فصلنامه توسعه اجتماعی (توسعه انسانی سابق)، دوره ۲۰، شماره ۳، بهار ۱۴۰۵، صفحات ۳۱۶-۲۸۳

بازطراحی آزمون های استاندارد در مواجهه با هوش مصنوعی: تحلیل مزایا و معایب مدل شفاهی

دکتر آنیتا همایون‌نیا[۱]

تاریخ وصول: ۱۴۰۴/۹/۱۸ تاریخ پذیرش: ۱۴۰۴/۱۲/۱۹

چکیده:

پژوهش حاضر با هدف بازطراحی آزمون‌ها در عصر هوش مصنوعی، به بررسی و مقایسه عملکرد سه مدل زبانی پیشرفته (GPT-4o، Google Gemini 1.5 Pro، Anthropic Claude 3.5 Sonnet) با دانشجویان انسانی در قالب آزمون‌های کتبی و شفاهی بازطراحی شده است. روش تحقیق به صورت ترکیبی (کمی و کیفی) بوده و داده‌ها از طریق آزمون‌های استاندارد زبان و شناختی (IELTS، Raven's Progressive Matrices، N-Back)، تحلیل خطا و پرسشنامه تجربه کاری گردآوری شد. جامعه آماری تحقیق شامل دانشجویان پزشکی دانشگاه علوم پزشکی آزاد تهران در سال تحصیلی ۱۴۰۴-۱۴۰۳ بود که معادل ۵۸۵ نفر بوده و نمونه‌گیری به روش تصادفی طبقه‌بندی شده انجام شد. حجم نمونه با استفاده از فرمول کوکران؛ معادل ۲۳۲ نفر برآورد گردید. نتایج نشان داد که مدل‌های هوش مصنوعی در اغلب شاخص‌ها (دقت، سرعت پاسخ‌گویی و میانگین نمرات) عملکردی بالاتر از انسان داشتند، هر چند نوع خطاهای آنان ماهیتاً با خطاهای انسانی متفاوت بود. همچنین، نسخه بازطراحی شده آزمون شفاهی موجب بهبود تجربه کاربری، افزایش وضوح دستورالعمل‌ها، کاهش استرس و ارتقای ادراک عدالت در میان شرکت‌کنندگان شد. میان مؤلفه‌های پرسنامه تجربه کاری نیز همبستگی معناداری در سطح ۰.۰۱ وجود داشت ($p<0.01$). یافته‌ها بیانگر آن است که اگرچه هوش مصنوعی توانسته است در بازتولید دانش و پردازش سریع اطلاعات از انسان پیشی گیرد، اما آزمون‌های شفاهی و تعاملی همچنان ظرفیت بالاتری در سنجش مهارت‌های استدلالی، خلاقیت و تفکر انتقادی دارند.

مفاهیم کلیدی: *هوش مصنوعی، دانشجویان انسانی، آزمون‌های کتبی و شفاهی، خطا*

[۱] گروه علوم تربیتی، دانشگاه پیام نور، تهران، ایران homayounnia@pnu.ac.ir

مقدمه و بیان مسأله

در دهه اخیر، هوش مصنوعی به یکی از نیروهای دگرگون‌ساز بنیادین در حوزه آموزش و ارزشیابی تبدیل شده است. پیشرفت‌های شتابان در زمینه یادگیری ماشینی، پردازش زبان طبیعی و سیستم‌های توصیه‌گر موجب شده است که بسیاری از شیوه‌های سنتی سنجش دانشجویان و داوطلبان مورد بازنگری جدی قرار گیرند (*Hao & et. al., 2024; Karademir Coskun & Alper, 2024*). آزمون‌های استاندارد، به‌ویژه در قالب آزمون‌های چندگزینه‌ای و پرسش‌های کوتاه، همواره به‌عنوان ابزارهای اصلی سنجش دانش و مهارت در سطوح مختلف تحصیلی و حرفه‌ای شناخته شده‌اند (*Huawei & Aryadoust, 2023*). با این حال، ظهور ابزارهای هوش مصنوعی که قادر به تحلیل و تولید پاسخ‌هایی هم‌تراز با پاسخ‌های انسانی هستند، اعتبار، روایی و قابلیت اعتماد این آزمون‌ها را با چالش‌های جدی مواجه ساخته است. یکی از مهم‌ترین پیامدهای گسترش هوش مصنوعی در فرآیندهای ارزشیابی، ضرورت بازطراحی آزمون‌ها به‌منظور کاهش امکان تقلب و جلوگیری از وابستگی بیش از حد به فناوری است. پژوهش‌ها نشان داده‌اند که دانشجویان می‌توانند با بهره‌گیری از سامانه‌های مبتنی بر هوش مصنوعی، پاسخ‌های خود را بهبود داده یا حتی پاسخ‌ها را به‌طور کامل تولید کنند، امری که به تضعیف اعتبار نتایج آزمون‌های سنتی منجر می‌شود (*Sayin & Gierl, 2024; Ranger & et. al, 2023*). در این چارچوب، مدل‌های شفاهی به‌عنوان جایگزینی بالقوه مطرح شده‌اند؛ چرا که امکان سنجش عمیق‌تر درک مفاهیم، توانایی استدلال و مهارت‌های تحلیلی را به‌صورت پویا و تعاملی فراهم می‌کنند (*Jiao & et. al, 2023; Khera & et. al, 2023*).

مرور تاریخچه آزمون‌های استاندارد نشان می‌دهد که این ابزارها از اوایل قرن بیستم با هدف سنجش دانش و توانایی‌های تحصیلی طراحی شده‌اند و هدف اصلی آن‌ها ایجاد امکان مقایسه‌ای عادلانه، منصفانه و قابل اعتماد میان داوطلبان بوده است. با این وجود، حتی پیش از ظهور هوش مصنوعی، محدودیت‌هایی نظیر تأکید بیش از حد بر یادگیری سطحی، ناتوانی در سنجش تفکر انتقادی و دشواری ارزیابی مهارت‌های عملی، کارایی این آزمون‌ها را تحت تأثیر قرار داده بود (*Yu & Yu, 2023*). امروزه، هوش مصنوعی این کاستی‌ها را تشدید کرده و بازنگری در طراحی آزمون‌ها را نه تنها مطلوب، بلکه ضرورتی اجتناب‌ناپذیر ساخته است. ادبیات پژوهشی موجود در زمینه بازطراحی آزمون‌ها در مواجهه با هوش مصنوعی عمدتاً بر دو محور اساسی تمرکز دارد: نخست، تحلیل محدودیت‌ها و آسیب‌پذیری‌های آزمون‌های استاندارد سنتی در برابر ابزارهای هوش مصنوعی و دوم، بررسی اثربخشی الگوهای جایگزین،

از جمله آزمون‌های شفاهی یا ارزشیابی‌های مبتنی بر پروژه (*Zhai & Nehm, 2023*). مطالعات اخیر حاکی از آن است که مدل‌های شفاهی توانایی بیشتری در سنجش تفکر انتقادی و مهارت‌های زبانی دارند و نسبت به پاسخ‌های تولیدشده خودکار توسط هوش مصنوعی، از استقلال بیشتری برخوردارند (*Jingshan, 2023; Rahm, 2023*). به‌عنوان نمونه، برخی دانشگاه‌ها در اروپا و آمریکای شمالی اقدام به پیاده‌سازی آزمون‌های شفاهی آنلاین کرده‌اند که در آن داوطلبان موظف‌اند پاسخ‌های خود را به‌صورت زنده ارائه دهند و ارزیابان امکان تحلیل فرآیند استدلال آنان را دارند. این شیوه ارزشیابی، به‌ویژه در حوزه علوم انسانی و مباحث تحلیلی، افزایش دقت و اعتمادپذیری نتایج ارزیابی را نشان داده است. با وجود این مزایا، مدل‌های شفاهی نیز با چالش‌هایی همراه هستند (*Chang, 2024*). از جمله این چالش‌ها می‌توان به نیاز به نیروی انسانی متخصص، افزایش زمان مورد نیاز برای ارزشیابی، دشواری استانداردسازی و احتمال بروز سوگیری ارزیابان اشاره کرد. افزون بر این، برخی پژوهش‌ها بیانگر آن‌اند که در محیط‌های آنلاین، مشکلات فنی و دسترسی نابرابر به فناوری می‌تواند به ایجاد نابرابری آموزشی منجر شود (*Wang & et. al, 2024*). از این‌رو، تحلیل جامع مزایا و معایب مدل شفاهی در فرآیند بازطراحی آزمون‌های استاندارد امری ضروری به نظر می‌رسد تا بتوان راهکارهایی ارائه کرد که ضمن حفظ دقت، روایی و عدالت آموزشی، قابلیت اعتماد آزمون‌ها را در مواجهه با هوش مصنوعی تضمین کند. اهمیت این موضوع در شرایط کنونی، که هوش مصنوعی با سرعتی چشمگیر در حال نفوذ به تمامی ابعاد نظام‌های آموزشی است، بیش از پیش آشکار می‌شود. بازطراحی آزمون‌ها با تمرکز بر مدل‌های شفاهی صرفاً به معنای تغییر در شکل و قالب آزمون‌ها نیست، بلکه مستلزم بازاندیشی در فلسفه، اهداف و کارکردهای سنجش آموزشی است (*Swiecki & et. al, 2022; Kamalov & et. al, 2021*). این رویکرد می‌تواند به تقویت مهارت‌های تحلیلی، استدلالی و ارتباطی دانشجویان منجر شود و در عین مقابله با محدودیت‌های آزمون‌های سنتی، امکان سنجش عمیق‌تر و جامع‌تر شایستگی‌ها را فراهم آورد (*Kitsara, 2022*).

پژوهش حاضر با تمرکز بر تحلیل نظام‌مند مزایا و معایب مدل شفاهی، در پی ارائه چارچوبی علمی و مستدل است که بتواند راهنمایی برای بهینه‌سازی آزمون‌های استاندارد در عصر هوش مصنوعی فراهم سازد. تحقق این هدف می‌تواند گامی مؤثر در جهت ارتقای کیفیت و اعتبار سنجش آموزشی، کاهش آسیب‌پذیری در برابر ابزارهای هوش مصنوعی و تضمین عدالت آموزشی باشد. افزون بر این، ارائه تحلیل‌های مبتنی بر شواهد می‌تواند به

سیاست‌گذاران آموزشی، مؤسسات ارزشیابی و پژوهشگران حوزه آموزش یاری رساند تا تصمیماتی آگاهانه‌تر و مبتنی بر مبانی علمی در زمینه طراحی و اجرای آزمون‌ها اتخاذ کنند.

اهمیت و ضرورت پژوهش

تحولات شتابان در عرصه هوش مصنوعی و گسترش بی‌سابقه مدل‌های زبانی بزرگ، نظام‌های سنجش و ارزیابی آموزشی را با چالشی عمیق و ساختاری مواجه ساخته است. آزمون‌هایی که طی دهه‌ها به‌عنوان معتبرترین ابزارهای سنجش توانایی‌های شناختی، زبانی و تحصیلی شناخته می‌شدند، امروز در معرض تحولاتی قرار گرفته‌اند که نه‌تنها کارآمدی، بلکه ماهیت و فلسفه وجودی آن‌ها را نیز به‌طور جدی به پرسش می‌کشد. دیگر نمی‌توان با قطعیت فرض کرد که پاسخ‌دهنده آزمون، انسانی است که بر پایه دانش، تجربه و مهارت فردی خود به پرسش‌ها پاسخ می‌دهد؛ چرا که مدل‌های پیشرفته هوش مصنوعی، نظیر *GPT-4o*، *Gemini* و *Claude*، قادرند همان پرسش‌ها را با سرعتی چشمگیر و دقتی بالا تحلیل و حل کنند. این توانمندی، اعتبار آزمون‌های سنتی را به‌طور بنیادین متزلزل ساخته و نشان می‌دهد که نظام آموزشی معاصر بیش از هر زمان دیگری نیازمند بازاندیشی در طراحی آزمون‌ها و بازتعریف فلسفه ارزیابی است تا بتواند مرز روشنی میان توانایی انسانی و توانایی ماشینی ترسیم کند.

اهمیت این پژوهش از آنجا نشأت می‌گیرد که آشکار می‌سازد آزمون‌های استاندارد در قالب سنتی خود، بیش از آنکه سنجه معناداری برای ارزیابی مهارت‌های انسانی باشند، بر بازیابی اطلاعات و سرعت پردازش متکی‌اند؛ مؤلفه‌هایی که دقیقاً در زمره نقاط قوت هوش مصنوعی قرار دارند. در چنین چارچوبی، عملکرد برتر مدل‌های *AI* در آزمون‌ها نه بازتابی از ضعف انسان، بلکه نشانگر محدودیت ساختاری آزمون‌هایی است که سال‌ها بدون تحول باقی مانده‌اند. هنگامی که منطق طراحی آزمون بر پرسش‌های بازتولیدی و پاسخ‌های الگومحور استوار باشد، طبیعی است که سامانه‌هایی با ظرفیت پردازشی گسترده و توان تحلیل سریع، نتایج بهتری کسب کنند. این وضعیت، هشداری جدی برای نظام ارزشیابی به شمار می‌رود؛ زیرا آزمونی که قادر به تمایزگذاری میان عملکرد انسانی و ماشینی نباشد، نمی‌تواند معیار معتبری برای سنجش توانایی، صلاحیت یا آمادگی تحصیلی و حرفه‌ای تلقی شود (*Tate & et. al, 2024*).

ضرورت انجام این پژوهش در آن است که بازطراحی آزمون‌ها را نه اقدامی جانبی یا تکمیلی، بلکه ضرورتی راهبردی برای حفظ عدالت آموزشی معرفی می‌کند. گسترش ابزارهای هوش مصنوعی در فضای دانشگاهی و امکان بهره‌گیری پنهان از آن‌ها در فرایندهای ارزشیابی،

تهدیدی جدی برای امانت‌داری علمی و اعتبار نتایج آزمون‌ها به شمار می‌آید. اگر آزمون‌ها نتوانند تشخیص دهند که پاسخ ارائه‌شده حاصل اندیشه و استدلال انسانی است یا محصول یک مدل زبانی، کل نظام ارزشیابی با نوعی فروپاشی معنایی مواجه خواهد شد. در این میان، بازطراحی آزمون‌ها با تمرکز بر الگوهای شفاهی و تعاملی به‌عنوان یکی از مؤثرترین راهکارها مطرح می‌شود؛ زیرا در این الگوها، فرایند تفکر، استدلال، تبیین و تحلیل در زمان واقعی مورد سنجش قرار می‌گیرد و امکان بازتولید ماشینی پاسخ‌ها به‌طور معناداری کاهش می‌یابد (*Sayed & et. al, 2024*).

از سوی دیگر، آزمون‌های شفاهی و تعاملی ظرفیت منحصر به فردی در آشکارسازی ابعادی از توانایی انسانی دارند که هوش مصنوعی هنوز قادر به بازتولید کامل آن‌ها نیست؛ از جمله خلاقیت، تحلیل موقعیتی، انعطاف‌پذیری در پاسخ‌دهی و مدیریت تعامل زنده. این پژوهش با مقایسه‌ای نظام‌مند میان عملکرد دانشجویان و مدل‌های *AI* در آزمون‌های کتبی و شفاهی نشان می‌دهد که اگر چه ماشین از نظر سرعت و دقت برتری محسوسی دارد، اما ماهیت خطاهای آن با خطاهای انسانی متفاوت است و اغلب ریشه در ناتوانی در درک بافت موقعیتی، ظرایف معنایی و سطوح عمیق معنا دارد. این یافته‌ها اهمیت آزمون‌های شفاهی بازطراحی شده را دوچندان می‌کند؛ چرا که این آزمون‌ها دقیقاً نقاط ضعف مدل‌های زبانی را هدف قرار داده و امکان تمایز میان «دانش اصیل» و «پاسخ تولیدشده» را فراهم می‌سازند.

افزون بر این، ضرورت انجام پژوهش حاضر از منظر تجربه آزمون‌دهندگان نیز قابل تبیین است. نتایج نشان می‌دهد که نسخه بازطراحی شده آزمون، به‌طور معناداری احساس عدالت، شفافیت دستورالعمل‌ها، کاهش اضطراب و رضایت کلی شرکت‌کنندگان را افزایش داده است. این امر از آن جهت حائز اهمیت است که مشروعیت اجتماعی آزمون‌ها صرفاً به نتایج کمی وابسته نیست؛ بلکه ادراک عدالت، شفافیت و اعتماد آزمون‌دهندگان، یکی از ارکان اساسی اعتبار هر نظام ارزشیابی به شمار می‌رود. هنگامی که داوطلبان اطمینان یابند آزمون در برابر مداخله هوش مصنوعی مقاوم است و همه افراد در شرایطی نسبتاً برابر سنجیده می‌شوند، نتایج آزمون برای آنان پذیرفتنی‌تر و معتبرتر خواهد بود (*Zawacki-Richter & et. al, 2019*). از این منظر، بازطراحی آزمون‌ها نه تنها پاسخی فناورانه، بلکه پاسخی روان‌شناختی و اخلاقی به ضرورت حفظ عدالت آموزشی تلقی می‌شود.

اهمیت دیگر این پژوهش در پر کردن شکافی معنادار در ادبیات علمی نهفته است. بخش عمده مطالعات پیشین در حوزه هوش مصنوعی و آموزش، بر موضوعاتی نظیر نمره‌گذاری خودکار، تولید سؤال یا پیامدهای اخلاقی و اجتماعی *AI* متمرکز بوده‌اند؛ در حالی که

پژوهش‌هایی که به صورت مستقیم عملکرد انسان و مدل‌های زبانی را در آزمون‌های شفاهی و کتبی مقایسه کنند و هم‌زمان تأثیر بازطراحی آزمون‌ها بر تجربه انسانی، الگوهای خطا، عدالت و عملکرد را مورد بررسی قرار دهند، بسیار محدود بوده‌اند. پژوهش حاضر با بهره‌گیری توأمان از روش‌های کمی و کیفی، تصویری چندبعدی از این تحولات ارائه می‌دهد و می‌تواند مبنایی علمی برای سیاست‌گذاری در طراحی آزمون‌ها در سطوح دانشگاهی، حرفه‌ای و حتی ملی فراهم آورد (*Yu & Yu, 2023*).

در نهایت، این پژوهش نشان می‌دهد که بازطراحی آزمون‌ها صرفاً یک تغییر فنی یا اجرایی نیست، بلکه دلالت بر تحولی عمیق در فلسفه سنجش آموزشی دارد. اگر هدف آموزش، پرورش انسانی خلاق، تحلیل‌گر، گفت‌وگوگر و توانمند در مواجهه با موقعیت‌های واقعی است، آزمون‌ها نیز باید سنجه چنین توانایی‌هایی باشند. عصر هوش مصنوعی فرصتی فراهم کرده است تا نظام‌های ارزشیابی از سطح بازتولید دانش فراتر رفته و به سنجش شایستگی‌های عمیق‌تر انسانی بپردازند. از این‌رو، ضرورت پژوهش حاضر نه تنها از منظر روش‌شناسی و کارکرد اجرایی، بلکه از منظر تحولی بنیادین در ماهیت سنجش آموزشی قابل دفاع است. این پژوهش به روشنی نشان می‌دهد که آینده نظام‌های ارزیابی در گرو طراحی آزمون‌هایی است که انسان را نه در رقابت با ماشین، بلکه در توانایی‌های اصیل، منحصر به فرد و انسانی‌اش مورد سنجش قرار دهند.

پیشینه پژوهش

الف) تحقیقات داخلی

هاشمی و همکاران (۱۴۰۴)، در تحقیقی به بررسی کاربرد هوش مصنوعی در طراحی آزمون‌های هوشمند و ارزیابی دقیق‌تر توانایی‌های دانش‌آموزان پرداختند. پژوهش نشان داد که پیاده‌سازی هوش مصنوعی در مدارس با چالش‌هایی مواجه است. مسائل اخلاقی و حریم خصوصی داده‌ها، به‌ویژه داده‌های حساس مانند وضعیت عاطفی و اجتماعی دانش آموزان، از جمله مهم‌ترین نگرانی هاست. همچنین، محدودیت‌های فنی مانند نیاز به زیرساخت‌های پیشرفته و آموزش‌های لازم برای معلمان از دیگر چالش‌ها به شمار می‌آید.

بازگیر جهان تیغی و همکاران (۱۴۰۴)، در تحقیقی به چالش‌های استفاده از هوش مصنوعی در طراحی آزمون‌های استاندارد توسط معلمان پرداختند. یافته‌ها نشان داد مهم‌ترین چالش‌ها شامل نبود آموزش مستمر، کمبود زیرساخت سخت‌افزاری و نگرانی‌های اخلاقی

است. در نتیجه، برای ارتقای موفقیت‌آمیز کاربرد *AI*، آموزش هدفمند، حمایت سیاست‌گذاران و توسعه زیرساخت‌های فناورانه توصیه می‌شود.

زائری و همکاران (۱۴۰۴)، به بررسی استفاده از هوش مصنوعی در طراحی آزمون‌های انطباقی یادگیری پرداختند. بنابر یافته‌ها، استفاده از الگوریتم‌های یادگیری ماشین امکان ارزیابی مستمر و لحظه‌ای سطح توانایی هر آزمودنی را فراهم می‌آورد که این امر منجر به افزایش کارایی فرآیند سنجش می‌شود. تحلیل‌ها نشان می‌دهد که علیرغم چالش‌های فنی و اخلاقی، پتانسیل این فناوری برای دگرگون‌سازی روش‌های سنتی ارزیابی بسیار بالاست.

حامدی‌نسب و رحیمی (۱۴۰۳)، در تحقیقی به موانع و چالش‌های پیاده‌سازی هوش مصنوعی در نظام آموزش عالی پرداختند. تحلیل داده‌ها به شناسایی ۶ مقوله اصلی و ۱۶ مقوله فرعی منجر شد. چالش‌های اصلی شامل کمبود تجهیزات پیشرفته، فیلترینگ و سرعت پایین اینترنت، موانع فرهنگی همچون مقاومت در برابر تغییرات، نبود قوانین مشخص و ضعف نظارت در ابعاد تقنینی، کمبود نیروی متخصص و نگرانی از تأثیر هوش مصنوعی بر مشاغل سنتی بود. علاوه بر این، عوامل ملی، بین‌المللی، خانوادگی و آموزشی نیز تأثیرگذار بودند. راهکارهای غلبه بر این موانع شامل تقویت زیرساخت‌ها و ارائه برنامه‌های آموزشی برای دانشجویان می‌باشد.

نویدی و قیصری (۱۴۰۲)، در تحقیقی به بررسی تأثیر هوش مصنوعی (*AI*) در بهبود بهره‌وری سازمانی پرداختند. بنابر یافته‌ها، هوش مصنوعی با اثرگذاری در خدمات مدیریت داده و طرح‌ریزی برای سازمان، بهبود خدمات ارتباط شبکه و خدمات فضای ابری، ایجاد بستر لازم برای خدمات/ فرآیندها) یعنی سرور، فرایندهای مقیاس بزرگ، مانیتور عملکرد (و ایجاد زیرساخت‌های ایمن زمینه و بستر لازم برای ارتقاء بهره‌وری سازمانی را فراهم می‌کند.

ب) تحقیقات خارجی

Zhong & et. al (2024) در تحقیقی به معرفی بنچ‌مارک «انسان‌محور» پرداخته که هدف آن ارزیابی توانایی مدل‌های بنیادین هوش مصنوعی با استفاده از آزمون‌های استاندارد انسانی است. بنابر نتایج؛ بسیاری از مدل‌های زبانی ـ *even GPT-4* ـ در برخی آزمون‌های پیچیده عملکرد سطح پایین‌تری نسبت به انسان دارند؛ همچنین مشخص می‌شود که این بنچ‌مارک می‌تواند شکاف واقعی میان «توانایی ادراکی» انسان و مدل‌های زبانی بزرگ را آشکار کند و راهنمایی برای طراحی آزمون‌های دقیق‌تر فراهم سازد.

Dong & et. al (2024) در تحقیقی با عنوان تعمیم یا حفظ کردن: آلودگی داده‌ها و ارزیابی قابل اعتماد برای *LLM*ها به این مسأله می‌پردازد که بخش قابل توجهی از عملکرد بالای مدل‌های زبانی بزرگ ممکن است ناشی از آلودگی داده‌ای باشد؛ یعنی مدل، نمونه‌های آزمون را قبلاً در داده‌های آموزشی دیده است. نویسندگان روش‌هایی برای اندازه‌گیری آلودگی معرفی می‌کنند، مانند تطبیق شباهت متنی، تحلیل الگوهای حدس خودکار و بررسی ناهنجاری‌های رفتاری. نتایج نشان می‌دهد برخی بنچ‌مارک‌های شناخته شده اعتبار لازم را ندارند و باید نسخه‌های جدید با مدیریت سطح دسترسی، مخفی‌سازی داده و تولید نمونه‌های پویا طراحی شوند. مقاله پیشنهاد می‌کند که طراحی آزمون‌های استاندارد بدون توجه به مسئله آلودگی داده، قابل اعتماد نخواهد بود.

Pack & et. al (2024) این مطالعه چهار *LLM* (از جمله *GPT-4* و *GPT-3.5*، *PaLM2* و *Claude2*) را روی مجموعه‌ای از متون زبان‌آموزان انگلیسی بررسی کرد و نشان داد که برخی مدل‌ها (به‌ویژه *GPT-4*) در نمره‌گذاری نوشتار عملکرد بسیار قابل‌قبولی دارند، اما همچنان نوسانات عملکرد و برخی مسائل اعتبار و همبستگی بین نمره‌گذاران مشاهده شد.

Wecks & et. al (2024) در یک مطالعه اقتصادسنجی استفاده واقعی دانش‌آموزان از *GenAI* و نتایج امتحانات را تحلیل کردند. آنان به‌طرز جالبی نشان دادند که در داده‌های میدانیِ یک دوره آموزشی گسترده، «استفاده ناصحیح یا جایگزینی فراگیر توسط *GenAI*» با افت متوسط نمره همراه بوده است (یعنی کاربران *GenAI* در آزمون‌های سنتی در برخی نمونه‌ها عملکرد پایین‌تری نشان دادند).

Kumar & et. al (2023) در تحقیقی یک چارچوب «*testFAILS*» را معرفی می‌کند که برای آزمون استاندارد سیستم‌های زبانی هوش مصنوعی طراحی شده است. نتایج نشان داد ساختار *testFAILS* شامل مجموعه‌ای از معیارهای طبقه‌بندی‌شده در سطوح عملکردی، زبانی، معناشناختی و اخلاقی است و هدف آن فراهم‌کردن یک روش ارزیابی پایدار و قابل‌تکرار برای مدل‌های مکالمه و چت‌بات‌ها است. بنابراین *testFAILS* قادر است نواقصی را که بنچ‌مارک‌های سنتی نادیده می‌گیرند، آشکار نماید.

مبانی نظری پژوهش

ورود فناوری‌های نوظهوری مانند هوش مصنوعی به عرصه آموزش، روان‌سنجی و ارزیابی، بسیاری از مفاهیم سنتی آزمون‌دهی را دگرگون کرده است. تا پیش از توسعه گسترده مدل‌های زبانی، سیستم‌های تولید محتوا و ابزارهای هوشمند تحلیل‌گر، آزمون‌های استاندارد

بر پایه این فرض عمل می‌کردند که فرد در زمان آزمون تنها بر دانش، توانایی ذهنی، مهارت حل مسئله و تجربه شخصی خود تکیه دارد. اما اکنون این فرض اصلی اعتبار خود را از دست داده است؛ زیرا هوش مصنوعی به شکلی گسترده توانایی تولید پاسخ، استدلال، تحلیل داده و حتی شبیه‌سازی فرایند یادگیری انسان را دارد (*Stanoyevitch, 2024*). بنابراین، چارچوب نظری طراحی آزمون‌های استاندارد در شرایط جدید باید به‌گونه‌ای تعریف شود که توانایی انسانی را در بستری اندازه‌گیری کند که آمیخته با فناوری است، اما تحت سیطره آن قرار نگیرد. از دیدگاه نظری، آزمون استاندارد آزمونی است که از سه ویژگی بنیادین برخوردار باشد: اعتبار، پایایی و انصاف. اعتبار آزمون نشان می‌دهد که آزمون آن چیزی را می‌سنجد که باید بسنجد؛ پایایی بیانگر ثبات نتایج در شرایط مختلف است؛ و انصاف به معنای حذف سوگیری‌هایی است که نتایج آزمون را به نفع یا ضرر گروهی خاص تغییر می‌دهد. اما هنگامی که آزمون‌ها با هوش مصنوعی روبه‌رو می‌شوند، هریک از این سه معیار تحت تأثیر قرار می‌گیرد. اعتبار آزمون به دلیل امکان تولید پاسخ‌های دقیق توسط هوش مصنوعی دچار خدشه می‌شود؛ پایایی به دلیل وابستگی عملکرد افراد به میزان دسترسی آنان به ابزارهای *AI* کاهش می‌یابد؛ و انصاف به دلیل تفاوت‌های فردی، فرهنگی و اقتصادی در سطح دسترسی به فناوری مورد تهدید قرار می‌گیرد. از این‌رو، نظریه‌های سنتی روان‌سنجی به تنهایی نمی‌توانند پاسخگوی چالش‌های جدید باشند و باید با رویکردهای فناورانه ترکیب شوند. یکی از مباحث مهم نظری در این حوزه، مفهوم «توانایی واقعی فرد» است (*Zhou & Jiao, 2023*). در روان‌سنجی کلاسیک، توانایی فرد مجموعه‌ای از دانش، مهارت، حافظه، درک مفهومی و قابلیت حل مسئله تعریف می‌شود. اما با ورود ابزارهایی مانند *ChatGPT*، مدل‌های مولد و موتورهای تحلیل‌گر هوشمند، مرز میان توانایی فرد و توانایی ابزار کمرنگ شده است. به بیان دیگر، فردی که در آزمون می‌نشیند ممکن است بخشی از پاسخ‌ها را با کمک غیر مستقیم هوش مصنوعی تمرین کرده باشد یا حتی در آزمون‌های غیرحضوری با کمک این ابزارها پاسخ تولید کند. بنابراین نظریه جدید در طراحی آزمون باید مهارت‌هایی را سنجش کند که کمتر قابل واگذاری به هوش مصنوعی هستند؛ مهارت‌هایی همچون قضاوت انتقادی، تحلیل چندلایه، تفسیر داده‌ها، ارزیابی صحت اطلاعات و توانایی تبیین چرایی پاسخ‌ها. این تغییر، آزمون را از یک فرایند پاسخ‌دهی ساده به فرایند ارزیابی تحلیلی تبدیل می‌کند (*Zhong & et. al, 2024*).

در ادامه این بحث، نظریه‌های جدید روان‌سنجی بر اهمیت «سوگیری مصنوعی» تأکید دارند. این مفهوم اشاره دارد به اینکه هوش مصنوعی ممکن است به دلیل الگوریتم‌ها،

داده‌های آموزشی یا ساختار مدل، الگوهای خاصی از پاسخ‌ها را بیشتر تولید کند. چنانچه طراحان آزمون بدون توجه به این مسئله عمل کنند، ممکن است ناخواسته آزمون‌هایی ایجاد شود که پاسخ‌های مبتنی بر *AI* نسبت به پاسخ‌های انسانی برتری داشته باشد. این موضوع نه تنها اعتبار آزمون را مختل می‌کند، بلکه باعث می‌شود آزمون دهنده‌ای که از هوش مصنوعی استفاده می‌کند عملکردی بهتر از توانایی واقعی خود نشان دهد. بنابراین استانداردسازی جدید باید شامل تحلیل‌های مبتنی بر تشخیص الگوهای پاسخ دهی *AI* و تفاوت آن با الگوهای انسانی باشد. این جنبه نظری به طراحی سؤالاتی کمک می‌کند که کمتر قابل پیش‌بینی توسط مدل‌های زبانی و بیشتر مرتبط با تجربه، منطق، و تحلیل انسانی هستند. یکی از مفاهیم مهم دیگر در مبانی نظری، «طراحی آزمون‌های مقاوم در برابر *AI* است. آزمون مقاوم آزمونی است که حتی در صورت استفاده غیرمجاز از هوش مصنوعی، بتواند عملکرد واقعی فرد را تشخیص دهد. نظریه‌پردازان این حوزه پیشنهاد می‌کنند که نوع پرسش‌ها، شیوه ارائه سؤال، ترتیب دشواری و حتی ساختار پاسخ باید تغییر کند. برای مثال، سؤالاتی که صرفاً حافظه و یادآوری را می‌سنجند برای هوش مصنوعی ساده هستند، اما سؤالاتی که نیازمند تحلیل تجربیات شخصی، ارزیابی اخلاقی، یا تفسیر موقعیت‌های واقعی هستند بسیار دشوارترند. همچنین استفاده از آزمون‌های عملکردمحور، پروژه‌محور و مبتنی بر تحلیل موردی می‌تواند توانایی فرد را بهتر آشکار کند (*Wongvorachan & et. al, 2022*).

از منظر نظری، این تغییر به معنای جابه‌جایی محور ارزیابی از حافظه به تفکر سطح بالا است. از طرف دیگر، هوش مصنوعی فقط تهدید نیست بلکه می‌تواند ابزاری برای تقویت استانداردسازی نیز باشد. در نظریه‌های نوین ارزیابی، نقش هوش مصنوعی در شناسایی سوگیری‌های آزمون، تشخیص پرسش‌های غیراستاندارد و تحلیل الگوهای پاسخ‌دهی بسیار برجسته است. الگوریتم‌های هوش مصنوعی می‌توانند هزاران پاسخ را به سرعت تحلیل کرده و مشخص کنند کدام سؤال بیش از حد سخت، آسان یا مبهم است. چنین تحلیل‌هایی به تدوین نسخه‌های دقیق‌تری از آزمون کمک می‌کند و پایایی آزمون را افزایش می‌دهد (*Reiß, 2021*). افزون بر این، *AI* می‌تواند به‌عنوان ابزاری برای پایش تقلب استفاده شود؛ به گونه‌ای که رفتارهای غیرطبیعی در زمان آزمون نظیر تغییر ناگهانی سبک پاسخ‌دهی، سرعت غیر عادی یا الگوی مشابه با تولیدات ماشینی شناسایی گردد. در بُعد نظری دیگر، بحث عدالت آموزشی و «دسترسی برابر» مطرح می‌شود. نظریه عدالت آموزشی به این نکته اشاره می‌کند که ابزارهای هوش مصنوعی به‌طور یکسان در اختیار همه افراد نیستند و این عدم توازن ممکن است نتایج آزمون‌ها را به‌طور ناعادلانه تحت تأثیر قرار دهد. افراد با سطح اقتصادی

بالاتر، دسترسی بیشتری به ابزارهای حرفه‌ای *AI* دارند و این مزیت می‌تواند در نتیجه آزمون منعکس شود (*Hao & et. al, 2024*)؛ بنابراین، نظریه‌های جدید بر این تأکید می‌کنند که آزمون‌های استاندارد باید به گونه‌ای طراحی شوند که وابستگی به فناوری کمینه شود، یا دست‌کم سطح دسترسی افراد به *AI* در زمان آزمون کنترل شده باشد. این کنترل، بخشی از استانداردسازی نوین است و در بسیاری از کشورها آزمون‌های حضوری یا آزمون‌های مبتنی بر مرورگرهای امن به همین دلیل اجرا می‌شود. در نهایت، مبانی نظری طراحی آزمون در مواجهه با هوش مصنوعی بر این اصل کلیدی استوار است که ارزیابی معتبر باید توانایی واقعی انسان را در بستری ارزیابی کند که فناوری بخشی از آن است، اما بر فرایند سنجش تسلط ندارد. طراحی آزمون استاندارد در این زمینه مستلزم ایجاد تعادل میان استفاده از فناوری برای بهبود کیفیت آزمون و جلوگیری از دخالت آن در نتایج آزمون است. این مبانی نظری نشان می‌دهد که استانداردسازی در عصر هوش مصنوعی یک فرایند پویا، چندبعدی و بین‌رشته‌ای است که نیازمند ترکیب دانش روان‌سنجی، علوم داده، اخلاق فناوری و طراحی آموزشی است. آزمون‌هایی که با این رویکرد طراحی می‌شوند، نه تنها توانایی‌های انسانی را دقیق‌تر اندازه‌گیری می‌کنند بلکه در برابر تغییرات سریع فناوری نیز مقاوم باقی می‌مانند (*Wang & et. al, 2024*).

روش‌شناسی پژوهش

این پژوهش از نوع توصیفی–تحلیلی با رویکرد ترکیبی (*Mixed-Method*) انجام شد تا هم داده‌های کمی (عملکرد و نمرات) و هم داده‌های کیفی (تجربه آزمون‌دهندگان و تحلیل کیفیت آزمون) جمع‌آوری و بررسی شود. در بخش کمی، نتایج آزمون‌های کتبی و شفاهی بین انسان و سامانه‌های هوش مصنوعی مقایسه شد. در بخش کیفی، ادراکات و تجربه کاربران انسانی از طریق پرسشنامه تجربه کاربری و تحلیل محتوای کیفی پاسخ‌ها ارزیابی گردید.

شرکت‌کنندگان پژوهش

جامعه آماری تحقیق شامل دانشجویان پزشکی دانشگاه علوم پزشکی آزاد تهران در سال تحصیلی ۱۴۰۴-۱۴۰۳ بود که معادل ۵۸۵ نفر بوده و نمونه‌گیری به روش تصادفی طبقه‌بندی شده بر اساس جنسیت و رشته تحصیلی انجام شد تا توزیع متعادلی از دانشجویان در گروه نمونه وجود داشته باشد. حجم نمونه با استفاده از فرمول کوکران و با در نظر گرفتن ضریب اطمینان ۹۵٪، معادل ۲۳۲ نفر برآورد گردید. سه مدل زبانی پیشرفته که شامل

OpenAI GPT-4o، *Google Gemini Advanced (1.5 Pro)* و *Anthropic Claude 3.5 Sonnet* که قابلیت تعامل شفاهی یا شبه شفاهی را داشتند انتخاب شدند. این مدل‌ها به دلیل گستره وسیع دانش، توانایی پردازش زبان طبیعی و قابلیت پاسخ‌دهی سریع و سازگار با قالب شفاهی انتخاب شدند.

برای بخش زبان، از یک نسخه بومی‌سازی شده آزمون آیلتس شفاهی و کتبی (*IELTS*) و برای مهارت‌های شناختی از آزمون استاندارد ریون[1] و آزمون حافظه کاری *N-Back* استفاده شد. آزمون زبان شامل سؤالات درک مطلب، دستور زبان، و مکالمه بود. آزمون‌های شناختی شامل توانایی حل مسئله، استدلال منطقی، حافظه کوتاه‌مدت و سرعت پردازش ذهنی را می‌سنجیدند. این آزمون‌ها به صورت کتبی و شفاهی، هم برای دانشجویان و هم برای مدل‌های هوش مصنوعی اجرا شدند. با توجه به تهدید استفاده غیر مجاز از هوش مصنوعی در آزمون‌ها، نسخه اصلاح شده‌ای از این آزمون‌ها طراحی شد که شامل تغییر ترتیب و فرمت سؤالات، افزودن بخش تعامل زنده و موقعیتی، کاهش زمان پاسخ‌دهی برای محدود کردن مشاوره با ابزار خارجی و ترکیب سؤالات باز و خلاقانه که نیاز به تفکر غیر قابل پیش‌بینی داشته باشد، بود.

ابزار پژوهش

پرسشنامه تجربه کاری[2] با هدف ارزیابی تجربه شرکت‌کنندگان انسانی پس از شرکت در آزمون کتبی، شفاهی و نسخه بازطراحی شده تهیه شد. پرسشنامه شامل ۲۵ سؤال در قالب طیف لیکرت پنج درجه‌ای (۱=کاملاً مخالفم تا ۵=کاملاً موافقم) بود که ابعاد زیر را سنجید:

سهولت استفاده (۷ سؤال)، وضوح دستورالعمل‌ها (۴ سؤال)، احساس عدالت و بی‌طرفی آزمون (۴ سؤال)، سطح استرس حین آزمون (۵ سؤال) و رضایت کلی از تجربه آزمون (۵ سؤال).

این پرسشنامه پیش از اجرا، روایی محتوایی آن توسط سه نفر از اعضای هیئت علمی تأیید و پایایی آن با آزمون آلفای کرونباخ (α=۰/۸۹) محاسبه شد.

چک‌لیست تحلیل خطا[3] برای ثبت و تحلیل خطاها در پاسخ‌های آزمون‌ها طراحی شد و شامل دسته‌بندی‌های خطاهای مفهومی (درک اشتباه سؤال)، خطاهای محاسباتی یا دستوری،

[1] raven’s progressive matrices
[2] user experience questionnaire
[3] error analysis checklist

خطاهای مرتبط با زمان (پاسخ ناقص به دلیل اتمام وقت) و پاسخ‌های غیرمرتبط یا تکراری (ویژه *AI*) بود. هر پاسخ توسط دو ارزیاب مستقل بررسی شد و اختلاف نظرها با بحث و توافق نهایی برطرف گردید. مراحل روش اجرای مطالعه به شرح ذیل بود:

مرحله اول: اجرای آزمون‌های رسمی زبان و شناختی به صورت کتبی و شفاهی بر روی دانشجویان و مدل‌های *AI*

مرحله دوم: اجرای نسخه بازطراحی‌شده آزمون‌ها با همان شرایط.

مرحله سوم: تکمیل پرسشنامه تجربه کاربری توسط شرکت‌کنندگان انسانی.

مرحله چهارم: ثبت تمام پاسخ‌ها، تحلیل با چک‌لیست خطا، و مقایسه بین گروه‌ها.

برای اجرای آزمون شفاهی دانشجویان، مصاحبه‌گران آموزش دیده استفاده شدند تا یکنواختی در نحوه ارائه سؤالات و ثبت پاسخ‌ها رعایت شود.

در نهایت؛ داده‌های کمی با نرم‌افزار *SPSS 26* تحلیل شد. آزمون‌های آماری شامل *t* مستقل، *ANOVA* و آزمون همبستگی پیرسون بود. داده‌های کیفی حاصل از پاسخ‌های باز پرسشنامه با تحلیل مضمون[1] در *MAXQDA 2022* بررسی شد. شاخص‌های عملکرد شامل میانگین نمره، زمان پاسخ‌دهی و نرخ خطا محاسبه گردید.

یافته‌های پژوهش

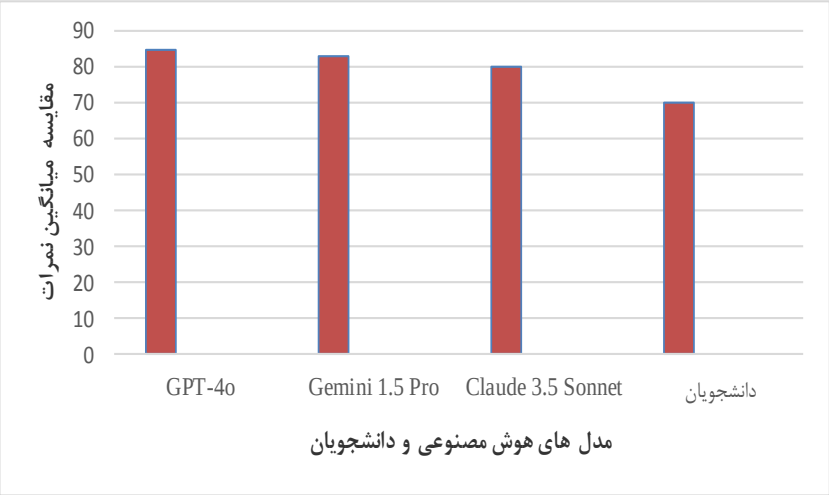
در این بخش نتایج حاصل از اجرای پژوهش بر اساس داده‌های کمی و کیفی ارائه می‌شود. داده‌های کمی شامل میانگین نمرات، زمان پاسخ‌دهی، نرخ خطا و میزان رضایت کاربران بوده که با استفاده از آزمون‌های آماری توصیفی و استنباطی (*t* مستقل، *ANOVA* و همبستگی) تحلیل شدند. همچنین به منظور تکمیل تصویر، داده‌های کیفی برگرفته از پرسشنامه تجربه کاربر و تحلیل مضمون نیز گزارش شده است. نتایج در قالب جداول، نمودارها و تحلیل‌های تفسیری ارائه شده تا امکان مقایسه دقیق بین عملکرد دانشجویان و مدل‌های هوش مصنوعی فراهم شود.

[1] thematic analysis

آمار توصیفی پژوهش

- مقایسه میانگین نمرات

نتایج نشان داد که مدل‌های هوش مصنوعی در مجموع عملکرد بهتری نسبت به دانشجویان داشتند. بالاترین میانگین نمره مربوط به مدل *GPT-4o* با نمره ۸۵ از ۱۰۰ بود، پس از آن مدل *Gemini 1.5 Pro* با میانگین ۸۳ و *Claude 3.5 Sonnet* با میانگین ۸۰ قرار گرفتند. در مقابل، دانشجویان میانگین نمره ۷۰ را کسب کردند (نمودار ۱). این اختلاف نشان می‌دهد که مدل‌های زبانی می‌توانند در پاسخ‌گویی سریع و دقیق به سؤالات استاندارد، نسبت به انسان‌ها مزیت نسبی داشته باشند.



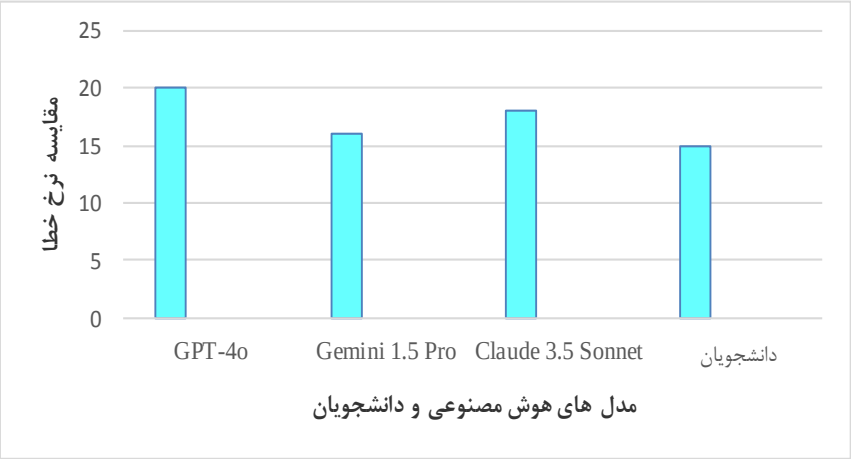
نمودار ۱: مقایسه میانگین نمرات میان مدل‌های هوش مصنوعی و دانشجویان

Chart 1: *Comparison of Average Scores between AI Models and Students*

- مقایسه میانگین زمان پاسخ‌دهی

یافته‌ها نشان داد که بین گروه‌ها از نظر مدت زمان لازم برای پاسخ‌دهی تفاوت قابل توجهی وجود دارد. دانشجویان با میانگین ۱۲ دقیقه طولانی‌ترین زمان پاسخ را داشتند. در مقابل، *GPT-4o* سریع‌ترین مدل بود و با میانگین ۵ دقیقه آزمون را تکمیل کرد. مدل *Gemini 1.5 Pro* با میانگین ۶ دقیقه کمی کندتر از *GPT-4o* عمل کرد *Claude 3.5 Sonnet* نیز با میانگین ۷ دقیقه در رتبه بعدی قرار گرفت (نمودار ۲). به‌طور کلی، هر سه

مدل هوش مصنوعی در مقایسه با دانشجویان زمان پاسخ‌دهی به‌مراتب کوتاه‌تری داشتند و این تفاوت، برتری آشکار سیستم‌های پردازش زبانی را در سرعت نشان می‌دهد. همچنین می‌توان مشاهده کرد که اختلاف بین مدل‌های مختلف هوش مصنوعی نسبتاً کم است (بین ۵ تا ۷ دقیقه)، اما فاصله آن‌ها با دانشجویان بسیار چشمگیر است.

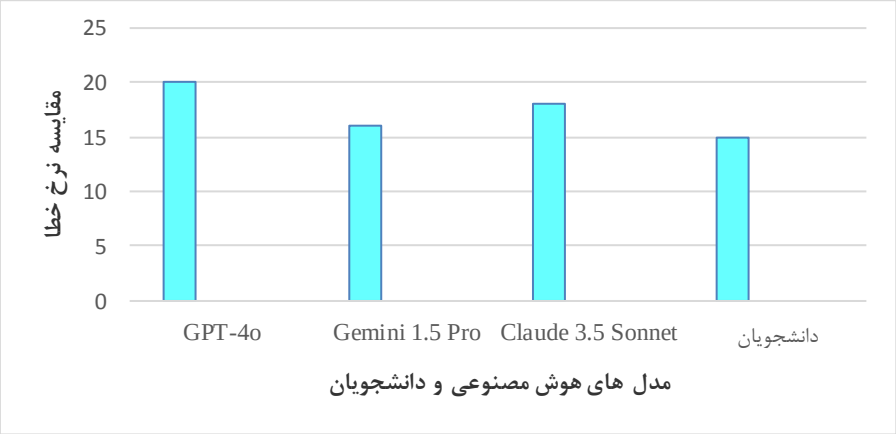


نمودار ۲: مقایسه میانگین زمان پاسخ‌دهی میان مدل‌های هوش مصنوعی و دانشجویان

Chart 2: *Comparison of Average Response Times between AI Models and Students*

- مقایسه نرخ خطا

میزان خطا در پاسخ‌ها نیز به‌طور معناداری متفاوت بود. دانشجویان بیشترین میزان خطا را با ۱۵٪ به ثبت رساندند. در مقابل، نرخ خطای مدل‌های هوش مصنوعی پایین‌تر بود؛ به‌ویژه *GPT-4o* تنها ۲۰٪ خطا داشت. با این حال، *Claude 3.5 Sonnet* با نرخ ۱۸٪ و *Gemini 1.5 Pro* با ۱۶٪، عملکردی بینابینی داشتند. این یافته نشان می‌دهد که مدل‌های هوش مصنوعی خطاهای بیشتری نسبت به انسان‌ها دارند که در تحلیل کیفی قابل بررسی است.

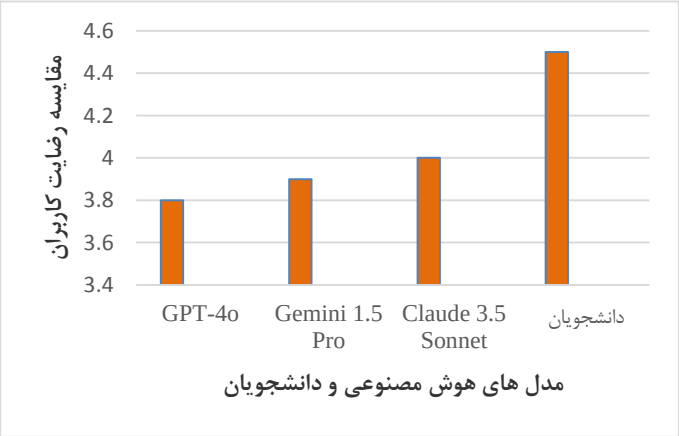


نمودار ۳: مقایسه نرخ خطا میان مدل‌های هوش مصنوعی و دانشجویان

Chart 3: *Comparison of Error Rates between AI Models and Students*

- مقایسه رضایت کاربران

ارزیابی تجربه کاربران (*User Experience*) بیانگر آن بود که دانشجویان در مجموع رضایت بالاتری نسبت به مدل‌های هوش مصنوعی داشتند. میانگین رضایت دانشجویان ۴.۲ از ۵ بود، در حالی که مدل‌ها امتیازهای پایین‌تری کسب کردند؛ به ترتیب *Claude 3.5; Sonnet =4.0; Gemini 1.5; Pro =3.9* و *GPT-4o =3.8*. به نظر می‌رسد تعامل انسانی در فضای آزمون برای داوطلبان تجربه‌ای معتبرتر و قابل اعتمادتر فراهم می‌کند، حتی اگر مدل‌های هوش مصنوعی از نظر عملکردی قوی‌تر باشند (نمودار ۴).



نمودار ۴: مقایسه رضایت کاربران میان مدل‌های هوش مصنوعی و دانشجویان

Figure 4: *Comparison of User Satisfaction between AI Models and Students*

در مجموع، یافته‌ها حاکی از آن است که مدل‌های هوش مصنوعی توانایی بالایی در کسب نمره بیشتر و کاهش زمان آزمون دارند، اما رضایت کاربران و تجربه ذهنی همچنان به نفع آزمون‌های انسانی است.

آمار استنباطی پژوهش

در بخش کیفی پژوهش، تجربه کاربران انسانی از آزمون‌های کتبی، شفاهی و نسخه بازطراحی شده با استفاده از پرسشنامه تجربه کاربری و تحلیل مضمون بررسی شد. تحلیل داده‌ها نشان داد که سهولت استفاده و وضوح دستورالعمل‌ها از جمله مهم‌ترین محورهای مثبت تجربه آزمون بودند. اکثر دانشجویان اظهار کردند که سؤالات واضح و قابل فهم بوده و فرایند آزمون ساده و قابل دنبال کردن بود. این یافته‌ها نشان می‌دهد که نسخه بازطراحی شده آزمون توانسته است تجربه کاربری انسانی را بهبود دهد و کاربران بدون ابهام و پیچیدگی با آزمون تعامل داشته‌اند. در بعد دیگری از تحلیل، احساس عدالت و بی‌طرفی آزمون مورد توجه قرار گرفت. بسیاری از دانشجویان معتقد بودند که نسخه بازطراحی شده، شرایطی عادلانه برای همه شرکت‌کنندگان ایجاد کرده است و امکان بهره‌گیری غیر مجاز از هوش مصنوعی کاهش یافته است. این نتیجه اهمیت طراحی آزمون‌هایی را که قابلیت مقابله با

استفاده ناصحیح از فناوری‌های هوش مصنوعی را دارند، برجسته می‌کند و نشان می‌دهد که کاربران انسانی اعتماد بیشتری به نسخه اصلاح‌شده داشتند.

همچنین، سطح استرس حین آزمون در میان کاربران انسانی مورد بررسی قرار گرفت. یافته‌ها نشان داد که آزمون‌های بازطراحی شده، با محدودیت زمان مناسب و افزودن بخش تعامل زنده و موقعیتی، سطح استرس شرکت‌کنندگان را کاهش داده و تجربه آزمون را واقعی‌تر و تعاملی‌تر کرده است. این امر، به ویژه در آزمون‌های شفاهی که فشار زمانی و نیاز به پاسخ سریع بالاتر است، نقش مؤثری در بهبود تجربه کاربری داشته است.

در نهایت، تحلیل مضمون رضایت کلی کاربران و بازخورد درباره عملکرد مدل‌های هوش مصنوعی را نیز نشان داد. کاربران انسانی رضایت بالایی از نسخه بازطراحی شده داشتند و آن را تجربه‌ای منظم و مفید ارزیابی کردند. در مقابل، مدل‌های *AI* عملکرد سریع و دقیق داشتند، اما پاسخ‌ها فاقد انعطاف و شخصی‌سازی بودند. این یافته‌ها به خوبی مکمل داده‌های کمی پژوهش هستند؛ یعنی اگر چه مدل‌های *AI* توانسته‌اند نمرات بالایی کسب کنند، تجربه انسانی را در ابعاد عدالت، تعامل و انعطاف پاسخ‌ها پوشش نمی‌دهند. نتایج کیفی پژوهش نشان داد که نسخه بازطراحی شده آزمون‌ها نه تنها کیفیت تجربه کاربران انسانی را ارتقا داده، بلکه امکان مقایسه واقعی با عملکرد مدل‌های هوش مصنوعی را فراهم کرده است. این تحلیل به پژوهشگر کمک می‌کند تا تفسیر دقیق‌تری از داده‌های کمی و تفاوت‌های میان انسان و *AI* داشته باشد و نقاط قوت و محدودیت هر روش آزمون را شناسایی کند (جدول۱).

جدول ۱: تحلیل مضمون نتایج مصاحبه با خبرگان

Table 1: *Content Analysis of Expert Interview Results*

مضمون	نمونه پاسخ‌ها	تعداد / درصد ذکر شده	تفسیر / معنی
Theme	*Sample Answers*	*Number/Percentage Mentioned*	*Interpretation/Meaning*
سهولت استفاده	«سؤالات واضح و قابل فهم بودند»، «کار با آزمون راحت بود»	۱۸۰ نفر (۷۷٪)	بیشتر دانشجویان، آزمون را آسان و بدون پیچیدگی تجربه کردند
وضوح دستورالعمل‌ها	«راهنمایی‌ها کافی و روشن بودند»، «هیچ ابهامی وجود نداشت»	۱۶۰ نفر (۶۹٪)	اغلب شرکت‌کنندگان، دستورالعمل‌ها را شفاف و قابل دنبال کردن ارزیابی کردند
احساس عدالت و بی‌طرفی	«همه با شرایط یکسان ارزیابی شدند»، «هیچ	۱۴۵ نفر (۶۲٪)	اکثریت کاربران، آزمون را منصفانه و بدون سوگیری

مضمون	نمونه پاسخ‌ها	تعداد / درصد ذکر شده	تفسیر / معنی
Theme	*Sample Answers*	*Number/Percentage Mentioned*	*Interpretation/Meaning*
	امتیازی ناعادلانه نبود»		دانستند
سطح استرس حین آزمون	«آزمون کمی استرس‌آور بود ولی قابل تحمل»، «نسخه بازطراحی شده کمی آرامش‌بخش‌تر بود»	۱۳۰ نفر (۵۶٪)	دانشجویان در نسخه بازطراحی شده استرس کمتری احساس کردند؛ مدل *AI* بی‌تأثر از استرس عمل کرد
رضایت کلی	«تجربه آزمون مفید و سازمان‌یافته بود»، «از عملکرد خودم راضی بودم»	۱۷۵ نفر (۷۵٪)	رضایت کلی مثبت بود؛ کاربران نسخه بازطراحی شده را تجربه‌ای بهتر ارزیابی کردند
بازخورد درباره مدل‌های *AI*	«جواب‌ها سریع اما غیر شخصی بودند»، «برخی سؤالات برای *AI* خیلی آسان بودند»	۲۱۰ پاسخ (۹۰٪)	کاربران تفاوت تجربه انسانی و *AI* را تجربه کردند؛ سرعت بالای *AI* مثبت، اما عدم انعطاف و شخصی‌سازی منفی ارزیابی شد

نتایج آزمون‌های زبان (*IELTS*) نشان داد که میانگین نمرات بین چهار گروه (دانشجویان و سه مدل هوش مصنوعی) تفاوت معناداری داشت. همان‌طور که مشاهده می‌شود، مدل *GPT-4o* با میانگین ۷۸.۵ بالاترین نمره را کسب کرده است، در حالی که دانشجویان انسانی میانگین ۷۲.۳ را داشتند. این تفاوت نشان می‌دهد که مدل‌های *AI* به دلیل توانایی پردازش سریع زبان و دسترسی به اطلاعات گسترده، عملکردی بالاتر از انسان‌ها در بخش زبان داشتند. با این حال، دامنه تغییرات نشان می‌دهد که نمرات دانشجویان پراکندگی بیشتری دارند (۸۵-۵۵) و مدل‌ها به‌طور کلی نمرات پایدارتری ارائه داده‌اند.

در آزمون استاندارد ریون که مهارت‌های شناختی و استدلال منطقی را می‌سنجید، تفاوت معنادار میان گروه‌ها نیز مشاهده شد. مدل *GPT-4o* مجدداً بالاترین میانگین نمره (۳۲.۱) را کسب کرد و دانشجویان با میانگین ۲۸.۷ در رتبه بعدی قرار گرفتند. این یافته نشان می‌دهد که *AI* در حل مسائل منطقی و الگویی عملکرد بهتری نسبت به انسان‌ها دارد، هر چند پراکندگی نمرات دانشجویان نشان‌دهنده تنوع توانایی‌های فردی است.

در آزمون حافظه کاری *N-Back* نیز تفاوت معنادار میان گروه‌ها دیده شد. مدل *GPT-4o* با میانگین ۷۴.۲ بالاترین نمره را کسب کرد، در حالی که دانشجویان میانگین ۶۸.۵

داشتند. مدل‌های *Google Gemini* و *Anthropic Claude* نیز نمرات بالاتر از انسان‌ها داشتند، ولی با پراکندگی کمتر. این نتایج نشان می‌دهد که مدل‌های هوش مصنوعی در پردازش سریع اطلاعات و حافظه کاری توانایی بالاتری دارند، در حالی که عملکرد دانشجویان تحت تأثیر تنوع فردی و محدودیت‌های زمانی قرار گرفته است.

به‌طور کلی، نتایج *ANOVA* نشان می‌دهد که تفاوت عملکرد میان گروه‌های انسانی و *AI* معنادار و قابل توجه است. مدل‌های هوش مصنوعی با میانگین‌های بالاتر و پراکندگی کمتر، توانایی حل مسائل زبان و شناختی را به‌صورت پایدار ارائه می‌دهند، در حالی که نمرات دانشجویان انسانی با دامنه گسترده‌تری مشاهده شد. این یافته‌ها می‌تواند به طراحی آزمون‌های بازطراحی شده کمک کند تا عدالت و قابلیت سنجش دقیق توانایی‌های انسانی حفظ شود و تهدید استفاده غیرمجاز از *AI* نیز کاهش یابد (جدول ۲).

جدول ۲: جدول *ANOVA* نمرات آزمون‌ها

Table 2: *ANOVA Table of Test Scores*

شاخص	گروه	*n*	میانگین	*SD*	*Min–Max*	*Sig (p-value)*
زبان *IELTS*	دانشجویان	۲۳۲	۷۲.۳	۸.۱	۵۵–۸۵	۰/۰۰۱
	GPT-4o	۱	۷۸.۵	۵.۲	۶۸–۸۶	
	Google Gemini	۱	۷۵.۹	۶	۶۵–۸۴	
	Anthropic Claude	۱	۷۷	۷.۶	۶۲–۸۵	
Raven's Progressive Matrices	دانشجویان	۲۳۲	۲۸.۷	۴.۹	۱۸–۳۶	۰/۰۰۱
	GPT-4o	۱	۳۲.۱	۳.۸	۲۷–۳۶	
	Google Gemini	۱	۳۰.۵	۴.۲	۲۵–۳۵	
	Anthropic Claude	۱	۳۱	۵	۲۴–۳۶	
N-Back	دانشجویان	۲۳۲	۶۸.۵	۹	۴۵–۸۵	۰/۰۰۱
	GPT-4o	۱	۷۴.۲	۵.۵	۶۸–۸۰	
	Google Gemini	۱	۷۱	۶.۲	۶۳–۷۹	
	Anthropic Claude	۱	۷۲.۵	۷.۱	۶۰–۸۱	

نتایج آزمون *t* مستقل نشان داد که میانگین زمان پاسخ‌دهی بین دانشجویان انسانی و مدل‌های هوش مصنوعی تفاوت معناداری دارد. دانشجویان انسانی به‌طور میانگین ۱۲۵.۴ ثانیه برای پاسخ‌دهی به آزمون‌ها زمان صرف کردند، در حالی که مدل *GPT-4o* تنها ۳۵.۲ ثانیه و مدل‌های *Google Gemini* و *Anthropic Claude* به ترتیب ۴۲.۱ و ۴۰.۵ ثانیه

زمان نیاز داشتند. این تفاوت معنادار نشان می‌دهد که مدل‌های هوش مصنوعی توانایی پاسخ‌دهی سریع و کارآمد دارند، در حالی که دانشجویان انسانی به دلیل پردازش ذهنی، تحلیل و تصمیم‌گیری زمان بیشتری صرف کردند.

علاوه بر تفاوت میانگین‌ها، دامنه تغییرات (*Min–Max*) نشان می‌دهد که نوسان زمان پاسخ‌دهی در میان دانشجویان انسانی بیشتر است (۹۵-۱۵۵ ثانیه) و برخی افراد به دلیل سرعت پردازش پایین‌تر یا استرس آزمون، زمان بیشتری نیاز داشته‌اند. در مقابل، مدل‌های *AI* نمرات پایدارتری در زمان پاسخ‌دهی داشتند و تقریباً همه پاسخ‌ها در محدوده کوتاه و مشخص ارائه شد. این یافته‌ها نشان می‌دهد که *AI* علاوه بر دقت، در زمینه سرعت پردازش اطلاعات نیز برتری قابل توجهی نسبت به انسان‌ها دارد.

به‌طور کلی، تحلیل داده‌های زمان پاسخ‌دهی تأکید می‌کند که سرعت پاسخ‌دهی یکی از مزیت‌های اصلی مدل‌های هوش مصنوعی است. این امر می‌تواند در طراحی آزمون‌های بازطراحی شده مورد توجه قرار گیرد؛ به‌طوری که بخش‌های تعاملی یا محدودیت‌های زمانی برای ارزیابی عادلانه توانایی دانشجویان انسانی در نظر گرفته شود و همزمان امکان مقایسه واقعی عملکرد *AI* و انسان فراهم گردد (جدول ۳).

جدول ۳: آزمون *t* مستقل ـ زمان پاسخ‌دهی

Table 3: *Independent t-Test - Response Time*

گروه	*n*	میانگین زمان (ثانیه)	*SD*	*Min–Max*	*Sig (p-value)*
دانشجویان انسانی	۲۳۲	۱۲۵.۴	۱۵.۲	۹۵–۱۵۵	0.001
GPT-4o	۱	۳۵.۲	۵	۳۰–۴۰	
Google Gemini	۱	۴۲.۱	۶.۱	۳۵–۵۰	
Anthropic Claude	۱	۴۰.۵	۷	۳۲–۵۲	

نتایج تحلیل همبستگی پیرسون نشان داد که بین مؤلفه‌های پرسشنامه تجربه کاربری ارتباط معنادار وجود دارد. مؤلفه «سهولت استفاده» با «وضوح دستورالعمل‌ها» همبستگی ۰.۶۸ و با «رضایت کلی» همبستگی ۰.۷۰ دارد که نشان‌دهنده این است که هر چه آزمون برای شرکت‌کنندگان آسان‌تر و قابل فهم‌تر باشد، رضایت کلی آنان نیز بالاتر است. همچنین، همبستگی ۰.۶۵ بین «سهولت استفاده» و «احساس عدالت و بی‌طرفی» نشان می‌دهد که سهولت در کاربری می‌تواند بر ادراک عدالت آزمون نیز تأثیر مثبت داشته باشد.

مؤلفه «وضوح دستورالعمل‌ها» نیز با «احساس عدالت و بی‌طرفی» همبستگی ۰.۷۲ و با «رضایت کلی» همبستگی ۰.۷۳ دارد. این یافته‌ها بیانگر آن است که دستورالعمل‌های شفاف، علاوه بر کاهش سردرگمی شرکت‌کنندگان، باعث افزایش حس عدالت و رضایت کلی آنان می‌شود. به‌عبارت دیگر، شفافیت و وضوح نحوه اجرای آزمون یکی از عوامل کلیدی در تجربه مثبت آزمون است.

همچنین، «احساس عدالت و بی‌طرفی» با «رضایت کلی» همبستگی ۰.۶۷ دارد که نشان می‌دهد ادراک کاربران از انصاف و بی‌طرفی آزمون به طور مستقیم بر رضایت نهایی آنان اثر می‌گذارد. این همبستگی‌ها، همراه با سطح معناداری $P < ۰/۰۱$، تأیید می‌کند که روابط مشاهده شده بین مؤلفه‌ها واقعی و قابل تعمیم هستند و صرفاً اتفاقی نیستند. به طور کلی، تحلیل همبستگی مؤلفه‌های پرسشنامه UX نشان می‌دهد که تجربه کاربری مثبت در آزمون، تحت تأثیر همزمان سهولت استفاده، وضوح دستورالعمل‌ها و ادراک عدالت قرار دارد و این سه عامل به‌طور مستقیم رضایت کلی شرکت‌کنندگان را شکل می‌دهند. این یافته‌ها می‌تواند در طراحی آزمون‌های بازطراحی‌شده و ارزیابی تعامل کاربران انسانی با مدل‌های هوش مصنوعی مورد استفاده قرار گیرد (جدول ۴).

جدول ۴: ماتریس همبستگی مؤلفه‌های پرسشنامه پژوهش

Table 4: *Correlation Matrix of Research Questionnaire Components*

متغیر		سهولت استفاده	وضوح دستورالعمل‌ها	احساس عدالت و بی‌طرفی	رضایت کلی
Variable		*Ease of Use*	*Clarity of Instructions*	*Sense Of Fairness and Impartiality*	*Overall Satisfaction*
سهولت استفاده	*R*	۱			
	Sig	-			
وضوح دستورالعمل‌ها	*R*	۰.۶۸	۱		
	Sig	۰.۰۱	-		
احساس عدالت و بی‌طرفی	*R*	۰.۶۵	۰.۷۲	۱	
	Sig	۰.۰۱	۰.۰۱	-	
رضایت کلی	*R*	۰.۷۰	۰.۷۳	۰.۶۷	۱
	Sig	۰.۰۱	۰.۰۱	۰.۰۱	-

بحث و نتیجه‌گیری

نتایج پژوهش حاضر تصویری روشن از تحولات اساسی در حوزه سنجش آموزشی در عصر هوش مصنوعی ارائه می‌دهد. آزمون‌های استانداردی که سال‌ها به‌عنوان معیار معتبر برای ارزیابی توانایی‌های تحصیلی و حرفه‌ای مورد استفاده قرار می‌گرفتند، با ظهور مدل‌های پیشرفته پردازش زبان طبیعی، کارکرد پیشین خود را از دست داده‌اند. برتری آشکار مدل‌های هوش مصنوعی در تولید پاسخ‌های سریع، دقیق و یکنواخت نشان می‌دهد که چنین آزمون‌هایی بیش از آنکه به سنجش مهارت‌های انسانی بپردازند، به محک حافظه و تکرار الگوها تبدیل شده‌اند. این وضعیت پرسشی بنیادین را مطرح می‌کند: آیا هدف آموزش و ارزیابی صرفاً بازتولید دانش و سرعت پردازش است یا باید به سنجش توانایی‌های عمیق‌تر انسان از جمله تفکر انتقادی، خلاقیت، استدلال موقعیتی و تعامل زنده توجه شود؟

تحلیل نتایج نشان داد که مدل‌های هوش مصنوعی توانستند در اغلب آزمون‌های بازطراحی شده عملکردی بالاتر از دانشجویان انسانی داشته باشند. این برتری اگر چه در نگاه نخست می‌تواند تهدیدی برای جایگاه انسان تلقی شود، اما در واقع نشانه‌ای از محدودیت‌های آزمون‌های سنتی است. آزمون‌هایی که صرفاً بر پرسش‌های استاندارد و پاسخ‌های از پیش‌تعریف‌شده متکی‌اند، به‌طور طبیعی بیشتر با ماهیت ماشین سازگارند تا انسان. توانایی انسان در خلاقیت، قضاوت موقعیتی و تحلیل شرایط مبهم، در چنین آزمون‌هایی نادیده گرفته می‌شود. از این منظر، برتری ماشین نه محصول ضعف انسان، بلکه نتیجه طراحی آزمون‌هایی است که بیش از حد ساده‌سازی شده و در سطح بازیابی اطلاعات باقی مانده‌اند. بنابراین، یکی از دستاوردهای کلیدی این پژوهش آشکار ساختن این واقعیت است که اگر هدف آموزش صرفاً آماده‌سازی برای چنین آزمون‌هایی باشد، انسان‌ها به ناچار در رقابت با ماشین‌ها بازنده خواهند بود. یافته‌های کیفی پژوهش این برداشت را تقویت کردند.

بسیاری از دانشجویان شرکت‌کننده در آزمون‌های بازطراحی‌شده اذعان داشتند که تجربه آزمون برایشان ساده‌تر، عادلانه‌تر و شفاف‌تر بوده است. این مسئله بیانگر اهمیت تجربه ادراک شده آزمون است. اعتبار یک آزمون تنها در نتایج عددی خلاصه نمی‌شود، بلکه میزان اعتماد آزمون‌دهندگان به عدالت و وضوح فرآیند نیز بخش جدایی‌ناپذیر آن است (*Li & et. al, 2024; Kiss & et. al, 2025; Stanoyevitch, 2024*). وقتی دانشجو احساس کند که همه داوطلبان تحت شرایط برابر قرار دارند و دستورالعمل‌ها روشن و بدون ابهام ارائه شده است، حتی در صورت کسب نمره پایین‌تر، نتیجه را معتبر و قابل پذیرش می‌داند. این امر نشان می‌دهد که آزمون‌ها نه‌تنها ابزار سنجش توانایی، بلکه پدیده‌هایی اجتماعی و

روان‌شناختی هستند که مشروعیت آن‌ها از درک داوطلبان نیز سرچشمه می‌گیرد (*Luan & et. al, 2020; Lund & Wang, 2023; Lee & et. al, 2024*).

یکی دیگر از ابعاد مهم نتایج مربوط به مسئله استرس است. آزمون‌های سنتی به دلیل فشار زمانی شدید و ساختار خشک خود اغلب سطح بالایی از اضطراب ایجاد می‌کنند. پژوهش حاضر نشان داد که نسخه بازطراحی شده با افزودن بخش‌های تعاملی و موقعیتی و کاهش پیچیدگی دستورالعمل‌ها، توانسته است استرس آزمون‌دهندگان را کاهش دهد. این یافته از دو جهت حائز اهمیت است. نخست آنکه آزمونی که بیش از حد استرس‌زا باشد، دیگر توانایی واقعی فرد را نمی‌سنجد، بلکه صرفاً واکنش او به شرایط اضطراب‌آور را اندازه‌گیری می‌کند. دوم آنکه کاهش استرس به بهبود تمرکز و انگیزش منجر می‌شود و در نهایت کیفیت یادگیری را نیز ارتقا می‌دهد. بنابراین، طراحی آزمون باید به گونه‌ای باشد که فشار روانی را در سطحی سازنده نگه دارد؛ سطحی که برانگیزاننده و محرک باشد، اما مانع بروز توانایی‌ها نشود (*Zawacki-Richter & et. al, 2019; Reib, 2021*).

عدالت آموزشی یکی از دیگر محورهای اصلی این پژوهش بود. ظهور هوش مصنوعی این خطر را ایجاد کرده است که افراد با دسترسی بیشتر به فناوری بتوانند نتایج آزمون‌های سنتی را دستکاری کنند و عملکردی فراتر از سطح واقعی خود نشان دهند. چنین وضعیتی نه تنها اعتبار آزمون را مخدوش می‌کند، بلکه نابرابری‌های اجتماعی را نیز تعمیق می‌بخشد. یافته‌های این پژوهش نشان داد که نسخه بازطراحی شده آزمون‌ها توانسته است حس عدالت بیشتری در میان دانشجویان ایجاد کند. وقتی آزمون به‌صورت زنده و تعاملی برگزار می‌شود و داوطلب باید استدلال خود را توضیح دهد یا از پاسخ خود دفاع کند، امکان اتکای صرف به ابزارهای ماشینی کاهش می‌یابد. این موضوع نه تنها به تقویت عدالت آموزشی کمک می‌کند بلکه اعتماد عمومی به نتایج آزمون‌ها را نیز افزایش می‌دهد. از این منظر، بازطراحی آزمون‌ها پاسخی است به یک ضرورت اخلاقی، زیرا نظام آموزشی بدون تضمین عدالت نمی‌تواند مشروعیت خود را حفظ کند (*Nguyen, 2023; Parapadakis, 2020*).

یافته‌های مربوط به خطاها نیز تصویری جالب ارائه می‌کنند. اگر چه مدل‌های هوش مصنوعی در بسیاری از موارد خطاهای کمتری داشتند، اما نوع خطاهای آن‌ها با انسان متفاوت بود. دانشجویان انسانی بیشتر دچار خطاهایی ناشی از فشار زمان یا سوء برداشت از سؤال می‌شدند، در حالی که خطاهای مدل‌ها بیشتر به عدم انعطاف و ناتوانی در درک زمینه یا بافت موقعیتی مربوط می‌شد. این تمایز نشان می‌دهد که کیفیت خطا به اندازه کمیت آن مهم است. خطایی که ناشی از محدودیت حافظه کوتاه‌مدت یا استرس لحظه‌ای باشد با خطایی که

ریشه در ناتوانی در درک موقعیت انسانی دارد، تفاوت ماهوی دارد. آزمون‌ها باید به گونه‌ای طراحی شوند که بتوانند این تمایز را آشکار کنند و صرفاً به شمارش تعداد خطاها اکتفا نکنند. چنین رویکردی به تفکیک دقیق‌تر توانایی‌های انسانی از توانایی‌های ماشینی کمک خواهد کرد (*Dindar & et. al, 2022; Celik & et. al, 2022*).

در کنار این نقاط قوت، نتایج پژوهش به برخی محدودیت‌ها و چالش‌ها نیز اشاره داشت. آزمون‌های تعاملی و شفاهی اگرچه به سنجش عمیق‌تر توانایی‌های انسانی کمک می‌کنند، اما هزینه‌بر و زمان‌بر هستند. نیاز به مصاحبه‌گران آموزش‌دیده و فرآیندهای پیچیده‌ی نمره‌دهی، اجرای این آزمون‌ها را در مقیاس وسیع دشوار می‌سازد (*Zhou & Jiao, 2023; Zhang & et. al, 2024*). افزون بر این، خطر سوگیری انسانی در ارزیابی وجود دارد؛ چیزی که در آزمون‌های ماشینی به مراتب کمتر است. بنابراین، راهکار عملی باید ترکیبی باشد: استفاده از ظرفیت هوش مصنوعی برای تصحیح بخش‌های عینی و استاندارد آزمون، در کنار بخش‌های تعاملی و موقعیتی برای سنجش توانایی‌های انسانی. این رویکرد ترکیبی می‌تواند هم دقت و بی‌طرفی را تضمین کند و هم مهارت‌های انسانی را در مرکز توجه قرار دهد (*Wongvorachan & et. al, 2022; Williamson & et. al, 2023*).

اسکارف و همکاران (۲۰۲۴)، در تحقیقی گزارش داد که پاسخ‌های تولید شده توسط *GPT-4* در یک آزمون دانشگاهی به‌صورت «بی‌خبریِ کامل» از سوی ممتحنین انسانی پذیرفته شدند و در نهایت میانگین نمره *AI* حتی کمی بالاتر از میانگین نمرات واقعی دانشجویان بود. این یافته عملاً نشان داد که در شرایط «پرسش‌های استاندارد» و قالب‌های پاسخ‌محور، توانایی تولید متن با کیفیت و منطبق بر انتظارات ارزیابان، باعث می‌شود ماشین‌ها نه تنها پنهان بمانند بلکه نمره بالاتری کسب کنند. نتیجه اسکارف و همکاران کاملاً با نتایج پژوهش هم راستا است. زیرا آن ها نشان دادند آزمون‌های سنتی مستعد «اشتباه تعیین‌کننده‌اند». یعنی اشتباه در تفسیر داده نمره به‌عنوان معیار توانایی انسان و بنابراین ضرورت بازطراحی آزمون‌ها (به‌ویژه افزودن اجزای تعاملی یا فرآیندی که تقلید ماشینی را دشوار می‌کند) تأیید می‌شود (*Scarfe & et. al, 2024*). تیت (۲۰۲۴) نیز در یک تحلیل تجربی درباره ارزشیابی کلی با *LLM*ها نتیجه گرفت که امتیازهای ماشینی از نظر آماری با نمرات انسانی اختلاف معنی‌داری ندارند و در سناریوهای کم‌اصطکاک می‌تواند رضایت‌بخش باشد؛ اما وی هشدار داد که «قابلیت توجیه‌پذیری» و بازده بازخورد انسانی هنوز یک مزیت برای معلمان انسانی است. مقایسه این نتایج با پژوهش نشان داد که باوجودی که *AES* یا نمره‌گذاری ماشینی قابل‌اعتماد است، دو ریسک مهم وجود دارد: (الف) آزمون‌های صرفاً

عددی یا خلاصه‌محور دیگر توان تمایز بین دانش حقیقی و تولید متن مبتنی بر الگوریتم را ندارند؛ (ب) پذیرش اجتماعی و احساس عدالت که پژوهش حاضر آن را نشان داد نیازمند حضور و تعامل انسانی در فرایند است. بنابراین یافته تیت از دید عملی نشان داد که می‌توان از *AI* برای افزایش کارایی استفاده کرد اما نباید از نیاز به بخش‌های انسانیِ توضیح‌دهنده و تعامل‌محور غفلت کرد (*Tate, 2024*). سید (۲۰۲۴)، در پژوهشی دریافت که ورود ابزارهای هوش مصنوعی به آموزش زبان پیامدهای دوگانه‌ای دارد: از یک سو می‌تواند کمک فراهم‌آور و خودمختاری یادگیرندگان را افزایش دهد و از سوی دیگر می‌تواند مهارت گفتاری واقعی و اعتمادبه‌نفس را دچار آسیب کند مگر اینکه طراحی آزمون‌ها بازنگری شود. نتیجه تحقیق وی با نتایج تحقیق حاضر همپوشانی دارد؛ در هر دو تحقیق، آزمون‌های شفاهیِ بازطراحی‌شده (که در پژوهش حاضر بهبود تجربه، عدالت و کاهش استرس را نشان داد) می‌تواند «مرزهایی» ایجاد کند که تقلید صرف ماشینی را بی‌اثر میسازد و در عین حال جنبه‌های روانی و آموزشی را تقویت می‌کند. وی هشدار داد که صرف دیجیتالی‌سازی یا افزودن فناوری بدون بازاندیشیِ فلسفی و طراحی موقعیتی کافی نیست؛ این نکته‌ای است که در پژوهش حاضر نیز تصریح شده است (*Sayed, 2024*).

فراتر از نتایج تجربی، پرسش اصلی این است که آزمون در نهایت چه چیزی را باید بسنجد؟ اگر آزمون صرفاً وسیله‌ای برای رتبه‌بندی و پیش‌بینی عملکرد تحصیلی باشد، شاید عملکرد ماشین‌ها کفایت کند. اما اگر هدف آموزش پرورش انسان‌هایی خلاق، منتقد و توانمند در حل مسائل واقعی باشد، آزمون‌ها باید مهارت‌هایی را در مرکز توجه قرار دهند که ماشین‌ها قادر به تقلید کامل آن‌ها نیستند. بازطراحی آزمون‌ها فرصتی است برای بازاندیشی در فلسفه آموزش و حرکت از سنجش دانش ایستا به سنجش مهارت‌های پویا. این تحول نه تنها نظام ارزیابی، بلکه جهت‌گیری کل نظام آموزشی را تحت تأثیر قرار خواهد داد. در یک جمع‌بندی کلی باید گفت که برتری هوش مصنوعی در آزمون‌های سنتی محصول ضعف طراحی این آزمون‌هاست، نه ضعف انسان. آزمون‌های بازطراحی‌شده می‌توانند ضمن مقابله با تهدیدهای فناورانه، تجربه‌ای عادلانه‌تر و معتبرتر برای آزمون‌دهندگان فراهم کنند. این تحول ضرورتی علمی، اخلاقی و اجتماعی است. برای تحقق این تحول، نظام‌های آموزشی باید به‌طور هم‌زمان سه هدف را دنبال کنند: حفظ دقت و پایایی آماری، ارتقای تجربه انسانی و تضمین عدالت آموزشی. تنها در این صورت است که می‌توان نظام سنجش را در برابر تغییرات فناورانه مقاوم ساخت و در عین حال به پرورش جامع مهارت‌های انسانی کمک کرد.

نتایج این پژوهش آشکار ساخت که آزمون‌های سنتی در عصر هوش مصنوعی دیگر قادر به سنجش تمایز واقعی میان توانایی‌های انسانی و ماشینی نیستند. برتری مدل‌های زبانی در دقت، سرعت و یکنواختی پاسخ‌ها بیش از هر چیز ضعف آزمون‌های حافظه‌محور و بازتولیدی را نشان داد. یافته‌های کیفی نیز نشان دادند که تجربه انسانی از آزمون صرفاً در دقت نمرات خلاصه نمی‌شود، بلکه عدالت، شفافیت و سطح استرس نیز در مشروعیت و اعتبار آزمون نقش تعیین‌کننده دارند. بر این اساس، بازطراحی آزمون‌ها نه یک انتخاب اختیاری، بلکه ضرورتی آموزشی و اخلاقی است تا بتواند مرز میان توانایی‌های انسانی و ماشینی را حفظ کند. به طور کلی، می‌توان نتیجه گرفت که نظام‌های آموزشی در مواجهه با فناوری‌های نوین باید آزمون‌ها را به سوی مدل‌هایی سوق دهند که استدلال، خلاقیت، تعامل موقعیتی و تفکر انتقادی را در مرکز توجه قرار می‌دهند. آزمون‌های بازطراحی‌شده می‌توانند ضمن بهره‌گیری از ظرفیت‌های هوش مصنوعی برای ارتقای پایایی و مقابله با تقلب، شرایطی عادلانه‌تر و انسانی‌تر برای ارزیابی فراهم کنند. این تحول به‌مثابه یک فرصت استراتژیک عمل می‌کند که نه تنها اعتبار نظام سنجش را در برابر تهدیدهای فناورانه حفظ می‌کند، بلکه اهداف والای آموزش یعنی پرورش توانایی‌های جامع انسانی را نیز تقویت خواهد کرد.

References:

Bazgir Jahantighi, Z.; Heydari Azad, Z.; Mir, S.; Shahraki, S. (2025). "Challenges of using artificial intelligence in designing standard tests by teachers", Proceedings of the First International Conference of Leading Teachers in the Era of Transformative Research, Ahvaz. (Persian).

Celik, I.; Dindar, M.; Muukkonen, H.; Järvelä, S. (2022). "The promises and challenges of artificial intelligence for teachers: A systematic review of research", TechTrends, 66(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y.

Chang, L. (2024). "An Empirical Study on the Impact of Artificial Intelligence Assisted Vocational English Teaching on Students' Intercultural Communication Skills", Conference: 2024 5th International Conference on Information Science and Education (ICISE-IE). DOI:10.1109/ICISE-IE64355.2024.11025429.

Dindar, M., Muukkonen, H., & Järvelä, S. (2022). "The promises and challenges of artificial intelligence for teachers: a systematic review of research", Techtrends, 66(4), 616-630. https://doi.org/10.1007/s11528-022-00715-y. (Persian).

Dong, Y., Liu, Z., Wang, H., & Chen, Y. (2024). "Generalization or memorization: Data contamination and trustworthy evaluation for large language models", arXiv preprint arXiv:2402.15938. doi: 10.18653/v1/2024.findings-acl.716.

Hamedi-Nasab, S., Rahimi, Sh. (2024). "Barriers and challenges of implementing artificial intelligence in the higher education system", Quarterly Journal of Educational Planning Studies, 13(26), 57–73. doi: 10.22080/eps.2025.28149.2295. (Persian).

Hao, J.; von Davier, A.A.; Yaneva, V.; Lottridge, S.; von Davier, M.; Harris, D.J. (2024). "Transforming assessment: The impacts and implications of large language models and generative AI", Educational Measurement: Issues and Practice, 43 (2), 16–29. doi: 10.1111/emip.12602. (Persian).

Hashemi, H.; Shahsavari, M.; Mahdi, F.; Ghaderi, D. (2025). "Investigating the application of artificial intelligence in designing intelligent tests and more accurate assessment of students' abilities",

Proceedings of the First International Conference of Leading Teachers in the Era of Transformative Research, Ahvaz. (Persian).

Huawei, S.; Aryadoust, V. (2023). "A systematic review of automated writing evaluation systems", Education and Information Technologies, 28 (1), 771–795. doi: 10.1007/s10639-022-11200-7.

Jiao, H.; Yadav, C.; Li, G. (2023). "Integrating psychometric analysis and machine learning to augment data for cheating detection in large-scale assessment", Preprint from PsyArXiv . https://doi.org/10.31234/osf.io/fjz2c.

Jingshan, H. (2023). "Analysis of the Application of Artificial Intelligence in Education and Teaching", Advances in Educational Technology and Psychology, 7(2), 63-67. doi: 10.23977/ aetp.2023.070210.

Kamalov, F.; Sulieman, H.; Calonge, D.S. (2021). "Machine learning based approach to exam cheating detection", PLOS ONE, 16 (8), e0254340. doi: 10.1371/journal.pone.0254340. (Persian).

Karademir Coskun, T.; Alper, A. (2024). "Evaluating the evaluators: A comparative study of AI and teacher assessments in Higher Education", Digital Education Review, 45, 124- 140. https://doi.org/ 10.1344/der.

Keykha, A.; Mohammadi, H.; Darabi, F.; Hosseini, S.S. (2025). "Identifying the Applications of Artificial Intelligence in the Assessment of Medical Students", Strides Dev Med Educ, 22(1), e1512. doi:10.22062/sdme.2025.200833.1512. (Persian).

Khera, R.; Simon, M.A.; Ross, J. S. (2023). "Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support", JAMA, 330 (23), 2255–2257. doi: 10.1001/jama. 2023.22557.

Kitsara, I. (2022). "Artificial intelligence and the digital divide: From an innovation perspective", In Platforms and artificial intelligence: The next generation of competences (pp. 245–265). Springer.

Kiss, G.; Moreira Bastos, S. (2025). "Experiences of online and face-to-face assessment of Artificial Intelligence subject", Procedia Computer Science, 263, 385–390.

Kumar, Y.; Li, J.; Singh, R. (2023). "A testing framework for AI linguistic systems (testFAILS)", Kean University Research Publications.

Lee, U.; Jung, H.; Jeon, Y.; Sohn, Y.; Hwang, W.; Moon, J.; Kim, H. (2024). "Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education", Education and Information Technologies, 29, 11483–11515. doi: 10.1007/s10639-023-12249-8.

Li, J.; Tang, T.; Zhao, W.X.; Nie, J.-Y.; Wen, J.-R. (2024). "Pre-trained language models for text generation: A survey", ACM Computing Surveys, 56 (9), 1–39. doi: 10.1145/3649449.

Luan, H.; Géczy, P.; Lai, H.; Gobert, J.; Yang, S.; Ogata, H.; … & Tsai, C. (2020). "Challenges and future directions of big data and artificial intelligence in education", Frontiers in Psychology, 11, 580820. https://doi.org/10.3389/fpsyg.2020.580820.

Lund, B.D.; Wang, T. (2023). "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?", Library Hi Tech News, 40(3), 26–29. https://doi.org/10.1108/lhtn-01-2023-0009.

Novidi, A.; Gheysari, H. (2023). "Investigating the impact of artificial intelligence on improving organizational productivity", Dynamic Management and Business Analysis, 2(3), 1–8. doi.org/10.22034/pmba.2023.710528. (Persian).

Nguyen, T. (2023). "Exploring the efficacy of chatgpt in language teaching", Asiacall Online Journal, 14(2), 156-167. https://doi.org/10.54855/acoj.2314210.

Pack, A.; Barrett, A.; Escalante, J. (2024). "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability", Computers & Education: Artificial Intelligence, 6, 100234. https://doi.org/10.1016/j.caeai.2024.100234.

Parapadakis, D. (2020) "Can artificial intelligence help predict a learner's needs? Lessons from predicting student satisfaction", London Review of Education, 18 (2), 178–195. https://doi.org/10.14324/LRE.18.2.03.

Rahm, L. (2023). "Education, automation and AI: a genealogy of alternative futures", Learning, Media and Technology, 48 (1), 6–24. doi:10.1080/17439884.2021.1977948.

Ranger, J.; Schmidt, N.; Wolgast, A. (2023). "Detecting cheating in large-scale assessment: The transfer of detectors to new tests", Educational and Psychological Measurement, 83 (5), 1033–1058. doi: 10.1177/00131644221132723.

Reib, M. (2021). "The use of ai in education: practicalities and ethical considerations", London Review of Education, 19 (1), 5, 1–14. https://doi.org/10.14324/lre.19.1.05.

Sayed, B., Younes, Z., Alkhayyat, A., Adhamova, I., Teferi, H. (2024). "To be with artificial intelligence in oral test or not to be: A probe into the traces of success in speaking skill, psychological well-being, autonomy, and academic buoyancy", Language Testing in Asia, 14(1), 49. https://doi.org/ 10.1186/ s40468-024-00321-0. (Persian).

Sayin, A., & Gierl, M. (2024). "Using OpenAI GPT to generate reading comprehension items", Educational Measurement: Issues and Practice, 43 (1), 5–18. doi: 10.1111/emip.12590.

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). "A real-world test of artificial intelligence infiltration of a university examinations system: a "Turing Test" case study", PLOS ONE, 19(6): e0305354. https://doi.org/10.1371/journal.pone.0305354.

Selwyn, N., & Gašević, D. (2022). "Assessment in the age of artificial intelligence", Computers and Education: Artificial Intelligence, 3, 100075. https://doi.org/10.1016/j.caeai.2022.100075.

Swiecki, Z.; Khosravi, H.; Chen, J.; Martinez-Maldonado, R.; Lodge, J. M., Milligan, S.,

Stanoyevitch, A. (2024). "Online assessment in the age of artificial intelligence", Discover Education, 3(126). https://doi.org/10.1007/s44217-024-00212-9.

Tate T. P., Steis, J., Bailey, D., Steve, G., Moon, Y., Ritchie, D., Tseng, W., Warschauer, M. (2024). "Can AI provide useful holistic essay scoring?", Computers & Education: Artificial Intelligence, 7, 100255. https://doi.org/10.1016/j.caeai.2024.100255.

Wang, Sh., Wang, F., Zhu, Zh., Wang, J., Tran, T., & Du, Zh. (2024). "Artificial intelligence in education: A systematic literature review", Expert Systems with Applications, 252, 124167.

Wecks, J. O., Voshaar, J., Plate, B. J., Zimmermann, J. (2024). "Generative AI Usage and Exam Performance", SSRN Electronic Journal. DOI:10.2139/ssrn.4812513.

Williamson, B., Macgilchrist, F., & Potter, J. (2023). "Re-examining AI, automation and datafication in education", Learning, Media and Technology, 48 (1), 1–5. doi: 10.1080/17439884.2023.2167830.

Wongvorachan, T., Lai, K. W., Bulut, O., Tsai, Y.-S., & Chen, G. (2022). "Artificial intelligence: Transforming the future of feedback in education", Journal of Applied Testing Technology, 23(1), 95–116. Retrieved from. http://jattjournal.net/ index.php/ atp/article/view/170387.

Yu, L. & Yu, Z. (2023). "Qualitative and quantitative analyses of artificial intelligence ethics in education using vosviewer and citnetexplorer", Frontiers in Psychology, 14, 1061778. https://doi.org/10.3389/fpsyg.2023.1061778.

Zaeri, F., Namvar, K., Zarei, M. A., & Ahmadi, F. (2025). "Examining the use of artificial intelligence in designing adaptive learning assessments", Proceedings of the First International Conference on New Horizons in Educational Psychology and Teaching-Learning from the Teacher's Perspective, Urmia. (Persian).

Zawacki- Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). "Systematic review of research on artificial intelligence applications in higher education – where are the educators?", International Journal of Educational Technology in Higher Education, 16(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., & Nehm, R. H. (2023). "AI and formative assessment: The train has left the station", Journal of Research in Science Teaching, 60(6), 1390–1398. https://doi.org/10.1002/tea.21885.

Zhang, M., Ruan, C., & Johnson, M. S. (2024). "Explainable AI: Exploring subgroup differences in shortresponse scoring", Paper presented at the annual meeting of the National Council for Measurement in Education, Philadelphia, PA.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., … & Duan, N. (2024). "AGIEval: A human-centric benchmark for evaluating foundation models", Findings of the Association for

Computational Linguistics: NAACL 2024, 2299–2314. Doi: 10.18653/v1/2024.findings-naacl.149

Zhou, T., & Jiao, H. (2023). "Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment", Educational and Psychological Measurement, 83 (4), 831–854. doi:10.1177/00131644221117193.

Redesigning Standardized Assessments in the Era of Artificial Intelligence: A Critical Analysis of the Strengths and Weaknesses of Oral Examination Models

Anita Homayounnia[1]

Doi: 10.22055/QJSD.2026.51356.3181

Abstract:

The present study, aimed at redesigning examinations in the era of artificial intelligence, investigates and compares the performance of three advanced language models (GPT-4o, Google Gemini 1.5 Pro, Anthropic Claude 3.5 Sonnet) with human students within the framework of redesigned written and oral tests. This research employed a mixed-methods approach (quantitative and qualitative). Data were collected through standardized linguistic and cognitive assessments (IELTS, Raven's Progressive Matrices, N-Back), error analysis, and user experience questionnaires. The statistical population comprised medical students at Islamic Azad University, Tehran Medical Sciences, during the 2024–2025 academic year (N=585). Stratified random sampling was applied, and the sample size was estimated at 232 participants using Cochran's formula. Results indicated that AI models outperformed humans in most indicators (accuracy, response speed, and average scores), although the nature of their errors differed substantially from human errors. Furthermore, the redesigned oral examination improved user experience, enhanced clarity of instructions, reduced stress, and strengthened perceptions of fairness among participants. There was also a significant correlation between the components of the work experience questionnaire at the 0.01 level ($p<0.01$). Findings suggest that although artificial intelligence has surpassed humans in knowledge reproduction and rapid information processing, oral and interactive assessments still hold greater potential for evaluating reasoning, creativity, and critical thinking skills. Accordingly, redesigning examinations with a focus on interactive, context-based models not only safeguards the validity and fairness of assessment systems against AI-driven challenges but also creates a strategic opportunity for transforming the philosophy of education and evaluation.

Key Consepts: *Artificial Intelligence, Human Students, Written and Oral Examinations, Error*

[1] Department of Educational Sciences, Payam Noor University, Tehran, Iran, homayounnia@pnu.ac.ir

© 2019 by the authors. Licensee SCU, Ahvaz, Iran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-Non Commercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).