش‌بینی» هستند. سیستم‌های بازشناسی خودکار گفتار با دسترسی به مدل‌های آکوستیک استاندارد، قادرند انحرافات تلفظی در سطح واج را با دقتی فراتر از گوش انسان تشخیص دهند. همچنین الگوریتم‌های پردازش زبان طبیعی می‌توانند به راحتی واژگان استفاده شده را با دیتابیس واژگان سطح مقدماتی مقایسه کنند.

نتیجه: در سطوح نوآموز و مقدماتی، هوش مصنوعی نه‌تنها جایگزین معتبری است، بلکه به دلیل عینیت، سرعت، حذف خستگی و کاهش احتمالی سوگیری‌های انسانی، می‌تواند به عنوان ابزار سنجش بهینه‌ای محسوب شود.

توجه به آینده‌پژوهی: در این سطوح، محدودیت‌های احتمالی فعلی (نظیر نرخ خطای کلمه بالا در مواجهه با برخی لهجه‌ها یا سروصدای محیط)، ماهیتی فنی و منابعی دارند. این چالش‌ها با پیشرفت‌های آتی در معماری‌های ASR به‌ویژه با جمع‌آوری پیکره داده‌های گفتار L2 فارسی و مدل‌های چندوجهی قابل رفع و اصلاح کامل خواهند بود.

۲-۴ .تحلیل سطوح میانی(B): ظهور تعارضات در راهبردهای ارتباطی و تعامل

با ورود به سطح میانی، الگوواره آزمون طبق سند بنیاد سعدی از «تولید فرم» به «مدیریت تعامل» تغییر می‌کند و شکاف‌های جدی پدیدار می‌شوند. لازم به تأکید است که استاندارد سعدی، سطح میانی را به سه زیرسطح «پیش‌میانی»، «میانی» و «فوق‌میانی» تفکیک کرده که نشان‌دهنده انتظار پیشرونده از پیچیدگی زبانی است، به‌طوریکه در لایه بالاترین سطح از این بخش انتظار «بحث انتزاعی» و «فرضیه‌سازی» از زبان‌آموز می‌رود. این تفکیک سه‌گانه نشان‌دهنده یک تفاوت کیفی در سازه‌ی ارتباطی است که بر افزایش پیچیدگی وظایف از کاربردی در« پایین ترین سطح میانی(B1) » به انتزاعی در «بالاترین سطح میانی(B3) »تاکید دارد، به‌طوری که سنجش توانایی درک منطق استدلال در سطح گفتمان در بالاترین سطح میانی برای ماشین به مراتب دشوارتر از سطوح پایین‌تر است.

استناد به سند: طبق جدول شماره ۲-۹ (صفحه ۴۲ سند استاندارد)، یکی از اهداف کلیدی سطح میانی چنین است: «**زبان‌آموز توانایی خوبی در جبران عدم درک برخی ساخت‌ها... به وسیله استفاده درست از راهبردهای ارتباطی مانند تفسیر، بیان غیرمستقیم و تشریح از خود نشان می‌دهد.**»

۱. تحلیل عینی شکاف (سناریوی راهبرد جبرانی):

[8]. **Low entropy**

فرض کنید یک زبان‌آموز در آزمون سطح پیش‌میانی کلمه «چتر» را فراموش می‌کند. او به جای سکوت، می‌گوید: «آن وسیله‌ای که وقتی باران می‌آید روی سرمان می‌گیریم تا خیس نشویم»

دیدگاه انسان (طبق استاندارد): این رفتار نشان‌دهنده «توانش راهبردی» است. ارزیاب انسانی، مطابق با توصیف جدول ۲-۸ (صفحه ۳۸) استاندارد که به «دشواری در بازیابی» و لزوم حفظ ارتباط اشاره دارد، متوجه می‌شود که داوطلب توانسته است با استفاده از «توصیف» ارتباط را حفظ کند و این یک نقطه قوت محسوب می‌شود.

دیدگاه ماشین (الگوریتم فعلی): نارسایی ماشین در این سناریو، لزوماً در قابلیت درک معنایی، که مدل‌های LLM پیشرفته ممکن است بفهمند («وسیله...» همان «چتر» است) نیست بلکه در معماری مدل‌های نمره‌دهی تفکیکی[9] نهفته است. الگوریتم‌های سیستم بازشناسی خودکار گفتار و پردازش زبان طبیعی عموماً مکث‌ها، طولانی شدن کلام و عدم استفاده از واژه هدف را ثبت می‌کنند. این فرآیند خلاقانه (راهبرد جبرانی) باعث افزایش طول کلام و کاهش روانی کلام در واحد زمان می‌شود و عدم استفاده از واژه هدف سطح بالا را در پی دارد. تا زمانی که مدل‌های نمره‌دهی، مولفه‌ی جدیدی به نام «اثربخشی راهبردی ارتباطی» را به عنوان یک نمره مثبت با وزن مناسب محاسبه نکنند، الگوریتم‌های فعلی، این افزایش روانی را به حساب تسلط نمی‌گذارند، بلکه طولانی شدن کلام و عدم استفاده از واژه کلیدی را در بخش سنجش واژگان جریمه می‌کنند و این فرآیند خلاقانه را عمدتاً به عنوان «عدم تسلط و نشانه‌ای از «دایره لغت پایین» تفسیر کرده و نمره کسر می‌کنند.

نتیجه: در اینجا ماشین دقیقاً رفتاری را جریمه می‌کند که استاندارد آن را به عنوان یک شایستگی کلیدی تشویق کرده است.

۲. شکاف در نقش ممتحن به عنوان تسهیل گر و شریک تعاملی:

استناد به سند: در توصیف سطوح پایه (به ویژه مقدماتی)، استاندارد سعدی تصریح می‌کند که زبان‌آموز می‌تواند ارتباط برقرار کند «**مشروط به این که مخاطب وی حاضر باشد جملاتش را آرام ادا کند و حتی آنها را به ترکیبی دیگر تکرار نماید و به زبان‌آموز در ساختن جملات کمک کند**» (جدول ۷-۲، صفحه ۳۶). این بیان، نقش ارزیاب انسانی را به صراحت به عنوان یک مخاطب تنظیم‌کننده و تسهیل‌گر تعامل تعریف می‌کند.

تحلیل شکاف :هوش مصنوعی فاقد توانایی ذاتی برای ایفای این نقش پویا و مبتنی بر درک لحظه‌ای است. یک سیستم ارزیابی خودکار نمی‌تواند سرعت گفتار خود را به صورت تطبیقی کاهش دهد، ساختار جمله زبان‌آموز را به شکل طبیعی بازسازی کند، یا با دادن سرنخ‌های فرازبانی (مانند تأیید کلامی بله، متوجهم) فضای روان‌شناختی امن برای تولید زبان ایجاد نماید. این ناتوانی، هسته «**سازه تعاملی**»را در سطوح پایه، که پیش‌نیاز اعتماد به نفس برای سطوح بالاتر است، نادیده می‌گیرد و ارزیابی را به یک فرآیند غیرواقعی و ماشینی تقلیل می‌دهد.

9. Analytic Scoring Models

چالش تعامل و زبان بینابینی: در صفحه ۳۹ سند ذکر شده که **زبان‌آموز باید بتواند با سایر غیرفارسی‌زبانان ارتباط برقرار کند.** سنجش این مهارت نیازمند سناریوی «چندگوینده» و ارزیابی «مذاکره معنا »[10] است. هوش مصنوعی در تفکیک منابع صوتی و درک گفتگوی چندنفره‌ای که در آن هر دو گوینده دارای لهجه و خطاهای دستوری هستند، دچار خطای پردازشی بالا می‌شود و قادر به تشخیص کیفیت تعامل نیست.

۳-۴ .تحلیل سطوح پیشرفته و ماهر(C) : بن‌بست‌های معرفت‌شناختی، فرهنگی و سوگیری

درسطوح عالی، زبان دیگر صرفاً ابزار انتقال پیام نیست، بلکه ابزاری برای مدیریت موقعیت‌های پیچیده اجتماعی، فرهنگی و عاطفی است.

استناد به سند: طبق جدول ۲-۱۲ (صفحه ۵۱ سند استاندارد)، انتظارات شامل موارد زیر است: «**توانایی درک اطلاعات ضمنی و استنباطی، لحن و دیدگاه‌ها... استفاده از گفتمان اغنایی... و درک ظرایف و نکات دقیقی که گوینده به کار می‌برد (مانند طنز و کنایه).**»

۱. شکاف فقدان دانش پیشینه و کوری فرهنگی:

مفهوم «زبان‌های پربافت» که نخست توسط ادوارد تی هال مطرح شد (به نقل از کیتلر، ریگل و مکی‌نون، ۲۰۱۱)، بر این اصل استوار است که درک کامل پیام در چنین زبان‌هایی مستلزم دانش پیش‌زمینه گسترده مشترک، توجه به عناصر موقعیتی و آشنایی عمیق با بافت فرهنگی است. زبان فارسی، با توجه به غنای ادبی و فرهنگی خود، واجد چنین ویژگی‌هایی است. این در حالی است که مدل‌های زبانی بزرگ فاقد «تجسد» در فرهنگ و این دانش پیشینهٔ انسانی هستند (بندر و کولر، ۲۰۲۰). برای نمونه، عبارت تعارفی «قدمتان روی چشم» ممکن است توسط یک مدل زبانی صرفاً به صورت تحت‌اللفظی (آناتومیک) تفسیر شود یا به عنوان گزاره‌ای بی‌معنی تلقی گردد، در حالی که در استاندارد ارزیابی بنیاد سعدی، استفاده بجا و طبیعی از چنین تعارفات فرهنگی، نشانه‌ای از تسلط زبانی در سطح عالی محسوب می‌شود.

مشاهدهٔ تجربی: این شکاف در آزمایشی عملی با یک مدل زبانی بزرگ (LLM) تأیید شد. در یک تعامل ابتدایی، هنگام استفاده از اصطلاحات کنایی مانند «دسته گل به آب دادم» یا «دلم را به دریا زدم»، پاسخ اولیه مدل، تفسیر تحت‌اللفظی و لغوی بود. تنها پس از تعامل متوالی و ارائه نمونه‌های راهنما (عملی شبیه یادگیری کم‌نمونه)، مدل موفق به تطبیق الگو و درک معنای ثانویه کنایی گردید. این مشاهده نشان می‌دهد که ارزیابی خودکار ماشین از درک کنایه، فاقد روایی سازه کافی است، زیرا توانایی آن وابسته به کمک و راهنمایی مستقیم کاربر انسانی است و از درک مستقل و مبتنی بر تجربه فرهنگی برخوردار نیست.

۲.شکاف ناتوانی در سنجش «صحتی» و خطر «توهم هوش مصنوعی»:

این شکاف حیاتی، مرتبط با «روایی‌فروشی» است. در سطوح پیشرفته و ماهر« C1 وC2 » که استانداردی انتظار استدلال قانع‌کننده و تولید گفتمان نوآورانه و انتزاعی را دارد، ناتوانی ماشینی در تشخیص حقیقت موضوعی به عنوان یک خطر جدی برای «روایی سازه» خود را نشان می‌دهد. هوش مصنوعی، که خود مستعد پدیده توهم است و صرفاً بر احتمالات آماری استوار است،(جی و همکاران، ۲۰۲۳)، فاقد صلاحیت

[10]. Negotiation of Meaning

ذاتی برای «ارزیابی اعتباری»[11] و «استحکام منطقی» در استدلال‌های انسان است. «اگر زبان‌آموز، یک استدلال غیرمنطقی یا اطلاعات غلط (که خود می‌تواند حاصل «توهم» منابع دیگر باشد) را با روانی بالا بیان کند، ماشین صرفاً بر فرم نمره می‌دهد. این امر به تقلیل معنا می‌انجامد و به گفتاری نمره بالا اعطا می‌کند که حاوی نظریات سست یا بی‌اساس است. این نقص دوگانه« توهم در تولید مدل و توهم در ارزیابی توسط همان مدل» ریشه در معماری یکسان آن‌ها دارد و خطری بنیادین برای اعتبار آزمون در سطوح عالی محسوب می‌شود.

۳. شکاف کوری عاطفی[12] و ناتوانی در سنجش «گفت‌وشنود تعاملی»[13]:

در سطح ماهر، آزمون‌گر انسانی ممکن است عمداً کلام داوطلب را قطع کند یا با او مخالفت تند کند تا «تاب‌آوری زبانی» و توانایی مدیریت گفتمان را بسنجد. این در حالی است که استاندارد بنیاد سعدی ماهیت ارتباط در این سطوح را «دیالوژیک» یا گفت‌وشنود می‌داند. هوش مصنوعی‌های فعلی عمدتاً «تک‌گویی و منفعلانه و نوبت‌گرا»[14] هستند؛ آن‌ها منتظر می‌مانند تا کلام تمام شود و نمی‌توانند واکنش‌های عاطفی، استرس یا تسلط روانی داوطلب را در هنگام یک تعامل پویا و چالشی بسنجند. همچنین تحلیل ویژگی‌های فرازبانی برای تشخیص حالاتی مانند تمسخر، تردید یا خشم در صدا، برای ماشین‌های زبان فارسی چالش‌برانگیز است.

٤. شکاف سوگیری لهجه [15] در تقابل با رویکرد ارتباط‌محور استاندارد:

طبق جدول ۲-۱۲ صفحه ۵۱، استاندارد بنیاد سعدی تصریح می‌کند که در سطح ماهر، «لهجه غیربومی» تا زمانی که مانع ارتباط مؤثر نشود و ساختار کلام صحیح باشد، پذیرفته است. این یک رویکرد کاملاً ارتباط‌محور و عملگراست. در مقابل، مدل‌های هوش مصنوعی، به ویژه در بخشASR ، اغلب بر پایه داده‌های گویشوران بومی با لهجه معیار (عمدتاً تهرانی) آموزش دیده‌اند. در نتیجه، هرگونه انحراف آوایی از این لهجه معیار را در موارد مختلف به عنوان خطا شناسایی کرده و نمره کسر می‌کنند. این رفتار، نقض صریح روح حاکم بر استاندارد در سطوح پیشرفته است و می‌تواند به نمره‌دهی ناعادلانه به برخی زبان‌آموزان با لهجه‌های دیگر منجر شود.

شواهد کمی و نظری: مطالعات گسترده‌ای در حوزه سوگیری الگوریتمی نشان داده‌اند که سیستم‌های ASR به طور سیستمی عملکرد ضعیف‌تری بر روی لهجه‌های غیرمعیار یا زبان‌آموزان (L2) دارند، که منجر به نرخ خطای کلمه (WER) بالاتری نسبت به گفتار معیار می‌شود. این پدیده، به‌ویژه به دلیل انحرافات قابل توجه گفتار غیربومی از گرایش‌های مرکزی (میانگین) تولیدات آوایی در مقایسه با گویندگان بومی رخ می‌دهد .کسی و ییگر، (۲۰۲۰).علی‌رغم آنکه پژوهش‌ها نشان داده‌اند این انحرافات لزوماً ناشی از افزایش «تنوع درون‌گروهی» یا ناپایداری در گفتار زبان‌آموزان مسلط نیست، مدل‌های ASR که بر الگوهای معیار بومی آموزش دیده‌اند، هرگونه انحراف از آن میانگین را به عنوان «عدم انطباق یا خطا» تفسیر می‌کنند. کسی و ییگر، (۲۰۲۰). در بافت زبان فارسی کم‌منابع، این سوگیری فنی، تهدیدی مستقیم و کمی‌پذیر برای عدالت آزمونی و روایی سازه استاندارد سعدی است، زیرا فرد ارتباط‌گر را به دلیل عدم انطباق با الگوی صوتی معیار جریمه می‌کند.

---

11. Content Validity
12. Emotional Blindness
13. Dialogic Interaction Assessment
14. Turn-based
15. Accent Bias

با توجه به فقر داده‌های L2 Persian، پژوهش حاضر بر مبنای تحلیل محتوایی داده‌های منتشر شده‌ی قیاسی در زبان‌های سامی و هندواروپایی مشابه (که از لحاظ ساختار آوایی و محدودیت‌های داده‌ای به فارسی نزدیک‌ترند) استدلال می‌کند که: به دلیل ساختار مشابه الگوریتم‌های ASR و ضعف ذاتی آن‌ها در برابر گونه‌های غیرمعیار، منطقاً پیش‌بینی می‌شود که WER برای زبان‌آموزان غیرمعیار فارسی به طور قابل توجهی (بین ۱۵ تا ۳۰ درصد) افزایش یابد. این میزان افزایش، با نتایج مشاهده‌شده‌ی افزایش WER تا حدود ٪۴۰ در سایر زبان‌های کم‌منابع هم‌خوانی دارد.

۵ .شکاف در ارزیابی «راهبردهای فراگفتمانی» و «مدیریت تعامل زنده»:

استناد به سند: در سطوح عالی بر «استفاده از راهبردهای تعاملی و گفتمانی گوناگونی مانند نوبت‌گیری» و «شرکت در هر مکالمه یا مباحثه‌ای» تأکید شده است.

تحلیل شکاف: ارزیابی این شایستگی‌ها مستلزم یک گفتگوی زنده و پویا است. هوش مصنوعی‌های کنونی اساساً در یک الگوی «نوبت‌گیری از پیش تعیین شده و منفعلانه»[16] عمل می‌کنند. آن‌ها قادر نیستند مانند یک ممتحن انسان، عمداً فضای گفتگو را پیچیده کنند (بعنوان نمونه با طرح یک دیدگاه مخالف غیرمنتظره)، نوبت گفتار را به صورت طبیعی قطع کنند تا تاب‌آوری زبانی را بسنجند، یا به نشانه‌های فرازبانی و غیرکلامی زبان‌آموز (مانند درخواست کمک با نگاه یا تغییر لحن) پاسخ دهند. بنابراین، هوش مصنوعی عملاً قادر به سنجش مهم‌ترین بخش «توانش ارتباطی» در سطح عالی، یعنی «مدیریت تعامل اجتماعی در زمان واقعی»، نیست.

۵ .بحث و نتیجه‌گیری

۵-۱ .بحث: خلاصه وضعیت روایی سنجش در سطوح مختلف

نتایج تحلیل تطبیقی فصل چهارم در جدول زیر خلاصه شده است که نمای کلی پژوهش را نشان می‌دهد:

سطح مهارت	مولفه کلیدی استاندارد سعدی	وضعیت عملکرد هوش مصنوعی	سطح چالش روایی سازه	ماهیت محدودیت
نوآموز/مقدماتی (N-A)	صحت تلفظ، واژگان پایه، جملات کلیشه‌ای	**بسیار کارآمد :**دقیق‌تر و سریع‌تر از انسان در تشخیص الگو	ناچیز	**فنی/قابل رفع**
میانی(B)	راهبردهای جبرانی، تعامل، نیاز به مخاطب تنظیم‌کننده ،بحث انتزاعی	**دچار کژفهمی :**تفسیر استراتژی ارتباطی به عنوان خطا؛ ناتوان در ایفای نقش تسهیل‌گر تعاملی؛ ناتوان در سنجش تعامل پویا.	متوسط	**بخشی فنی، بخشی معرفت‌شناختی**

[16] .Turn-based & Passive

پیشرفته/ماهر(C)	طنز، کنایه، اقناع، ادب، تناسب اجتماعی و انعطاف سبکی، مدیریت گفتمان زنده، پذیرش لهجه غیربومی.	**ناتوان :**فقدان شعور فرهنگی و عاطفی، کوری نسبت به هنجارهای موقعیتی، ناتوانی در مدیریت یا ارزیابی گفتگوی پویا، سوگیری علیه لهجه.	بسیار بالا	**معرفت‌شناختی/ذاتی**

تحلیل تطبیقی ما نشان می‌دهد که محدودیت‌ها به دو دسته عمده تقسیم می‌شوند: محدودیت‌های «فنی و منابعی» در سطوح مقدمات و بخشی از میانی که با تزریق داده‌های بیشتر و پیشرفت‌های آتی قابل رفع کامل هستند، و محدودیت‌های «ذاتی و معرفت‌شناختی» (مانند فقدان تجسد، ناتوانی در ارزیابی صحت محتوایی و استدلال، و تفسیر راهبرد جبرانی) که ریشه در «ماهیت آماری مدل‌های فعلی» دارند. تا زمانی که معماری هوش مصنوعی از الگوی تطبیق الگو به هوش تجسد یافته و موقعیتی تحول نیابد، پر کردن کامل این شکاف‌ها و جایگزینی مطلق ارزیاب انسانی در سنجش کامل «توانش ارتباطی» در سطوح عالی، بدون تحول الگوواره در معماری، نامحتمل است. به این ترتیب، ما با قاطعیت بین آنچه فناوری «می‌تواند» و آنچه «نمی‌تواند» انجام دهد، تمایز قائل می‌شویم.

درجه‌بندی فوق بر اساس تحلیل تطبیقی و پیش‌بینی نظری نویسنده از میزان انحراف میان «سازه تعریف شده در استاندارد» و «خروجی فنی الگوریتم‌های فعلی» تعیین شده است. یافته‌های این پژوهش را می‌توان در چارچوب اعتبار مبتنی بر استدلال **چاپل و واس، (۲۰۲۱)** خلاصه کرد: شواهد تحلیل تطبیقی نشان می‌دهد که برای ادعای استفاده انحصاری از نمره هوش مصنوعی در سطوح پیشرفته به عنوان شاخص صلاحیت ارتباطی کامل، پیوند استنباطی بین «خروجی الگوریتم» و «تفسیر میزان تسلط ارتباطی» به دلیل کوری فرهنگی، عاطفی و تعاملی سیستم‌های فعلی، ضعیف یا گسسته است. این ضعف، بنیان اعتبار چنین استفاده‌ای از نمره را متزلزل می‌سازد.»

۵-۲ . چالش حیاتی: اثر بازگشتی منفی[17]

یکی از مهم‌ترین یافته‌های تحلیلی این پژوهش، تأثیر منفی احتمالی بر اکوسیستم آموزشی است. در آموزش زبان، آزمون‌ها نقش «فرمان» را بازی می‌کنند. اگر آزمون‌های سطوح عالی و سرنوشت‌ساز به ماشین سپرده شود، مدرسان و زبان‌آموزان به صورت ناخودآگاه تلاش می‌کنند تا «ماشین‌پسند» صحبت کنند. این به معنای استفاده از جملات کتابی، حذف مکث‌های طبیعی تفکر، پرهیز از خلاقیت‌های زبانی پیچیده، اجتناب از تعارفات فرهنگی و حتی تغییر لهجه به سوی معیار مصنوعی است. این امر روح حاکم بر استاندارد بنیاد سعدی را که بر «ارتباط طبیعی، موثر و فرهنگی» تأکید دارد، نقض می‌کند و کلاس‌های درس را به کارگاه‌های «تست‌زنی ماشینی» تبدیل می‌نماید.

۵-۳ .نتیجه‌گیری نهایی

پژوهش حاضر با واکاوی دقیق سند بنیاد سعدی نشان داد که فناوری هوش مصنوعی فعلی، علی‌رغم پیشرفت‌های شگرف، شایستگی معرفت‌شناختی لازم برای جایگزینی کامل ارزیاب انسانی در سطوح میانی و عالی را ندارد. ماشین در سنجش «فرم» استاد است اما در سنجش

17. Negative Washback Effect

«معنا در بافت[18] ناتوان». شکاف‌های شناسایی‌شده در حوزه‌های راهبردهای ارتباطی، تعامل پویا، فرهنگ، عاطفه و عدالت در قبال لهجه‌ها، استفاده انحصاری از آن را در سطوح بالای آزمون‌های سرنوشت‌ساز ناموجه و حتی خطرناک می‌سازد.

۴-۵ .پیشنهادات کاربردی و اجرایی

به منظور بهره‌گیری از مزایای فناوری (سرعت، عینیت، مقیاس‌پذیری) و پرهیز از معایب آن (شکاف‌های فوق)، مدل «سنجش ترکیبی سلسله‌مراتبی» به عنوان خروجی عملیاتی پژوهش پیشنهاد می‌شود:

اتوماسیون کامل سطوح پایه: آزمون‌های سطوح نوآموز و مقدماتی به طور کامل به ماشین واگذار شود. این کار باعث کاهش چشمگیر هزینه‌های عملیاتی و افزایش سرعت و دسترسی می‌گردد.

غربالگری هوشمند با نظارت انسانی در سطح میانی: در سطوح میانی(B) ، ماشین نقش غربالگر اولیه را ایفا کند. با توجه به حجم پیش‌بینی شده آموزش (حدود ۷۲۰ ساعت تا سطح ماهر در سند استاندارد)، استفاده از ماشین برای ارزیابی‌های تکوینی و کلاسی ضروری است تا بار شناختی مدرس کاهش یابد. اما در آزمون‌های جامع پایان سطح، نمرات توسط ماشین صادر شده و فایل‌های دارای الگوهای غیرعادی یا در مرز تصمیم‌گیری، حتماً توسط ارزیاب انسانی بازبینی می‌شوند. عملیاتی‌سازی بازبینی انسانی: معیار «مرز تصمیم‌گیری» باید به صورت عددی مشخص شود؛ برای نمونه نمراتی که در دامنه از آستانه قبولی قرار می‌گیرند، نیازمند بازبینی انسانی هستند. معیار «الگوی غیرعادی نیز بر اساس دو شاخص کلیدی خواهد بود: الف) تفاوت شدید نمرات فرعی ماشین (نظیر تفاوت بیش از ۲۰ درصدی بین نمره روانی کلام و نمره دقت واژگانی/دستوری) یا) بالا بودن غیرطبیعی زمان مکث‌های جبرانی که می‌تواند نشان‌دهنده استفاده از راهبردهای جبرانی ارتباطی باشد که توسط ماشین به درستی تفسیر نشده است.

عملیاتی‌سازی بازبینی انسانی: معیار «مرز تصمیم‌گیری» باید به صورت عددی مشخص شود؛ برای مثال، نمراتی که در دامنه ±٪۵ تا ±٪۱۰ از آستانه قبولی قرار می‌گیرند، نیازمند بازبینی انسانی هستند. معیار «الگوی غیرعادی» نیز بر اساس دو شاخص کلیدی خواهد بود: الف) تفاوت شدید نمرات فرعی ماشین (نظیر تفاوت بیش از ۲۰ درصدی بین نمره روانی کلام و نمره دقت واژگانی/دستوری). این ناسازگاری نشان می‌دهد که داوطلب علی‌رغم مشکلات صوری، پیام را به طور مؤثر منتقل کرده و نیازمند قضاوت انسانی درباره «توانش استراتژیک» است. ب) بالا بودن غیرطبیعی زمان مکث‌های جبرانی که می‌تواند نشان‌دهنده استفاده از راهبردهای جبرانی ارتباطی باشد که توسط ماشین به درستی تفسیر نشده است.

منابع فارسی

۱. 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/2247368 .

18. Meaning in Context

۲. مظهرپور، دیار و سیدکلان، سید محمد (۱۴۰۳). سنتزپژوهی کاربرد چت‌بات‌ها (نرم‌افزار هوش مصنوعی) در آموزش زبان انگلیسی. پژوهش در مطالعات برنامه درسی، ۱۴(۱)، ۴۳-۶۴. https://doi.org/10.48310/jcdr.2024.17527.1115 .

۳. صبوری، س.، و حاج ملک، م. م. (۱۴۰۲). استفاده از ظرفیت‌های هوش مصنوعی در آموزش تلفظ زبان‌های خارجی. نهمین کنفرانس بین‌المللی وب پژوهی.

منابع انگلیسی

1. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.ACL-MAIN.463

2. Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press. https://assets.cambridge.org/97811084/84022/frontmatter/9781108484022_frontmatter.pdf

3. Huth, T. (2020). Testing interactional competence: Patterned yet dynamic aspects of L2 interaction. *Papers in Language Testing and Assessment*, 9(1), 1-25.

4. Ie, X., & Jaeger, T. F. (2020). Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5), 3322-3347. https://doi.org/10.1121/10.0001141

5. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38. https://doi.org/10.1145/3571730

6. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

7. Kittler, M. G., Rygl, D., & Mackinnon, A. (2011). Beyond culture or beyond control? Reviewing the use of Hall's high-/low-context concept. *International Journal of Cross Cultural Management*, 11(1), 63-82. https://doi.org/10.1177/1470595811398797

8. Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *Information Systems Frontiers*, 24(5), 1321-1340. https://doi.org/10.1080/0960085X.2021.1927212

9. Liu, X. J., Wang, J., & Zou, B. (2025). Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback. *Journal of English for Academic Purposes*, 75, 101505.

10. Manggiasih, L. A., et al. (2023). Strengths and limitations of SmallTalk2Me app in English language proficiency evaluation. *TELL Journal*, 11(2), 146-157.

11. Nigmatulina, I., Kew, T., & Samardžić, T. (2020). 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 7th Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 15-24). International Committee on Computational Linguistics (ICCL). https://aclanthology.org/2020.vardial-1.2/

12. Raud, N. (2025). *Automatic assessment of L2 interactional competency* [Master's thesis, Aalto University].

13. Santos, S. C., Kapadia, A., & Feinberg, D. R. (2025). Hearing people speak in different accents biases voice discrimination. *Scientific Reports*, 15, 30775. https://doi.org/10.1038/s41598-025-13117-w

14. Zhang, M., Bridgeman, B., & Davis, L. (2019). Validity considerations for using automated scoring in speaking assessment. In *Automated speaking assessment* (pp. 174-185). Routledge.

15. Zou, B., et al. (2024). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system. *System*, 126, 103497.