بررسی سامانه‌های امتیازدهی خودکار مقالات با تکیه بر مدل‌های زبانی بزرگ: نقش و چالش‌های استفاده از OpenAI ChatGPT در سیستم‌های AES

سیدعلی حسینی* [۱]،محمد رمضانی نیارکی[۲]،آسیه پیرمردوند چگینی[۳]، بهاره پیرمردوند چگینی[۴]

۱-دانشجوی کارشناسی ارشد تکنولوژی آموزشی Al-hosseini@atu.ac.ir

۲- دانشجوی کارشناسی ارشد تکنولوژی آموزشی mohamadramezani97@gmail.com

۳-کارشناسی علوم تربیتی asiye.pm.ch@gmail.com

۴-کارشناسی آموزش زبان انگلیسی chegini.bahare.p@gmail.com

خلاصه

در دهه‌های اخیر، گسترش فناوری‌های دیجیتال به‌ویژه هوش مصنوعی (AI) تأثیر چشمگیری بر نظام‌های آموزشی و روش‌های ارزشیابی داشته است. یکی از مهم‌ترین کاربردهای این فناوری، سیستم‌های امتیازدهی خودکار مقالات (Automated Essay Scoring - AES) است که با بهره‌گیری از الگوریتم‌های پردازش زبان طبیعی(NLP)، یادگیری ماشین (ML) و شبکه‌های عصبی عمیق، فرآیند ارزیابی نوشتار را به سطحی دقیق‌تر، سریع‌تر و استانداردتر ارتقا داده‌اند. تاریخچه AES از پروژه PEG در دهه ۱۹۶۰ آغاز شد و با توسعه مدل‌هایی چون e-rater و سیستم‌های مبتنی بر تحلیل معنایی و آماری ادامه یافت. با پیشرفت مدل‌های یادگیری عمیق مانندBERT ، RoBERTa وGPT، این حوزه دستخوش تحولی بنیادین شد.

مدل‌های زبانی بزرگ (LLM) به‌ویژهChatGPT ، توانسته‌اند جنبه‌های پیچیده‌ای از زبان طبیعی مانند انسجام، منطق استدلال و سبک نگارش را در سطحی نزدیک به انسان تحلیل کنند. این توانایی موجب شده AES نه‌تنها در ارزیابی سریع متون بلکه در ارائه بازخورد توصیفی، تحلیل تحلیلی و کل‌نگر، و حتی شبیه‌سازی بازخورد همتایان مورد استفاده قرار گیرد. با این حال، چالش‌هایی همچون ماهیت جعبه‌سیاه مدل‌های عمیق، تعصبات زبانی و فرهنگی، ناپایداری پاسخ‌ها و محدودیت در شفافیت الگوریتمی، همچنان مطرح هستند.

این مقاله با روش مرور نظام‌مند منابع علمی، به بررسی تاریخچه، مبانی نظری، رویکردهای سنتی و مدرن AES پرداخته و جایگاه مدل‌های مبتنی بر هوش مصنوعی، به‌ویژهChatGPT ، در این حوزه را تحلیل می‌کند. یافته‌ها نشان می‌دهد که AES می‌تواند ابزار مکمل ارزشمندی در آموزش نوشتاری باشد، به شرط آنکه همراه با نظارت انسانی و طراحی دقیق پرامپت‌ها به کار رود. در نهایت، پژوهش تأکید می‌کند که بهره‌گیری متعادل از هوش مصنوعی در ارزشیابی آموزشی، می‌تواند هم به بهبود کیفیت یادگیری و هم به کارآمدی نظام‌های آموزشی منجر شود.

کلیدواژه:

هوش مصنوعی، امتیازدهی خودکار مقالات(AES) ، پردازش زبان طبیعی(NLP) ، یادگیری عمیق، ChatGPT

مقدمه

در این جهان به سرعت جهش یافته ، فناوری دیجیتال، تأثیر شگرفی بر تکامل اقتصادی، اجتماعی و فرهنگی همه جوامع دارد در حالی که اشکال جدیدی از فناوری زندگی ما را فرا گرفته و جوانان ما را مجذوب خود میکند. دانشگاه‌ها و مدارس چاره ای ندارند جز اینکه جایی برای فناوری‌های دیجیتال ایجاد کنند. هوش مصنوعی همیشه موضوعی داغ برای بحث بوده است زیرا در قرن بیست و یکم جهان تقریباً در همه زمینه های زندگی توسط فناوری اداره می شود. هوش مصنوعی (AI) به هوشمندی نشان داده شده توسط ماشین‌ها در شرایط مختلف اطلاق میشود که در مقابل هوش طبیعی در انسان‌ها قرار میگیرد. (Prendes, 2021 & González)

نمونه هایی از کاربست های هوش مصنوعی در ارزشیابی آموزشی (Zhang, 2023 & ,Tingting, Su)

دستیار عملیاتی مجازی جهت ارائه بازخورد به دانش آموزان بر اساس معیارهای اجرایی.

آنها واقعیت مجازی و هوش مصنوعی را برای طبقه بندی دانش آموزان در رابطه با معیارهای عملکرد مهارت ادغام می کنند و سیستم برای کمک به بهبود آنها بازخورد می دهد.

آنها استفاده از هوش مصنوعی را برای تولید تست در محیط های آموزش الکترونیکی تجزیه و تحلیل می کنند و یک نرم افزار هوشمند برای انتخاب سوالات امتحانات آنلاین توسعه می دهند.

بازخورد ارائه شده توسط سیستم های دیجیتال در موقعیت های یادگیری دارای مشکلاتی مانند برانگیختن احساسات منفی (این‌ها عبارتند از غفلت، ناامیدی، عدم اطمینان، نیاز به تایید و ناراحتی).

امتیازدهی خودکار مقالات AES(automated essay scoring)

AES یکی از کاربرد های اساسی هوش مصنوعی که در ارزشیابی آموزشی بوده که در این قسمت می‌خواهیم به ارزشیابی خودکار توسط هوش مصنوعی بپردازیم

تعریف و مبانی نظری سیستم‌های امتیازدهی خودکار :(AES)

سیستم‌های امتیازدهی خودکار (Automated Essay Scoring - AES) ابزارهای هوشمندی هستند که با استفاده از الگوریتم‌های رایانه‌ای، توانایی دارند نوشته‌های آزاد و مقالات دانش‌آموزان را از نظر ساختار، محتوا، دستور زبان و انسجام ارزیابی کرده و نمره‌دهی کنند. این سیستم‌ها با هدف افزایش کارایی، کاهش خطای انسانی و استانداردسازی ارزیابی نگارش به‌وجود آمده‌اند .(Shermis & Burstein, 2013) به‌طور کلی، AESترکیبی از روش‌های پردازش زبان طبیعی (NLP)و مدل‌های آماری یا یادگیری ماشین برای تحلیل متن است.

تاریخچه AES به دهه ۱۹۶۰ میلادی بازمی‌گردد، زمانی که پروژه اولیه‌ای به نام Project Essay Grade (PEG) توسط Ellis Page در دانشگاه Duke توسعه یافت. این سیستم با استفاده از ویژگی‌های سطحی نوشتار مانند طول جمله و تعداد واژگان عمل می‌کرد .(Page, 1966) در دهه‌های بعد، مدل‌های آماری پیچیده‌تری مانند e-rater توسط ETS معرفی شدند که مبتنی بر تحلیل ساختار گرامری، واژگان سطح بالا و انسجام معنایی بودند .(Burstein et al., 1998) این تحولات منجر به بهبود چشمگیر در دقت و روایی نمره‌دهی نسبت به ارزیابی انسانی شدند.

با پیشرفت فناوری‌های یادگیری ماشین، AES به‌طور گسترده‌ای از مدل‌های پیشرفته‌تری مانند Random Forests، Support Vector Machinesو به‌ویژه شبکه‌های عصبی عمیق (Deep Learning) بهره‌مند شد. سیستم‌هایی همچون TextEvaluatorو Intelligent Essay Assessor اکنون توانایی تحلیل معناشناختی و سبک نگارشی را

دارند و در محیط‌های آموزشی رسمی استفاده می‌شوند .(Yoon et al., 2020) این سیستم‌ها با داده‌های بزرگ تمرین داده می‌شوند و می‌توانند نوشتار را با دقت بالا ارزیابی کنند.

یکی از ابزارهایی که در به کارگیری جهت استفاده در زمینه AES در سطح بالا ، هوش مصنوعی Chatgpt3.5 و - Chatgpt۴O است. در ۳۰ نوامبر ۲۰۲۲، ChatGPT،OpenAI را منتشر کرد، یک ربات گفتگوی همه‌منظوره که پیش‌بینی می‌شود تأثیر قابل‌توجهی بر تمام جنبه‌های جامعه داشته باشد. اثرات احتمالی این ابزار بر آموزش، هنوز نامشخص است. ظرفیت ChatGPT ممکن است بر تنظیم فعالیت‌های یادگیری، اهداف یادگیری آموزشی و روش‌های ارزیابی و ارزیابی تأثیر بگذارد که ممکن است تأثیر قابل‌توجه‌ی داشته باشد. ChatGPT، می تواند به دانشگاهیان کمک کند تا مقالاتی منظم، منسجم، (عمدتا) درست و آموزنده بنویسند. (Tomer, 2023 & ,Ratnam, Sharma)

روش تحقیق:

روش این پژوهش، کتابخانه‌ای است که از نوع مرور ادبیات است و از راه فیش برداری ،یاداشت برداری و همچنین جستجو در پایگاه ها و منابع اطلاعات علمی اینترنتی چون سیویلیکا ، نورمگز ، SID و... جمع شده است. ابتدا اطلاعات لازم گردآوری شده، سپس به تفسیر آنها پرداخته میشود.در مرحله بعد اطلاعات بدست آمده،طبقه بندی و خلاصه بندی می شوند و آنگاه تشابهات و تفاوت‌ها مورد بررسی قرار گرفته، به پرسش های پژوهش پاسخ داده می شود.

ابزاری که در این پژوهش برای گرد آوری اطلاعات مورد استفاده قرار گرفته،روش سندکاوی است.بر اساس کلیه کتاب‌ها ، مدارک ، اسناد ،مقاله ها و پژوهش های در دسترس پژوهشگران که به بررسی این مفاهیم پرداخته اند. مورد مطالعه قرار گرفته و از آنها فیش برداری شده است.

امتیازدهی خودکار مقاله (AES)

امتیازدهی خودکار مقاله (AES) از فناوری برای ارزیابی و امتیازدهی مقالات نوشته شده استفاده می‌کند. این فرآیند با برنامه‌های رایانه‌ای انجام می‌شود که مقالات را بر اساس معیارهایی مانند صحت زبانی، غنای واژگانی، انسجام نحو و ارتباط معنایی تجزیه و تحلیل می‌کنند. AES، با وجود مزایا و معایب، به دلیل جلوگیری از خطاهای انسانی و صرفه‌جویی در زمان، توصیه می‌شود. اولین سیستم‌های AES در دهه ۱۹۶۰ با پروژه نمره مقاله (PEG) شکل گرفتند که از تحلیل رگرسیون برای پیش‌بینی نمرات استفاده می‌کردند، اما به دلیل تمرکز بر ساختارهای سطحی و نادیده گرفتن محتوای مقاله، مورد انتقاد قرار گرفتند و در برابر تقلب آسیب‌پذیر بودند. (Eguchi, 2023 & Mizumoto)

با رشد فزاینده داده‌های متنی و نیاز به ارزیابی سریع و دقیق متون، سامانه‌های امتیازدهی خودکار مقاله (Automated Essay Scoring – AES) به یکی از زمینه‌های کلیدی در کاربرد هوش مصنوعی تبدیل شده‌اند. ریشه‌های توسعه این سامانه‌ها در به‌کارگیری روش‌های یادگیری ماشین (ML) و پردازش زبان طبیعی (NLP) نهفته است که به مدل‌ها اجازه می‌دهند با آموزش بر روی حجم بالایی از متون ارزیابی‌شده توسط انسان، به سطحی از دقت قابل قبول در پیش‌بینی نمره مقالات جدید برسند (Shermis & Burstein, 2013). در این فرآیند، انتخاب ویژگی‌های زبانی مانند طول جمله، تنوع واژگان و انسجام معنایی نقش تعیین‌کننده‌ای دارد، زیرا انتخاب مناسب ویژگی‌ها به کاهش نویز و افزایش دقت کمک می‌کند. هم‌زمان، NLPبه رایانه‌ها امکان می‌دهد تا زبان انسان را تحلیل و تفسیر کنند، که این امر برای کاربردهایی چون ترجمه ماشینی، تحلیل احساسات، پرسش و پاسخ خودکار و تولید محتوا حیاتی است.(Mizumoto & Eguchi, 2023)

الگوریتم‌های سنتی AES (مانندLSA ، SVM، E-Rater)

در مراحل اولیه، سامانه‌هایی نظیر E-Rater که توسط ETS توسعه داده شد، از مدل‌های آماری نظیر تحلیل معنایی نهفته (Latent Semantic Analysis - LSA) و ماشین بردار پشتیبان (Support Vector Machine - SVM) بهره می‌بردند. این مدل‌ها با تمرکز بر استخراج ویژگی‌های سطحی متن، مانند طول جملات و ساختار نحوی، به ارزیابی می‌پرداختند؛ اما محدودیت عمده آن‌ها در ناتوانی در درک عمیق معنا و زمینه متنی بود (Shermis & Burstein, 2013; Jaleel et al., 2024).

مدل‌های یادگیری عمیق مدرن (BERT، RoBERTa، GPT)

ظهور مدل‌های یادگیری عمیق موجب تحولی بنیادین در AES شد. یادگیری عمیق به عنوان زیرشاخه‌ای از یادگیری ماشین، از شبکه‌های عصبی چندلایه برای حل مسائل پیچیده مانند پردازش زبان طبیعی استفاده می‌کند. مدل‌هایی مانند BERT و RoBERTa با بهره‌گیری از معماری‌های attention-based قادر به استخراج روابط معنایی پیچیده بین واژگان هستند (Chatrola, 2024; Pilicita & Barra, 2025). همچنین مدل GPT و دیگر مدل‌های زبانی تولیدی، با استفاده از رویکرد autoregressive، در درک و تولید زبان انسانی عملکرد چشم‌گیری نشان داده‌اند (Mizumoto & Eguchi, 2023).

یادگیری عمیق این قابلیت را دارد که ویژگی‌های زبانی را به‌طور خودکار از داده‌های خام استخراج کند، در حالی که روش‌های سنتی به ویژگی‌هایی نیاز دارند که به صورت دستی طراحی شوند. این امر، دقت و توانایی تعمیم سیستم را به شکل چشم‌گیری افزایش داده و راه را برای استفاده گسترده‌تر از AES در زبان‌ها و کاربردهای مختلف باز کرده است.

تفاوت میان LLM-based و Feature-based AES

در جدول زیر، تفاوت‌های کلیدی بین دو رویکرد اصلی در توسعه سیستم‌های AES ، یعنی رویکرد مبتنی بر ویژگی‌ها (Feature-based) و رویکرد مبتنی بر مدل‌های زبانی بزرگ (LLM-based) آورده شده است:

ویژگی‌ها	Feature-based AES	LLM-based AES (مانند GPT/BERT)
رویکرد	استخراج ویژگی دستی	یادگیری از داده بدون ویژگی دستی
تفسیرپذیری	نسبتاً بالا	اغلب یک جعبه سیاه (black box)
نیاز به داده	نیازمند داده‌های ساختاریافته	نیازمند داده‌های عظیم برای آموزش
درک معنا	محدود به ویژگی‌های استخراجی	درک عمیق از معنا و زمینه
زبان	وابسته به زبان	اغلب چندزبانه یا قابل تنظیم‌پذیر
توسعه‌پذیری در سایر وظایف NLP	محدود	بالا (قابل استفاده در سایر وظایف NLP)

همان‌طور که مشاهده می‌شود، مدل‌های مبتنی بر LLM با بهره‌گیری از داده‌های گسترده و ساختارهای پیشرفته یادگیری عمیق، دقت، انعطاف‌پذیری، و قابلیت توسعه بالاتری نسبت به روش‌های سنتی دارند (Jaleel et al., 2024; Beseiso et al., 2021).

یادگیری ماشین و یادگیری عمیق

زمینه تحقیقات امتیازدهی خودکار مقاله (AES) از توسعه روش‌های یادگیری ماشین (ML) بهره برده است. یادگیری ماشین، به عنوان زیرشاخه‌ای از هوش مصنوعی، مدل‌های AES را با آموزش بر روی مجموعه بزرگی از مقالات ارزیابی‌شده توسط متخصصان انسانی می‌سازد و اعتبارسنجی می‌کند. این مدل‌ها الگوهایی را از داده‌های آموزشی استخراج کرده و امتیاز مقالات جدید را پیش‌بینی می‌کنند، به طوری که پیش‌بینی‌های آنها با امتیازهای انسانی همسو باشد. انتخاب ویژگی‌های حیاتی متن در این فرآیند اهمیت زیادی دارد، زیرا انتخاب مناسب باعث بهبود نتایج و کاهش نویز می‌شود. همچنین، یادگیری ماشین همراه با ابزارهای پردازش زبان طبیعی (NLP) امکان استخراج خودکار ویژگی‌های زبانی مانند انسجام متن را فراهم می‌کند.

پردازش زبان طبیعی (NLP) فناوری یادگیری ماشین است که به رایانه‌ها امکان تفسیر، دستکاری و درک زبان انسان را می‌دهد. این فناوری به سازمان‌ها کمک می‌کند تا حجم زیادی از داده‌های صوتی و متنی از منابع مختلف مانند ایمیل، پیام‌های متنی، شبکه‌های اجتماعی، ویدئو و صدا را پردازش کنند. NLP به تعامل بین زبان انسان و رایانه می‌پردازد و رایانه‌ها را قادر می‌سازد تا متن و گفتار را مانند مغز انسان درک کنند. کاربردهای NLP شامل مدیریت محتوا، ترجمه ماشینی، تحلیل احساسات، پرسش و پاسخ خودکار و تشخیص گفتار است.

در سال‌های اخیر، سیستم‌های امتیازدهی خودکار مقاله (AES) با ترکیب یادگیری عمیق تکامل یافته‌اند. یادگیری عمیق، زیرشاخه‌ای از یادگیری ماشین، از شبکه‌های عصبی چندلایه برای حل مشکلات پیچیده مانند تشخیص تصویر و گفتار، پردازش زبان طبیعی (NLP) و تصمیم‌گیری استفاده می‌کند. تفاوت اصلی یادگیری عمیق با سایر تکنیک‌های یادگیری ماشین در استفاده از چندین لایه برای یادگیری و تصمیم‌گیری است، که دقت و توانایی تعمیم بیشتری را به همراه دارد. این پیشرفت‌ها در ابزارهایی مانند DeepL و Google Translate مشاهده می‌شود. یادگیری عمیق ویژگی‌ها را به‌طور خودکار از داده‌های خام استخراج می‌کند، در حالی که سایر تکنیک‌های یادگیری ماشین به ویژگی‌های ساخته‌شده توسط متخصصان متکی هستند.(Mizumoto & Eguchi, 2023)

نقش OpenAI ChatGPT در AES

تحقیقات جدید در سال‌های اخیر، به‌ویژه در مورد GPT-2 و GPT-3 ، نشان داده‌اند که این مدل‌ها می‌توانند متون را با دقتی نزدیک به ارزیابی انسانی نمره‌گذاری کنند .(Ludwig et al., 2021) با این حال، از آنجایی که GPT در اصل برای تولید متن طراحی شده، استفاده مؤثر از آن در امتیازدهی نیازمند تنظیم دقیق (fine-tuning) با داده‌های خاص وظیفه AES است.(Ludwig et al., 2021)

مدل زبانی ChatGPT ، توسعه‌یافته توسط OpenAI ، به‌عنوان یکی از پیشرفته‌ترین نمونه‌های یادگیری عمیق در حوزه پردازش زبان طبیعی(NLP) ، توانسته است تحولی بنیادین در سامانه‌های امتیازدهی خودکار متون (AES) ایجاد کند. این مدل با تکیه بر معماری ترانسفورمر و حجم عظیمی از داده‌های متنی، توانایی بالایی در تولید پاسخ‌هایی شبیه انسان دارد. پژوهش‌ها نشان می‌دهد که ChatGPT قادر است جنبه‌های مهم نوشتار مانند انسجام متن، سازمان‌دهی ایده‌ها، و منطق استدلال را به‌دقت تحلیل کند و از این نظر فراتر از سیستم‌های سنتی مبتنی بر قواعد و معیارهای ایستا عمل کند. برای مثال، مطالعه‌ای توسط Mizumoto و Eguchi (2023) نشان داد که ChatGPT توانسته است عملکردی نزدیک به ارزیابی انسانی در تحلیل متون تحلیلی و تفسیری ارائه دهد، که نشان از پتانسیل قابل توجه آن در محیط‌های آموزشی و پژوهشی دارد .(Mizumoto & Eguchi, 2023)

یکی از برجسته‌ترین ویژگی‌های ChatGPT ، توانایی آن در درک و تحلیل دقیق زبان طبیعی است. این مدل از طریق یادگیری عمیق و پیش‌بینی توزیع واژگان، قادر است الگوهای معنایی پیچیده و روابط نحوی را شناسایی و تحلیل کند. در آزمون‌های میدانی که بر روی متون تولیدشده توسط زبان‌آموزان انجام شده، مشخص شده است که ChatGPT می‌تواند

به‌صورت دقیق مؤلفه‌هایی چون صحت دستوری، سبک نگارشی، و انسجام مفهومی را ارزیابی کند. نتایج این بررسی‌ها نشان می‌دهد که مدل در بسیاری از موارد عملکردی هم‌سطح یا حتی بالاتر از ارزیاب‌های انسانی داشته است، که این موضوع قابلیت آن را برای استفاده در محیط‌های آموزشی چندزبانه تقویت می‌کند .(Song et al., 2024)

قابلیت دسترسی به ChatGPT از طریق API رسمیOpenAI ، این امکان را برای پژوهشگران و توسعه‌دهندگان فراهم کرده است تا از این مدل در قالب سامانه‌های آموزشی تعاملی و خودکار استفاده کنند. این دسترسی به صورت بلادرنگ و مقیاس‌پذیر، امتیازدهی متون نوشتاری را به بخشی از فرایندهای آموزشی دیجیتال تبدیل کرده است. در مطالعه‌ای که توسط Liew و Tan (2024) انجام شده است، مشخص شد که استفاده از API مدل‌های بزرگ زبانی می‌تواند با بهره‌گیری از ساختار Rubric های آموزشی، فرایند ارزیابی را با دقت بالا و در زمان کمتر اجرا کند. این موضوع به‌ویژه در محیط‌های یادگیری آنلاین که نیاز به بازخورد فوری وجود دارد، از اهمیت ویژه‌ای برخوردار است(Liew & Tan, 2024).

ساختارهای prompt مهندسی‌شده برایAES

یکی از عوامل کلیدی در بهبود عملکرد ChatGPT در سامانه‌هایAES ، طراحی مؤثر و دقیقprompt هاست. مهندسی prompt شامل ایجاد دستورالعمل‌های شفاف، تعیین قالب‌های ارزیابی، و تعریف انتظارات صریح از مدل است. پژوهش Mansour و همکاران (۲۰۲۴) نشان می‌دهد که ساختارهای استانداردشده‌یprompt ، به‌ویژه آن‌هایی که بر اساسRubric های انسانی طراحی شده‌اند، می‌توانند باعث افزایش چشمگیر در دقت امتیازدهی شوند. این یافته اهمیت طراحی هدفمند ورودی به مدل‌های زبانی را برای بهره‌گیری بهینه از قابلیت‌های آن‌ها در ارزیابی نوشتار نشان می‌دهد (Mansour et al., 2024).

دسته بندی انواع پرامپت AES

در ادامه برای هر بخش از سیستم‌های امتیازدهی خودکار مقاله(AES) ، یک جدول اختصاصی ارائه می‌شود که شامل تعریف هر معیار، توضیح نحوه استفاده درChatGPT ، پرامپت نمونه برای استفاده در ارزیابی خودکار است.

امتیازدهی تحلیلی(Analytical Scoring)

امتیازدهی تحلیلی یک رویکرد جزءبه‌جزء برای ارزیابی مقالات است که هر بخش کلیدی مانند پایان‌نامه، مقدمه، نتیجه‌گیری، شواهد، سازمان متن، انتقال ایده‌ها، زبان و گرامر به‌طور مستقل سنجیده می‌شود. این روش که در پژوهش‌های Yamashita (2024)، Uyarو Büyükahıska (2025) و دیگران مطرح شده، با تجزیه ساختاری مقاله به مؤلفه‌های مشخص، امکان بازخورد دقیق‌تر و هدفمندتر را فراهم می‌کند. چنین ارزیابی‌ای در محیط‌هایی مانند ChatGPT به کمک پرامپت‌های دقیق، می‌تواند بازده بالایی در تشخیص نقاط ضعف و قوت هر بخش از متن داشته باشد.

معیار	توضیح عملکرد	ChatGPTپرامپت پیشنهادی برای	منابع
بیانیه پایان‌نامه	وضوح هدف مقاله و قدرت آن در هدایت استدلال‌ها	"پیگیری آیا این مقاله دارای یک پایان‌نامه روشن و قابل .".است؟ لطفاً آن را ارزیابی و نمره‌دهی کنید	(Yamashita, 2024)
مقدمه	درگیرکننده بودن و شفافیت دید کلی	"کرده و مقدمه مقاله را بررسی کن. آیا خواننده را جذب دید روشنی از موضوع می‌دهد؟"	(Uyar & Büyükahıska, 2025)
نتیجه‌گیری	جمع‌بندی مؤثر و حس بسته بودن	"نکات نتیجه‌گیری مقاله را بررسی کن. آیا مؤثر است و کلیدی را جمع‌بندی می‌کند؟"	(Obata et al., 2023)

شواهد و پشتیبانی	استفاده از منابع معتبر و یکپارچه	استدلال شواهد پشتیبان مقاله را از نظر اعتبار و ربط به بررسی کن."	(Shin & Lee, 2024)
سازمان و ساختار	نظم منطقی ایده‌ها	ارائه آیا ایده‌ها در مقاله به‌صورت منظم و قابل پیگیری شده‌اند؟ ساختار را ارزیابی کن."	(Mansour et al., 2024)
انتقال	روانی انتقال بین پاراگراف‌ها	را کیفیت استفاده از عبارات انتقالی و پیوستگی متن بررسی کن."	(Chen et al., 2025)
زبان و سبک	تناسب لحن و شفافیت	ایجاز زبان استفاده شده را از نظر مناسب بودن، وضوح و بررسی کن."	(Tate et al., 2024)
گرامر و مکانیک	خطاهای زبانی	آنها را آیا مقاله دارای خطاهای گرامری یا نگارشی است؟ مشخص و امتیازدهی کن."	(Kundu & Barbosa, 2024)

امتیازدهی کل‌نگر(Holistic Scoring)

در امتیازدهی کل‌نگر، مقاله به‌عنوان یک کل واحد ارزیابی می‌شود و معیارهایی مانند تأثیر کلی، تسلط بر محتوا، سطح تفکر انتقادی و میزان تعامل با خواننده سنجیده می‌شوند. این روش که در کارهای Tate et al. (2024) و Mansour et al. (2024)نیز توصیه شده، بیشتر بر برداشت کلی ارزیاب از کیفیت و اثرگذاری متن متمرکز است تا بررسی جزءبه‌جزء آن. در استفاده باChatGPT ، چنین رویکردی می‌تواند برای سنجش سریع کیفیت کلی متن پیش از ورود به جزئیات به‌کار رود.

معیار	توضیح عملکرد	پرامپت پیشنهادی	منابع
تأثیر کلی	تأثیر کلی مقاله بر خواننده	بده و بر اساس خواندن کامل مقاله، نمره‌ای از ۰ تا ۱۰ توضیح بده چرا."	(Tate et al., 2024)
تسلط بر محتوا	درک نویسنده از موضوع	ارزیابی آیا نویسنده تسلط کاملی بر موضوع دارد؟ لطفاً و نمره بده."	(Bui & Barrot, 2024)
تفکر انتقادی	تحلیل عمیق و استدلال	سطح تفکر انتقادی و استدلال در مقاله را بررسی کن."	(Mansour et al., 2024)
تعامل	جذابیت و ارتباط با خواننده	می‌دارد؟ خواننده را درگیر نگه آیا مقاله جذاب است و"	(Uyar & Büyükahıska, 2025)

امتیازدهی مبتنی بر روبریک(Rubric-Based Scoring)

این شیوه با استفاده از چارچوب‌های ازپیش‌تعریف‌شده (روبریک) انجام می‌شود که برای هر معیار، سطوح مختلف عملکرد و امتیاز مشخص شده است. طبق پژوهش‌های Xia et al. (2024) وYamashita (2024) ، معیارهایی چون محتوا، سازمان، ادبیات و قواعد زبانی، هر یک با مقیاس عددی (مثلاً ۱ تا ۴) سنجیده می‌شوند. استفاده از روبریک باعث

می‌شود ارزیابی‌ها استانداردتر، قابل‌مقایسه و کمتر وابسته به نظر شخصی ارزیاب باشد و ChatGPT می‌تواند با پرامپت‌های عددی و شفاف، این نوع ارزیابی را شبیه‌سازی کند.

عنصر	شرح معیار	پرامپت نمونه برای ChatGPT	منابع
محتوا	پوشش نکات کلیدی	"بر اساس روبریک، محتوا را از ۱ تا ۴ امتیاز بده."	(Xia et al., 2024)
سازمان	ساختار منطقی و توالی	"ساختار مقاله را امتیازدهی کن طبق معیار روبریک."	(Yamashita, 2024)
ادبیات	واژگان و تنوع زبانی	"کیفیت واژگان و جملات را از ۱ تا ۴ نمره بده."	(Chen et al., 2025)
قواعد زبانی	گرامر و قالب‌بندی	"از نظر قواعد زبانی، مقاله را امتیازدهی کن."	(Shin & Lee, 2024)

امتیازدهی بر اساس معیار(Trait-Based Scoring)

در این مدل، به جای تمرکز بر کل یا اجزای ریز، ویژگی‌های کیفی متن مانند تمرکز و هدف، آگاهی از مخاطب، توسعه ایده‌ها و انسجام کلی ارزیابی می‌شود Khademi (2023) . و Li et al. (2024) نشان داده‌اند که این روش به‌ویژه در محیط‌های آموزشی برای تشویق نویسندگان به بهبود مهارت‌های خاص نگارشی مؤثر است ChatGPT .می‌تواند از این معیارها برای تحلیل کیفی متن و ارائه بازخورد هدفمند استفاده کند.

معیار	شرح عملکرد	پرامپت نمونه	منابع
تمرکز و هدف	انسجام و تداوم موضوع	"آیا مقاله روی موضوع اصلی متمرکز مانده؟"	(Khademi, 2023)
آگاهی مخاطب	تناسب لحن با مخاطب	"آیا مقاله متناسب با مخاطب خاص نوشته شده است؟"	(Tate et al., 2024)
توسعه ایده‌ها	عمق و دقت استدلال	"ایده‌ها چگونه توسعه یافته‌اند؟ لطفاً تحلیل کن."	(Li et al., 2024)
انسجام	پیوستگی ایده‌ها	"آیا بخش‌های مقاله به‌خوبی به هم متصل‌اند؟"	(Obata et al., 2023)

بازخورد توصیفی(Descriptive Feedback)

بازخورد توصیفی بر شناسایی نقاط قوت، زمینه‌های نیازمند بهبود و پرسش‌های برانگیخته‌شده توسط متن تمرکز دارد. بر اساس Chen et al. (2025) و Xia et al. (2024) ، این نوع بازخورد نه تنها به نویسنده می‌گوید کجا اشتباه کرده،

بلکه دلیل آن را نیز توضیح می‌دهد و پیشنهاد مشخص برای بهبود ارائه می‌کند. این شیوه برای استفاده در ChatGPT بسیار مناسب است زیرا می‌توان پرامپت‌هایی طراحی کرد که بازخورد غنی و توضیحی تولید کنند.

عنصر	شرح	پرامپت پیشنهادی	منابع
نقاط قوت	جنبه‌های موفق مقاله	"چه نقاط قوتی در این مقاله وجود دارد؟"	(Li et al., 2024)
زمینه‌های بهبود	نقاط نیازمند اصلاح	"چه بخش‌هایی از مقاله نیاز به بهبود دارند؟"	(Chen et al., 2025)
سوالات مطرح‌شده	ابهامات و گره‌ها	آیا نکات مبهم یا سوال‌برانگیزی در مقاله " هست؟"	(Xia et al., 2024)

درخواست‌های بررسی همتایان(Peer Review Prompts)

در این دسته، بازخورد توسط همتایان (خوانندگان یا نویسندگان دیگر) ارائه می‌شود و شامل ارزیابی وضوح، جذابیت، ارائه انتقاد سازنده و بیان نقاط مثبت است. بر اساس یافته‌های Bui و Barrot (2024) و Mansour et al. (2024) ، بررسی همتایان نه‌تنها کیفیت متن را ارتقا می‌دهد بلکه مهارت‌های انتقادی و تحلیلی ارزیاب را نیز تقویت می‌کند. در ChatGPT، این رویکرد می‌تواند با شبیه‌سازی نقش یک همتا اجرا شود تا دیدگاه‌های متنوع‌تری به نویسنده ارائه گردد.

معیار	توضیح عملکرد	پرامپت نمونه	منابع
وضوح	شفافیت کلی متن	آیا مقاله قابل فهم و واضح است؟ چه بخش‌هایی مبهم" هستند؟"	(G., 2023)
علاقه‌مندی	جذابیت برای خواننده	"آیا مقاله توجه تو را حفظ کرد؟ چرا؟"	(Bui & Barrot, 2024)
انتقاد سازنده	پیشنهادات بهبود	"چه نکاتی می‌توانند در مقاله بهتر شوند؟"	(Mansour et al., 2024)
بازخورد مثبت	نکات موفق مقاله	چه چیزی در مقاله بیش از همه نظرت را جلب " کرد؟"	(Tate et al., 2024)

مزایا و محدودیت‌ها

استفاده از ChatGPT در ارزیابی متون نوشتاری مزایای متعددی دارد. این مدل قادر است سبک‌های نوشتاری مختلف را درک کرده، با زبان‌های گوناگون سازگار شود، و در تحلیل زمینه‌ای متون عملکرد خوبی داشته باشد. با این حال، محدودیت‌هایی نیز وجود دارد. یکی از چالش‌های اصلی، طبیعت «جعبه‌سیاه» مدل است که موجب می‌شود فرآیند تصمیم‌گیری آن برای کاربران شفاف نباشد. علاوه‌براین، ظرفیت حافظه محدود و ناپایداری در پاسخ‌های تکراری ممکن است منجر به عدم ثبات در ارزیابی شود Yavuz .و همکاران (۲۰۲۵) نیز به این نکته اشاره کرده‌اند که در مواردی که consistency اهمیت دارد، ChatGPT ممکن است پاسخ‌هایی با تنوع بالا ولی دقت پایین ارائه دهد (Yavuz et al., 2025).

در مطالعات مقایسه‌ای بین ChatGPT و سایر سامانه‌های امتیازدهی خودکار مانند OpenEssayist و Write & Improve، مشخص شده است که هر سیستم مزایا و نقاط ضعف خاص خود را دارد. برای نمونه، در مطالعه‌ای توسط Bui

وChatGPT در ارزیابی برخی از متون از لحاظ دقت نمره‌دهی) بر اساس همبستگی (Pearson Barrot (2024) ، عملکرد بهتری داشت. با این حال، در حوزه ثبات و تکرارپذیری، سامانه‌های سنتی مانند Write & Improve در مواردی دقیق‌تر ظاهر شدند. این یافته‌ها نشان می‌دهد که گرچه ChatGPT ابزار قدرتمندی است، اما برای استفاده گسترده در محیط‌های رسمی آموزشی باید با احتیاط و همراه با نظارت انسانی به‌کار گرفته شود.[(Bui & Barrot, 2024)]

یکی از چالش‌های مهم در استفاده از مدل‌های زبانی بزرگ مانند ChatGPT برای امتیازدهی خودکار، تعصبات زبانی و فرهنگی موجود در داده‌های آموزشی است که می‌تواند منجر به ارزیابی‌های ناعادلانه شود. همچنین، بسته‌بودن کد منبع و عدم شفافیت الگوریتم‌های داخلی مدل، اعتمادپذیری آن را کاهش می‌دهد. مطالعه‌ای از Pack و همکاران (۲۰۲۴) به این نکته اشاره دارد که استفاده مکرر ازprompt های خاص ممکن است منجر به overfitting شده و در نتیجه، توانایی تعمیم مدل به انواع جدید متون کاهش یابد. این چالش‌ها ضرورت نظارت انسانی و تحلیل انتقادی در استفاده از چنین ابزارهایی را برجسته می‌سازد .(Pack et al., 2024)

خلاصه و جمع‌بندی

سیستم‌های امتیازدهی خودکار مقالات (AES) از مهم‌ترین کاربردهای هوش مصنوعی در آموزش محسوب می‌شوند که با تکیه بر NLP و الگوریتم‌های یادگیری ماشین، امکان ارزیابی سریع، دقیق و استاندارد متون نوشتاری را فراهم کرده‌اند. بررسی تاریخی و تحولات اخیر نشان می‌دهد که این فناوری از تحلیل ویژگی‌های سطحی متن آغاز شده و اکنون با ظهور مدل‌های یادگیری عمیق وLLM ها، قادر به درک معنا و ساختار متون در سطحی نزدیک به انسان است ChatGPT . به‌عنوان یکی از برجسته‌ترین نمونه‌ها، توانسته است قابلیت‌هایی فراتر از سیستم‌های سنتی ارائه دهد؛ از جمله تحلیل انسجام متن، سازمان‌دهی ایده‌ها و بازخورد توصیفی.

با وجود این مزایا، محدودیت‌هایی همچون تعصبات داده، عدم شفافیت الگوریتم‌ها و ناپایداری در خروجی‌ها ضرورت بهره‌گیری محتاطانه از این ابزار را نشان می‌دهد. نتیجه این پژوهش حاکی از آن است کهAES ، به‌ویژه در ترکیب با مدل‌های زبانی بزرگ و همراهی نظارت انسانی، می‌تواند نقش مهمی در ارتقای یادگیری و بهبود کیفیت آموزش نوشتاری ایفا کند.

منابع

Bui, M. N., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. Education and Information Technologies, 30, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Chen, X., Zhou, Z., & Prado, M. (2025). ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams. International Journal of Assessment Tools in Education. https://consensus.app/papers/chatgpt35-as-an-automatic-scoring-system-and-feedback-chen-zhou/e64dbc9aa34a5f70bdb68577be80e29b

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. arXiv. https://consensus.app/papers/can-chatgpt-and-bard-generate-aligned-assessment-items-a-khademi/a1ce2ba721165d83bdbd43b318116fcd

Kundu, A., & Barbosa, D. (2024). Are large language models good essay graders? arXiv. https://consensus.app/papers/are-large-language-models-good-essay-graders-kundu-barbosa/0c4ef67a4a9b5cf1aab391f5598a78da

Li, Z., Huang, Q., & Liu, J. (2024). ChatGPT analysis of strengths and weaknesses in English writing and their implications. Applied Mathematics and Nonlinear Sciences, 9, 1–15. https://doi.org/10.2478/amns-2024-0912

Mansour, W., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can large language models automatically score proficiency of written essays? arXiv. https://consensus.app/papers/can-large-language-models-automatically-score-mansour-albatarni/67747492f8145ae2bd5c3ac9dcb6b37b

Obata, A., Tagawa, T., & Ono, Y. (2023). Assessment of ChatGPT's validity in scoring essays by foreign language learners of Japanese and English. 2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter), 105–110. https://doi.org/10.1109/IIAI-AAI-Winter61682.2023.00028

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. Education and Information Technologies, 29, 24735–24757. https://doi.org/10.1007/s10639-024-12817-6

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? Computers and Education: Artificial Intelligence, 100, 255. https://doi.org/10.1016/j.caeai.2024.100255

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. International Journal of Assessment Tools in Education. https://doi.org/10.21449/ijate.1517994

Xia, W., Mao, S., & Zheng, C. (2024). Empirical study of large language models as automated essay scoring tools in English composition: Taking TOEFL Independent Writing Task for example. arXiv. https://doi.org/10.48550/arXiv.2401.03401

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. Research Methods in Applied Linguistics, 3(2), 100133. https://doi.org/10.1016/j.rmal.2024.100133

González, C., & Prendes, E. (2021). Artificial intelligence for student assessment. Education Sciences, 11(9), 542. https://doi.org/10.3390/educsci11090542

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100025. https://doi.org/10.1016/j.rmal.2023.100025

Ratnam, M., Sharma, B., & Tomer, A. (2023). ChatGPT: Educational artificial intelligence. International Journal of Advanced Trends in Computer Science and Engineering, 12(2), 587–595. https://doi.org/10.30534/ijatcse/2023/111222023

Page, E. B. (1966). The imminence of grading essays by computer. The Phi Delta Kappan, 47(5), 238–243.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 206–210. https://doi.org/10.3115/980845.980867

Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge. https://doi.org/10.4324/9780203122761

Tingting, B., Su, Y., & Zhang, F. (2023). Exploration of the application of artificial intelligence in the intelligent evaluation system of information and innovation education. Advances in Computer, Signals and Systems Clausius Scientific Press, Canada.

Yoon, S., Zadrozny, W., & Flor, M. (2020). A deep neural network approach to automated essay scoring. Educational Measurement: Issues and Practice, 39(2), 34–47. https://doi.org/10.1111/emip.12267

Perelman, L. (2014). When “the state of the art” is counting words. Assessing Writing, 21, 104–111. https://doi.org/10.1016/j.asw.2014.05.001