applied sciences

MDPI

Systematic Review

Large Language Model-Powered Automated Assessment: A Systematic Review

Emrah Emirtekin

Center for Distance Education Application and Research, Ege University, İzmir 35040, Turkey; emrah.emirtekin@ege.edu.tr

Abstract: This systematic review investigates 49 peer-reviewed studies on Large Language Model-Powered Automated Assessment (LLMPAA) published between 2018 and 2024. Following PRISMA guidelines, studies were selected from Web of Science, Scopus, IEEE, ACM Digital Library, and PubMed databases. The analysis shows that LLMPAA has been widely applied in reading comprehension, language education, and computer science, primarily using essay and short-answer formats. While models such as GPT-4 and fine-tuned BERT often exhibit high agreement with human raters (e.g., QWK = 0.99, r = 0.95), other studies report lower agreement (e.g., ICC = 0.45, r = 0.38). LLMPAA offers benefits like efficiency, scalability, and personalized feedback. However, significant challenges remain, including bias, inconsistency, hallucination, limited explainability, dataset quality, and privacy concerns. These findings indicate that while LLMPAA technologies hold promise, their effectiveness varies by context. Human oversight is essential to ensure fair and reliable assessment outcomes.

Keywords: automated essay scoring; short answer grading; artificial intelligence; human–AI alignment

Academic Editors: Tymoteusz I. Miller and Yoonsik Choe

Received: 14 April 2025
Revised: 9 May 2025
Accepted: 12 May 2025
Published: 20 May 2025

Citation: Emirtekin, E. Large Language Model-Powered Automated Assessment: A Systematic Review. *Appl. Sci.* **2025**, *15*, 5683. https://doi.org/10.3390/10.3390/app15105683

1. Introduction

Assessment in education is a crucial element used to determine whether learning has been achieved. Various types of questions are utilized in assessment processes, which can be categorized into two main groups: objective questions (multiple choice, true–false, fill-in-the-blank) and subjective questions (open-ended questions, essays) [1]. Today, objective questions are often preferred because they are easier to grade than subjective questions. However, subjective questions are considered to contribute to more long-term and effective learning than objective questions, as they encourage students to think critically, explain their thought processes, and apply their knowledge [2]. However, subjective questions present several challenges, including no single correct answer, the requirement for evaluators to analyze student responses, the time-consuming nature of assessment, increased workload for educators, and potential evaluator bias [3]. Considering these challenges, there has been growing interest in automated scoring systems, which are often assumed to improve consistency and reduce human bias. This assumption has been questioned. Recent research indicates that automated assessment tools can also perpetuate or amplify inherent biases [4]. As a result, there is a need for new tools to facilitate the evaluation of subjective questions. In this context, automated essay scoring (AES) and automated short answer grading (ASAG) have emerged as key components of automated assessment [5].

AES is a software-based system developed to automatically score essays [6]. Similarly, ASAG is a system that builds a model for automated short-answer scoring [7].

Although both systems use natural language processing (NLP) and machine learning (ML) to evaluate text, AES focuses on long-form essay scoring, whereas ASAG is designed for short-answer responses.

Recently, with the emergence of LLMs, their potential applications in education have become a focal point for researchers. The ability of LLMs to be adapted to various assessment tasks (e.g., grading, scoring) with minimal training data and their lack of need for technical expertise make them promising tools for assessing student learning more quickly and effectively while providing personalized feedback [8,9]. Thus, there has been a noticeable increase in research on LLM-powered AES and ASAG in the recent literature. While these capabilities are promising, they raise significant ethical and practical concerns. Emerging studies highlight that LLMs can exhibit inherent biases that may affect grading fairness [10]. Mendonça et al. found that algorithmic bias in LLMs can lead to inequitable scoring for minority and non-standard language students. Additionally, data privacy and security have been identified as pressing issues, since student information processed by LLMs requires robust protection. Thurzo emphasizes the need for provable ethical safeguards and transparency to build trust in AI educational tools [11]. Moreover, human oversight remains crucial: educators should review and contextualize automated scores to ensure reliability and address nuances that AI might miss [4]. However, it remains unclear how LLM-powered models influence assessment processes, their agreement with human scoring, the overall reliability and validity of these tools, and whether they can effectively conduct automated assessments.

In summary, this study systematically analyzes research on Large Language Model-Powered Automated Assessment (LLMPAA). Throughout this document, AES and ASAG studies incorporating LLMs are collectively referred to as LLMPAA. Within this scope, the following research questions are explored:

(RQ1) What are the usage domains of studies with LLMPAA?

(RQ2) What are the details of using LLMPAA?

(RQ3) What are the challenges in LLMPAA?

2. Materials and Methods

2.1. Search Procedures

In terms of methodology, this study adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [9]. The PRISMA guidelines serve to enhance the reliability of research findings by ensuring that systematic reviews are reported in accordance with internationally recognized standards. The PRISMA guidelines provide researchers with a checklist and a flow diagram to transparently document each stage of the review process. Adherence to these guidelines serves to enhance the methodological rigor of this systematic review and ensures the replicability of the research process. To promote transparency and reproducibility, the protocol for this review was preregistered on the Open Science Framework (OSF) and is publicly accessible at: https://osf.io/rwnsm (accessed on 8 May 2025).

2.2. Inclusion and Exclusion Criteria

The selection of studies for inclusion in this research was determined by the criteria outlined in Table 1. The inclusion criteria encompassed studies published between 2018 and 30 October 2024, appearing in peer-reviewed journals, written in English, and utilizing LLMs for grading or scoring. Conversely, studies that were excluded from the analysis comprised those that were duplicates, did not mention the use of LLMs in their abstract, were not written in English, did not utilize LLMs, were literature reviews, or had no accessible full text.

Table 1. Inclusion and exclusion criteria.

Include	Exclude
between 2018 and 30 October 2024	duplicates
peer-reviewed journals	not using LLM in the abstract section
written in English	not written in English
grading or scoring using LLM	not using LLM
	literature reviews
	no accessible full text

2.3. Information Sources and Search Strategy

This study focuses on recent research conducted on the role and impact of LLMs in assessment processes between 2018 and 30 October 2024. Within the scope of the study, the academic databases Web of Science, Scopus, IEEE Explore, ACM Digital Library, and PubMed were searched. The relevant databases were searched across all fields using the search terms provided in Table 2. The rationale for selecting 2018 as the initial year for the search is that it signifies the year in which the Bidirectional Encoder Representations from Transformers (BERT) model was unveiled and commenced utilization [12]. After the proclamation of BERT, research in the domain of natural language processing (NLP) underwent a marked acceleration, and studies on language models became more pervasive. Consequently, a comprehensive search was conducted, yielding 816 studies that met the specified criteria (Table 3).

Table 2. Search strings.

Topic	Search Terms (Boolean Operator: AND)
application	scoring OR grading OR evaluation
question types	short answer OR essay OR open-ended
language models	large language model OR llm OR chatgpt OR gemini OR gpt OR claude OR llama OR bert
date range	2018–2024 (30 October)

Table 3. Numbers of papers by database searched.

Databases	Number
Web of Science	287
Scopus	142
IEEE	49
ACM Digital Library	161
PubMed	177
Total	816

2.4. Screening and Filtering Process

The flow diagram illustrating the screening process is presented in Figure 1. According to Figure 1, 219 of the 816 studies were found to be duplicates and were consequently removed. The abstracts of the remaining studies were then reviewed based on the inclusion criteria, with those not aligned with the research objective excluded. The full texts of the remaining articles were then thoroughly evaluated. Following a comprehensive evaluation of the full texts of the 60 articles, 11 were excluded for the following reasons: non-use of LLMs, literature reviews, or inaccessibility of the full text. Consequently, the systematic analysis included a total of 49 studies.

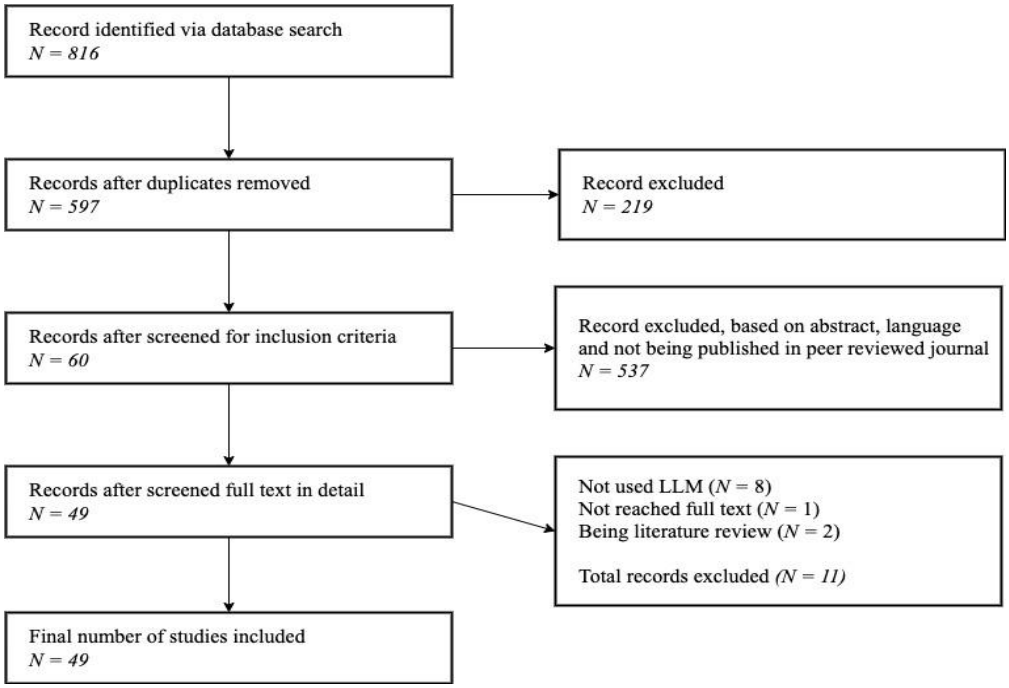

Figure 1. PRISMA flow diagram.

3. Results

3.1. Usage Domains

The analysis of LLMPAA usage revealed that these models have been applied across various disciplines, with higher concentrations in some fields. As shown in Table 4, LLMPAA were most used in reading comprehension (n = 14), language education (n = 13), and computer science (n = 10). They were also applied in medical education (n = 5) and writing skills (n = 5). Other disciplines, such as mathematics, physics, and chemistry, had only a limited number of studies (see Table 4).

Table 4. Dataset.

References	Domains	Datasets	Languages	Question Types	Sizes	Additional Info
[13]	Workplace Safety	ASAG-IT	Italian	Open-ended	10,560	Collected from final-year dental students.
[14]	Writing Skill	SELF	English	Essay	10,000	Essays collected from ESL learners during a university admissions test.
[15]	RC, CT, AT	ASAP-AES	English	Essay	12,978	Essays collected from students in grades 7–10 in the USA.
[16]	Computer Science	SPRAG	English	Short-answer	4039	Essays collected from students in grades 7–10 in the USA.
[17]	Computer Science, Language Education	ICNALE	English	Essay	200	Collected from college students across ten Asian countries/territories.
[18]	RC, CT, AT	ASAP-AES, ETS	English	Essay	25,078	Essays collected from students in grades 7–10 in the USA.
[19]	RC, CT, AT, Language Education	ASAP-AES, TOEFL iBT	English	Essay	28,989	Essays collected from students in grade 7 in the USA.

Table 4. *Cont.*

References	Domains	Datasets	Languages	Question Types	Sizes	Additional Info
[20]	Medical Education	SELF	English	Short-answer	37,346	Examined university entrance exams from 2020 to 2023.
[21]	Chemistry Education	SELF	English	Short answer	128	Collected from 199 undergraduate students at Universidad del Norte.
[6]	RC, CT, AT	ASAP-AES	English	Essay	12,978	Collected from final-year dental undergraduate students at the National University of Singapore.
[22]	Language Education	MERLIN	German, Italian, Czech	Essay	2267	Collected from students at the University of Montenegro.
[23]	Medical Education	SELF	English, French, German	Short-answer	2288	Collected from undergraduate medical education courses at the University of Luxembourg.
[8]	Reading Compre-hension	ROARS	English	Short-answer	1068	Collected from 130 students aged 9–18 in Ghana.
[24]	Psychology	SELF	English	Essay	378	Collected from non-native English speakers who took the TOEFL test in 2006 and 2007.
[1]	Computer Science	MOHLER, SELF	English	Short-answer	2273	Collected from the University of North Texas.
[25]	Language Education	SELF	English	Essay	78	Collected from Mega Italia Media's e-learning platform.
[26]	Writing Skill	SELF	English	Essay	20	Collected from national exam essays by the Indonesian Ministry of Education and Culture.
[27]	Computer Science	MOHLER, SemEval-2013	English	Short-answer	6214	Collected from 15 diverse science domains and the University of North Texas.
[28]	Physics	SELF	English	Essay	25	Collected from multilingual learners.
[29]	Computer Science	SciEntsBank	English	Short-answer	14,186	Collected from college students at Mindanao State University.
[30]	Writing Skill	UKARA	Indonesian	Short-answer	638	Collected from 5th- and 6th-grade primary school students in Tunisia.
[31]	Language Education	SELF	English	Essay	586	Collected from undergraduate computer science students and IT division employees.
[32]	RC, CT, AT	ASAP-AES, ELLIPSE	English	Essay	12,978	Collected from undergraduate students.
[33]	Language Education	I-JAS	Japanese	Short-answer	1400	Collected from non-native Japanese learners with 12 different first languages.
[34]	Medical Education	SELF	English	Essay	500	Collected from grade 3–6 students.

Table 4. *Cont.*

References	Domains	Datasets	Languages	Question Types	Sizes	Additional Info
[35]	RC, CT, AT, Writing Skill	SELF	Arabic	Essay	380	Generated by GPT-4 for grading physics proE30.
[36]	RC, CT, AT	SELF	Spanish	Short-answer	3772	Collected from students in grades 7–10 in the USA, EFL learners (ETS dataset: 12,100).
[37]	Cybercrime, History	AR-ASAG, SELF	Arabic	Short-answer	2183	Collected from a cybercrime and history course exam.
[38]	Language Education	TOEFL11	English	Essay	12,100	Collected from high school students in Bandung, Indonesia.
[39]	Divergent Thinking	SELF	English	Open-ended	27,217	Responses from 2039 participants across nine datasets.
[40]	Language Education	SELF	English	Essay	119	Collected from essays on 'Healthy Diet'.
[41]	Medical Education	SELF	English	Essay	69	Essays collected from students in grades 7–10 in the USA and EFL learners.
[42]	RC, CT, AT, Computer Science	ASAP-AES, SELF	English	Essay	15,368	Essays collected from students in grades 8 and 10 in the USA.
[43]	RC, CT, AT, Computer Science	ASAP-AES, SELF	English	Essay	15,959	Collected from 8th- to 10th-grade students in the USA.
[44]	Computer Science	SELF	Indonesian	Short-answer	3977	Collected from Asian learners of English, rated by 80 human raters and ChatGPT-4.0.
[45]	Medical Education	SELF	English	Essay	10	Responses collected from medical students in a high-stakes test context.
[46]	RC, CT, AT	ASAP-AES	English	Essay	12,978	Essays collected from students in grades 7–10 in the USA.
[47]	Writing Skill	SELF	Chinese	Essay	600	Collected from 2870 Chinese primary school students in grade 3.
[48]	RC, CT, AT	ASAP-AES	English	Essay	1730	Essays collected from students in grades 7–10 in the USA and collected from Chinese EFL learners.
[49]	Language Education, History	PATHWAY 2.0, Crossley, WRITE CFT, SELF	English	Essay	1786	Collected from students at Universitas Terbuka, Indonesia.
[2]	Mathematics Education	SELF	Spanish	Open-ended	677	Responses from fourth-grade students.
[50]	Language Education	ELLIPSE	English	Essay	480	Collected from TOEFL test takers.
[51]	Computer Science	MOHLER	English	Short-answer	2273	Collected from the University of North Texas.

Table 4. *Cont.*

References	Domains	Datasets	Languages	Question Types	Sizes	Additional Info
[52]	Computer Science	SELF	Indonesian	Short-answer	480	Collected from foreign learners at the Intermediate 1 level.
[53]	RC, CT, AT	ASAP-AES	English	Essay	5152	Essays collected from students in grades 7–10 in the USA.
[54]	Computer Science	MOHLER, SemEval-2013	English	Short-answer	13,077	Collected from 15 diverse science domains and the University of North Texas.
[55]	RC, CT, AT, Language Education	ASAP-AES, CELA	Chinese	Essay	13,122	Collected engineering faculty students.
[56]	Language Education	ICNALE-GRA	English	Essay	136	Collected from language, history, and writing courses.
[57]	Language Education	SELF	English	Essay	3	Rated by 15 human raters.

"Domains" indicates the academic fields where each dataset was applied (e.g., reading comprehension, language education). "Datasets" includes either publicly available benchmark datasets or custom datasets created by the researchers, indicated with SELF In such cases, manual annotation was typically performed by domain experts using holistic or trait-based rubrics. "Languages" refers to the primary language(s) used in student responses. "Question Types" are categorized as essay, short-answer, or open-ended. "Sizes" denote the total number of responses per study; in cases where multiple datasets were used, the combined total is reported. "Additional Info" provides methodological details including the data source (e.g., school level, subject area) and annotation process (e.g., number of raters). Public datasets such as ASAP-AES and MOHLER are pre-annotated and commonly used as benchmarks in LLMPAA research. Domain abbreviations: RC = reading comprehension, CT = creative thinking, AT = analytical thinking.

3.2. Details of LLMPAA Research

3.2.1. Datasets

An analysis of the datasets used in LLMPAA studies reveals that a significant share of the research (n = 23) involved creating custom datasets tailored to the specific study context. This suggests that each study requires specialized datasets for its educational context, student profile, or exam type. On the other hand, some studies have been found to use pre-existing datasets (Table 4). ASAP-AES (n = 11) and MOHLER (n = 4) were preferred more frequently than other pre-existing datasets. The ASAP-AES dataset is an automated essay scoring (AES) dataset published as part of the Automated Student Assessment Prize (ASAP) competition on Kaggle (https://kaggle.com/competitions/asap-aes, accessed on 15 December 2024). This dataset contains over 10,000 essays written by 7th-, 8th-, and 10th-grade students on eight different topics. Each essay has been scored by multiple independent evaluators, and the final scores are included in the dataset. The ASAP-AES dataset is a widely used benchmark for LLMPAA development and comparison and is referenced in most studies in this field. The MOHLER dataset, specifically developed for computer science, consists of short-answer questions and the corresponding grades provided by experts. Table 4 details the question types, languages, and dataset sizes used in the studies. These data enable a comprehensive evaluation of different dataset parameters and yield clearer insights for the analysis.

3.2.2. Languages, Question Types, Sizes, and Additional Info

The studies classified question types into three groups: essay (n = 30), short-answer (n = 16), and open-ended (n = 3). Essay questions were particularly prevalent in high school and university studies. This format was preferred for assessing students' critical thinking and writing skills in depth. investigated the writing proficiency of high school students using a corpus of 28,989 essays. A comprehensive analysis conducted in study [19]

investigated the writing proficiency of high school students using a corpus of 28,989 essays. Short-answer questions have been more commonly used in university-level studies and scientific fields. This question type was preferred to assess students' ability to provide concise and clear answers. One study [20] evaluated student performance in university entrance examinations by analyzing a dataset of 37,346 short-answer responses. Open-ended questions, although less commonly used, have been preferred particularly in studies at the elementary and middle school levels. This question type was used to assess students' creativity and open-ended thinking skills. A study [2] examined elementary school students' writing skills using 677 open-ended responses.

When examining the language of the datasets used in the studies, it was found that 67% of the studies used English, while 33% used other languages (Spanish, Japanese, Arabic, Chinese, Indonesian, German, Italian, Czech). This heavy reliance on English raises concerns about applicability in multilingual contexts and may disadvantage non-native English speakers [58]. A study [33] assessed a dataset consisting of 1400 essays written in Japanese by students with 12 different native languages to investigate the impact of the native language on the language learning process. Similarly, Firoozi and colleagues [22] used the MERLIN dataset, which contains a total of 2267 essays written in German, Italian, and Czech, to compare students' writing skills in multilingual education systems. Furthermore, studies using languages such as Arabic and Chinese have focused on the teaching and learning of these languages. In study [37], students' understanding of history and cybercrime was assessed using Arabic short-answer questions.

When examining the dataset sizes in the studies, it was found that the dataset sizes used in the 49 studies varied between 3 and 37,346. This large variance indicates that the studies were designed for different purposes and target audiences. The dataset size for essays ranges from 3 [57] to 28,989 [19]. Large-scale essay datasets were used to assess the writing skills of high school and university students. The dataset sizes for short-answer questions range from 128 [21] to 37,346 [20]. These types of datasets have been primarily used in university entrance exams and scientific assessments. The dataset sizes for open-ended questions range from 677 [2] to 27,217 [39]. These types of datasets have been primarily used to assess the creativity and open-ended thinking skills of elementary and middle school students. When examining the relationship between dataset sizes and question types, it is observed that short-answer questions are used in larger datasets. This is because short-answer questions are more suitable for automated assessment, whereas essay and open-ended questions have been used in smaller datasets. Geographically, studies in North America and Europe tend to use English datasets, while studies in Asia and the Middle East focus on local languages. For example, a study conducted in Indonesia [30] evaluated 5th- and 6th-grade students' responses to short-answer questions in Indonesian.

3.2.3. LLM's and Evaluation Metrics

In this systematic review, the full texts of the 49 selected studies were carefully analyzed to extract the metrics reported for evaluating LLMPAA systems. Rather than conducting new statistical tests, the study focused on identifying and categorizing the validation metrics used in each study. These metrics were descriptively compiled and classified according to their function—whether they assessed model accuracy, agreement with human raters, or inter-rater consistency. This approach enabled a comparative understanding of how performance and reliability are reported across different research contexts and also revealed areas where statistical validation practices could be strengthened. Based on the analysis, the LLMs utilized and the corresponding evaluation metrics are summarized in Table 5. The metrics employed across the 49 studies have been categorized into three groups:

- Metrics for determining model performance (accuracy, F1-score, precision, and recall).
- Metrics based on measuring the agreement of LLMs with human (Cohen's kappa—CK, linear weighted kappa—LWK, quadratic weighted kappa—QWK, correlation coefficient—r).
- Metrics based on measuring the reliability and consistency between different evaluators (human or LLM) (intraclass correlation coefficient—ICC).

Table 5. Overview of LLMs and evaluation metrics.

References	LLMs	Evaluation Metrics	Results
[13]	BERT, ELECTRA, Multilingual BERT, RoBERTa	F1: 0.72	BERT model performed the highest F1 score.
[14]	BERT-XLNET, CNN-BiLSTM, BERT-BiLSTM, R2BERT, ATT-CNN-LSTM	CK: 0.91	BERT-BiLSTM achieved better agreement compared to other models.
[15]	30-manually extracted features + 300-word2vec + 768-BERT, 300-word2vec + 768-BERT, 768-BERT, 30-manually extracted features + 768-BERT, 300-word2vec	ACC: 0.75, QWK: 0.77	The best agreement was achieved with the 30-manually extracted features + 300-word2vec + 768-BERT model.
[16]	all-MiniLM-L6-v2, paraphrase-MiniLM-L6-v2, paraphrase-albert-small-v2, quora-distilbert-base, stsb-roberta-large, multi-qa-MiniLM-L6-cos-v1, multi-qa-distilbert-cos-v1, stsb-distilbert-base	ACC: 82.56, *Abstract*ρ: 0.64, r: 0.69/ACC: 56.11, *Abstract*ρ: 0.76, r: 0.73	Different models showed accuracy up to 82.56% in binary classification and up to 56.11% in multi-class classification.
[17]	GPT-4	r: 0.382; ICC: 0.447	GPT-4 showed weak correlation with human and low consistency.
[18]	ABC-BERT-FTM, BERT	QWK: 0.9804	The ABC-BERT-FTM model achieved the highest agreement.
[19]	BigBird, BERT, DualBERT-Trans-CNN, Considering-Content-XLNet, MTL-CNN-BiLSTM, Trans-BERT-MS-ML-R	QWK: 0.791	The highest performance was achieved with the Trans-BERT-MS-ML-R model.
[20]	ACTA, BERT	QWK: 0.99	The ACTA model achieved the highest agreement with human.
[21]	GPT-4	r: 0.85	GPT-4 showed successful results with high correlation and low error rate.
[6]	BERT	QWK: 0.78	BERT achieved high agreement after fine-tuning.
[22]	LaBSE, mBERT	ACC: 0.88, CK: 0.83, QWK: 0.85	LaBSE performed better in all languages, showing superior performance in QWK and accuracy.
[23]	Gemini 1.0 Pro, GPT-4	ACC: 0.68 (binary), 0.59 (multi)	Gemini performed better in binary classification, while GPT-4 showed better performance in multi-class classification.
[8]	GPT-3.5 Turbo (Few-Shot, Zero-Shot), GPT-4 (Few-Shot, Zero-Shot)	LWK: 0.94 (2-class), QWK: 0.91 (3-class)	GPT-4 (Few-Shot) achieved the highest agreement.
[24]	Qwen-max-0428	r: 0.491	Qwen-max-0428 achieved low agreement with humans.
[1]	BERT (fine-tuned), LSTM	r: 0.761	BERT + LSTM showed the highest correlation.

Table 5. *Cont.*

References	LLMs	Evaluation Metrics	Results
[25]	GPT-3.5	ICC: 0.81	GPT-3.5 achieved high consistency with human.
[26]	GPT-3.5	ICC: 0.56	GPT-3.5 achieved medium consistency in language criteria.
[27]	BERT, RoBERTa, XLNet, SBERT	ACC: 0.76 (3-class), 0.82 (2-class)	BERT showed the most consistent and highest performance as the recommended model across all datasets; other models were competitive in specific tasks but generally performed lower.
[28]	GPT-4	r: 0.84	GPT-4 showed high correlation.
[29]	GPT-4	P: 0.788	GPT-4 showed high precision in the SciEntsBank dataset.
[30]	CNN-LSTM, IndoBERT	QWK: 0.465	IndoBERT achieved a 14.47% increase in agreement compared to CNN-LSTM.
[31]	BERT	ACC: 0.958	BERT performed with high accuracy.
[32]	CNN-LSTM-ATT, LSTM-MoT, SkipFlow LSTM, TSLF, Sentence-BERT, CNN-LSTM-ATT, Tran-BERT-MS-ML-R	QWK: 0.852	TSLF model achieved the highest agreement.
[33]	BERT, GPT-4, OCLL	QWK: 0.819	GPT-4 achieved the highest agreement.
[34]	GPT-4	ACC: 0.75	GPT-4 performed with high accuracy with humans.
[35]	AraBERT	r: 0.88	AraBERT showed high correlation with humans.
[36]	BERT-1-EN, BERT-1-ES, BERT-2-EN, BERT-1-MU, BERT-2-MU, Skip-Thought, BERT-2-ES	r: 0.83	BERT-1-ES showed the best results.
[37]	BERT, Word2vec, WordNet	r: 0.841	BERT showed the highest result across all datasets.
[38]	GPT-3 text-davinci-003	CK: 0.682	GPT-3 achieved agreement with humans.
[39]	BERT, GPT-3 (ada, babbage, curie, davinci), GPT-4	r: 0.813	Davinci showed the highest correlation.
[40]	GPT-3.5, GPT-4, Claude 2, PaLM 2	ICC: 0.927, r: 0.731	GPT-4 showed the highest consistency.
[41]	GPT-4	ICC: 0.858	GPT-4 showed excellent consistency with humans.
[42]	BERT, sentence embedding-LSTM	QWK: 0.766	The best agreement was achieved with the BERT model.
[43]	Bi-LSTM, Sentence-BERT	QWK: 0.76	The best agreement was achieved with the Sentence-BERT model.
[44]	ALBERT, BERT	r: 0.950	ALBERT showed the highest performance.
[45]	GPT-3.5	-	It showed high agreement with humans.
[46]	Aggregated-BERT + handcrafted + SVR, BERT-64 + LSTM, HISK + BOSWE(SVR), Skipflow, TSLF-all	QWK: 0.81	The highest agreement was achieved by Aggregated-BERT + Handcrafted + SVR.
[47]	ChatGLM2-6B, Baichuan-13B, InternLM-7B	QWK: 0.571	ChatGLM2-6B achieved the highest agreement.

Table 5. *Cont.*

References	LLMs	Evaluation Metrics	Results
[48]	Claude 2, GPT-3.5, GPT-4	QWK: 0.567	The highest agreement was achieved with GPT-4.
[49]	GPT-4	CK: 0.88	GPT-4 achieved the best agreement.
[2]	BLOOM, GPT-3, YouChat	P: 0.488, R: 0.912	GPT-3 showed high recall and low precision.
[50]	GPT-3.5 (Base), GPT-3.5 (fine-tuned: FTP1, FTP2, FTPB), GPT-4 (Base)	QWK: 0.78	Fine-tuned models FT-P1 achieved the highest QWK value.
[51]	all-distilroberta-v1, all-MiniLM-L6-v2, paraphrase-albert-small-v2, paraphrase-MiniLM-L6-v2, multi-qa-distilbert-cos-v1, multi-qa-mpnet-base-dot-v1, stsb-distilbert-base, stsb-roberta-large	r: 0.9586	All-distilroberta-v1 showed the highest correlation.
[52]	BERT	CK: 0.75	BERT achieved agreement with humans.
[53]	BERT, Bi-LSTM, CNN-T BERT, fine-tuned BERT, Xsum-T BERT, YAKE-T BERT	QWK: 0.789	Xsum-T BERT and YAKE-T BERT achieved the highest agreement.
[54]	BERT, Bi-LSTM, Capsule Network, fine-tuned BERT	r: 0.897	BERT showed high correlation with the extended dataset.
[55]	BERT-MTL-fine-tune, BERT-MTL-fine-tune (CELA dataset), BERT-fine-tune	QWK: 0.83	BERT-MTL-fine-tune achieved the highest agreement.
[56]	GPT-4	r: 0.67–0.82	GPT-4 showed medium correlation with humans.
[57]	Bard, GPT-3.5 (default, fine-tuned)	ICC: 0.972	Fine-tuned GPT-3.5 performed with the highest consistency.

This table summarizes the LLMs used in the reviewed studies and the evaluation metrics applied. "Model" refers to the specific LLM or its variant (e.g., GPT-4, BERT), and indicates whether it was pre-trained, fine-tuned, or used in a zero-shot setting. "Evaluation Metrics" are grouped into three categories: (1) performance (accuracy, F1-score, precision, recall), (2) agreement with human raters (QWK, CK, LWK, r), and (3) inter-rater reliability (ICC). Studies often used multiple metrics to strengthen validity. Fine-tuned models were typically trained on labeled datasets aligned with scoring rubrics. In some cases, prompt engineering and temperature adjustments were used to improve scoring consistency and control output variability.

It has been determined that the performance of LLMs is measured using the following metrics: ACC (n = 7), F1-score (n = 2), precision—P (n = 2), and recall—R (n = 1). These metrics are employed to assess the models' ability to predict actual assessment outcomes, which are represented on a scale from 0 to 1. A performance value closer to 1 indicates higher model effectiveness. According to the analysis, one study [31] reported a comparatively high accuracy, with a BERT-based model achieving an ACC of 95.8% in evaluating the written responses of foreign students learning Korean. Furthermore, some studies utilized multiple metrics in addition to ACC. For example, in a study aimed at developing an ASAG system to assess student responses in distance education objectively and quickly, it was found that the BERT model provided a good level of accuracy (ACC = 0.80) and correlation (r = 0.747) [27]. On the other hand, in a study [16] which aimed to create and assess the SPRAG dataset for ASAG related to programming languages, the developed model was reported to achieve high accuracy (ACC: 0.8256; r = 0.69). In studies investigating the agreement of LLMs with human (N = 40), it has been found that the metrics QWK (n = 18), CK (n = 5), LWK (n = 1), and r (n = 16) were the most frequently used. One study [20], which focused on assessing ASAG performance in medical education, reported a notably high QWK value of 0.99. Analysis of the studies employing the QWK metric (n = 15) revealed that 40% achieved excellent agreement (QWK = 0.81–1.00), 33% demon-

strated good agreement (QWK = 0.61–0.80), and 20% reported moderate agreement (QWK = 0.41–0.60). In studies that only used the CK metric (n = 4), it was found that in half of the cases, the model showed excellent agreement, while in the other half, the model demonstrated good agreement. Among the studies that exclusively examined the correlation coefficient (n = 12), 91% reported strong agreement (r = 0.70–1.00), while one of the studies [24] reported low agreement (r = 0.30–0.50). A review of the literature identified three studies that assessed the consistency and reliability between different evaluators (human- or LLM-based). These studies employed the intraclass correlation coefficient (ICC, n = 6) to evaluate the alignment between the model-generated scores and those provided by various raters. In two of these studies, the model demonstrated a good level of consistency (ICC = 0.75–0.90), whereas in one study, a moderate level of consistency was observed (ICC = 0.50–0.75). Additionally, there are studies in which both ICC and correlation coefficient (r) metrics were used concurrently. For example, study [17], which investigated the reliability of ChatGPT as an AES tool by comparing its performance with that of an experienced human evaluator, reported that GPT-4 exhibited weak correlation and low consistency with human ratings (r = 0.359; ICC = 0.447). In contrast, study [40], which compared the reliability and consistency of four different LLMs in assessing the writing of English language learners, found that GPT-4 demonstrated the highest consistency among the models examined (ICC = 0.927; r = 0.731).

When examining the LLMs used in the studies, BERT and its variants (RoBERTa, XLNet, etc.) have consistently shown high performance across various tasks. Especially when fine-tuned, BERT has been observed to give the best results in various text evaluation metrics (F1, QWK, r, etc.). It has been determined that GPT-4 and fine-tuned BERT models stand out as LLMs achieving the best results in LLMPAA, demonstrating high performance and positioning themselves among the most powerful and effective automated assessment tools. As a result,

- While LLMPAA have frequently demonstrated good or excellent agreement with human raters, some studies have reported moderate or low alignment, suggesting that their effectiveness may vary depending on the context and application.
- BERT and its variants have generally provided a strong foundation for assessment tasks, although performance can differ based on task type and fine-tuning strategies.
- GPT-4 has shown promising results for human-like assessment in many studies, but findings also highlight variability in accuracy and consistency across different scenarios.
- The effectiveness of each model depends on the specific requirements of the task, dataset characteristics, and implementation context.
- Fine-tuning has been shown to significantly improve model performance in several cases, though its impact may depend on the quality and diversity of training data.

3.3. Challenges in LLMPAA

3.3.1. Bias and Fairness

Research has found that the LLMPAA risks reflecting biases and inconsistencies in human assessments. This can lead to ethical issues and fairness concerns in assessment processes. Assessments, particularly those conducted in different linguistic and cultural contexts, may be influenced by this risk. For example, it was observed in a study [36] that, despite LLMs not being influenced by humans' subjective perceptions, they sometimes have stricter grades in certain cases. It has been noted that prior knowledge of human grading increases the risk of bias in LLMs [23]. It has been suggested that LLMs may show variations in assessing student responses based on demographic factors, such as gender or the language spoken at home [2]. Additionally, Lin and Chen found that AI-driven feedback often imposes rigid frameworks that limit creative thinking and cause student

frustration, suggesting negative effects on learner motivation [59]. Similarly, algorithmic bias can exacerbate inequity: Mendonça et al. found that LLM scoring unfairly penalizes students from minority or non-standard linguistic backgrounds [10]. Thurzo emphasizes that human oversight and clear governance frameworks are essential when integrating AI into assessment [11]. These issues highlight the importance of inclusive design and strong governance in LLMPAA deployments. For these reasons, identifying and minimizing biases is crucial for achieving fair outcomes [13,20]. Moreover, the risk of cultural and linguistic bias in LLMPAA deserves special attention. Although LLMs are designed to be neutral, several studies have shown that these systems may favor dominant linguistic norms embedded in the training data. Grévisse reports that automated scoring models demonstrated inconsistencies when evaluating responses from multilingual or non-Western students, particularly when the writing style deviated from standard academic English [23]. Likewise, Urrutia and Araya found that learners whose home language differed from the language of instruction were more likely to receive lower scores, despite similar content quality [2]. These findings suggest that LLMPAA may unintentionally penalize students who use regionally influenced grammar, vocabulary, or culturally grounded rhetorical styles. This has ethical and pedagogical implications, especially for non-native English speakers and students from underrepresented cultural backgrounds. Given that 67% of LLMPAA studies rely on English-language datasets, the models may be inadvertently trained to assess only a narrow range of linguistic expressions. To address this issue, researchers advocate for the inclusion of multilingual and culturally diverse datasets, bias-aware training practices, and regular fairness evaluations. Human raters with cultural sensitivity can also play a vital role in validating model output. Such practices are essential not only for enhancing model performance but also for ensuring equity and inclusivity in educational assessments.

3.3.2. Consistency

Research has indicated that one of the most common issues with LLMPAA is the problem of consistency. Consistency refers to the ability to produce similar outputs for similar inputs. It is critically important, especially when the same response is evaluated multiple times, as it ensures the same results [23]. This issue is a consequence of the probabilistic nature of LLMs. Parameters like temperature (which controls how random the model's assessments will be) can introduce variability in the model's outputs, reducing its deterministic behavior. Additionally, the patterns the model learns from training data lead to instability in assessments for certain topics or types of questions. Bui and colleagues [17] found that LLMPAA results did not align fully with human assessments, showing low consistency. Moreover, another study demonstrated lowering the temperature setting improved the model's consistency, although this effect was not always significant [48].

3.3.3. Dataset Quality

Research indicates that the quality of the datasets used in the assessment process, including their type, size, and diversity, directly influences the performance of the model and the assessment outcomes. High-quality and representative datasets enable the model to understand different writing styles and content types [36], while low-quality or imbalanced datasets may negatively affect the model's generalization ability [54]. Bonthu and colleagues [16] highlight that small-scale and limited datasets used in assessments may pose challenges in assessing more complex or specialized topics. Zhu and colleagues [54] suggested that the limited performance of LLMs is due to small datasets. Firoozi and colleagues [22] argue that if training data lacks diversity in terms of language, cultural context, or writing styles, the model may struggle to assess responses accurately from

diverse student groups. Similarly, Meccawy and colleagues [37] argue that the results of their study using only Arabic datasets are insufficient for generalization, and that the model they developed needs to be replicated on different datasets for broader results. Mardini and colleagues [36] created a dataset consisting of student responses to Spanish reading comprehension questions but found that the dataset lacked sufficient examples in the high-grade range, rendering it imbalanced, and suggested that the model could be improved with a balanced dataset. Therefore, it can be stated that for the success of LLMPAA, the training data must be of high quality, diverse, and representative of the target audience.

3.3.4. Explainability and Transparency

The concepts of explainability and transparency are emphasized as being of critical importance in LLMPAA studies. These concepts enhance the reliability of model decisions by enabling a better understanding of how assessment processes occur. However, the assessment processes of LLMs are typically "black box" in nature, presenting various challenges in terms of explainability and transparency [60]. Urrutia and colleagues [2] specifically noted that in the context of assessing the mathematical reasoning skills of fourth-grade students, LLMs perform worse than traditional ML models in terms of their ability to explain decision-making processes. Yavuz and colleagues [57] highlighted that the lack of ability to explain how LLMs make decisions when students dispute automated assessment results could jeopardize student rights and accountability in education. For example, one study [50] found that fine-tuned models tend to assign higher scores than appropriate to texts in lower score categories. For example, The FTP1 model gave lower scores than human raters to many texts rated three, but its assessments aligned more closely with human ratings for higher-scoring texts. Consequently, LLMs must present the rationale behind the scores (or assessments) they assign to answers in a clear and transparent way. However, due to the nonlinear and high-dimensional data representations inherent in artificial neural networks that underlie LLMs, justifying the decisions made by these models is quite challenging. This presents a significant issue, especially for educators and students questioning the explainability and transparency of assessment results [60]. While LLMs possess high accuracy and generalization capabilities, previous research has often neglected the explainability dimension, leading to opacity in decision-making processes [48]. Although it is difficult to control the concepts of "explainability" and "transparency" in LLMPAA studies, Song and colleagues [47] emphasized in their study that assessments should be rubric-based to increase the reliability of evaluations. Wu and colleagues [53] proposed a model called Topic-Aware BERT, which aims to enhance the explainability and transparency of LLMPAA by considering the relationships between student essays and their scores, while also incorporating subject-specific information from the assignment instructions.

3.3.5. Hallucinations

LLMs have a strong potential due to their summarization abilities [61,62]. While this is seen as a significant advantage of LLMPAA compared to traditional assessment systems, LLMs can "hallucinate". Hallucination refers to the phenomenon where LLMs generate information that does not exist, is incorrect, or is out of context [21,41]. This occurs because the model fills in information that is not present in its training data, which can lead to incorrect assessments [63]. The model generates hypothetical data to complete missing information, which poses significant issues in fields requiring scientific accuracy [64–66]. In LLMPAA, hallucination occurs when the model incorrectly analyzes student responses, uses non-existent information as a scoring criterion, or generates inaccurate feedback. For

example, one study [40] reported that the model treated a paragraph structure as if it were present in a student's essay, even though it was not. This results in a flawed representation of the student's actual performance [21].

3.3.6. Language and Human Characteristics

In research, various issues related to language and human characteristics are highlighted. One of the significant problems encountered in LLMPAA is the inability of LLMs to understand the structure of language and the complexity of language itself [46]. It has been pointed out that the scoring algorithms of LLMs may not fully comprehend complex texts [13,20], fail to identify spelling errors [2], struggle to recognize irrelevant or incorrect content, and face difficulties in assessing the subtleties and creativity of writing [51]. Additionally, due to their inability to provide human empathy and emotional intelligence, this may create a negative experience in the assessment process [41]. For instance, Bui and Barot [17] state that ChatGPT cannot fully capture the multidimensionality of writing, and that humans may be needed in final assessments to grasp the nuances and creativity in the writings of students whose first language is not English. Kortemeyer, in a study on the automatic assessment of handwritten solutions to physics problems by students using the GPT-4 model, found a high correlation (r = 0.84) with human assessment [28]. However, it was determined that GPT-4 tends to give higher scores compared to humans. Therefore, it is stated that while GPT-4 is less stringent in assessing incorrect solutions, it shows some inconsistencies in assessing symbolic and numerical calculations. Thus, it is indicated that the inability of LLMs to fully comprehend human language limits their ability to perform effective assessments [45], and there is a risk that LLMs may fail to adequately assess certain language features and provide inconsistent results in certain situations [33].

3.3.7. Privacy and Security

In the existing research, it is noted that LLMPAA brings significant privacy and security concerns. These concerns focus on data anonymization, the security of student-generated content, and the potential risks associated with cloud-based LLM services. When commercial LLMs such as ChatGPT or Gemini are used in LLMPAA studies, the anonymization of student responses becomes necessary. These models typically process data on external servers. Beyond data anonymization, security concerns arise because platforms conducting LLMPAA online upload student content to external servers. Song [47] points out that such platforms not only charge high fees but also expose student data to security vulnerabilities. The terms of service of certain commercial LLM providers, such as OpenAI's ChatGPT model, indicate that user interactions may be used "for the purposes of providing, maintaining, improving, and enhancing the services". This leads to uncertainty regarding the destination of the uploaded data. Kasneci and colleagues [67] emphasize that the compliance of AI models with ethical and legal standards should be thoroughly assessed before their widespread use in classrooms. This highlights the importance of developing institutional policies to ensure the privacy of student data and protect it from potential misuse. To address this issue, it is recommended that LLMs be deployed offline and locally, thereby enhancing the security of student data. In response to these challenges, it has been emphasized that LLMPAA processes should be conducted on local servers using open-source LLMs [8,47]. Furthermore, recent discussions emphasize that commercial cloud-based models like GPT-4 and Gemini introduce additional privacy risks due to their external data handling and unclear storage policies. These models may store and process student inputs in ways that are not fully transparent to educational institutions. As highlighted by Kasneci et al., such risks can include unintended data retention, breaches, or misuse [67]. To address these concerns, it is strongly recommended to implement privacy-preserving

approaches such as federated learning, which enables decentralized model training without transmitting raw student data. Additionally, deploying open-source LLMs on institutional servers, using strong encryption and anonymization techniques, and ensuring compliance with GDPR or similar data protection regulations are vital to safeguarding student information [68]. These strategies should be considered essential when evaluating the suitability of commercial AI tools in education. Kortemeyer emphasizes that although open-source LLMs exhibit lower performance compared to GPT-4, they may still be a viable alternative, highlighting the need for further advancements in their performance [29].

4. Discussion

In recent years, research on the use of LLMPAA in education has expanded. In this study, the literature on LLMPAA was systematically reviewed. Based on findings from 49 studies, various limitations, problems, and advantages related to the use of LLMPAA in educational assessment were identified. Research results reveal that LLMPAA has been extensively applied in reading comprehension, language education, medical education, and computer science (Table 4) [69]. However, LLMPAA studies in fields such as physics, chemistry, and mathematics education are limited. This can be attributed to the abstract and symbolic nature of these disciplines, the difficulty language models face in processing mathematical and physical problems effectively, and the complexity of the data [70,71]. Polverini and Gregorcic investigated ChatGPT's performance in solving kinematics problems and found that the model, despite producing coherent language outputs, struggled to interpret graphical data and lacked the capacity for symbolic reasoning [71]. Their findings illustrate a broader limitation of LLMPAA: while these systems are proficient in language-based tasks, they face significant challenges when applied to domains that require reasoning with formulas, symbols, or diagrams. This suggests that current LLMPAA technologies may not be directly generalizable to all disciplines and should be complemented with domain-specific tools or multimodal models for reliable assessment in such areas.

While LLMPAA demonstrate strong capabilities in automating assessments, they are not without limitations, particularly in subjective, creative, or culturally nuanced evaluation contexts. Recent research suggests that hybrid models, which integrate human feedback with model predictions, can mitigate issues such as bias, inconsistency, and hallucination [72,73]. One notable approach is reinforcement learning with human feedback (RLHF), which allows models to better align with human judgment by iteratively learning from human-provided rewards or preferences. RLHF replaces fixed, engineered reward functions with dynamic, human-centered feedback, making it especially valuable in educational settings where scoring criteria may be ambiguous or context-dependent. Such integration not only enhances fairness and reliability, but also preserves interpretive depth in complex assessments. Incorporating human oversight through mechanisms like review checkpoints or rule-based constraints allows LLMPAA to offer both scalability and nuanced judgment. Therefore, future implementations of LLMPAA should consider hybrid frameworks as a strategy to balance automation with human-centered values [72].

Among the studies reviewed, Pack et al. stands out as the only one that included a Claude model—specifically Claude 2—in its comparative evaluation of LLMPAA tools [40]. Their findings showed that Claude 2 achieved high inter-rater consistency (ICC = 0.927) and strong correlation with human raters (r = 0.731), placing it on par with GPT-4 in assessment reliability. Although Claude 3.5 was not available during the study period, and therefore not evaluated in these 49 articles, it represents a significant advancement over its predecessor. Anthropic reports that Claude 3.5 outperforms previous Claude versions and demonstrates competitive performance with GPT-4, particularly in tasks involving long-context reasoning and factual accuracy [74]. Given this trajectory, future research

should systematically include Claude 3.5 to ensure that performance comparisons reflect the most recent state-of-the-art capabilities in LLMPAA [40,74].

When the datasets were analyzed, it was determined that more than half of the studies used pre-existing datasets, while the others used custom datasets. The use of pre-existing datasets helps to establish comparability between studies because these datasets have been tested in previous studies and their results have been widely evaluated in the literature. This allows for a consistent performance comparison [27,53]. Another important reason is that, beyond being accessible, they provide a solid foundation for research by providing labeled data [32]. Labeled data are critical for accurate training and testing of systems. With labeled data, researchers can measure the performance of their systems more effectively. Other reasons these datasets are preferred include that they have large and diverse content [18], have been published and validated for accuracy, and are easily adaptable to language learning, reading comprehension, text analysis, and similar fields. In addition, the fact that studies use a wide range of models, development strategies, and testing processes across different datasets is important for monitoring progress in this field. On the other hand, it has also been reported that these datasets are often quantitatively unbalanced, i.e., some datasets have many examples while others have insufficient number of examples [52]. Such imbalances may affect the learning process of the model and may negatively affect the performance of the model in some categories.

The analysis reveals that the quality of the dataset affects model performance and parameters such as agreement, consistency, and reliability among different raters. In some studies, data insufficiency [16], data diversity [22], and unbalanced data distribution [36,54] directly affect model performance and evaluation results. To ensure that the model provides consistent results across different contexts, balanced, diverse, and comprehensive datasets should be created. Data augmentation techniques can be applied to overcome the data insufficiency. Synthetic data generation or data collection from different sources can be used, especially for underrepresented samples [16]. On the other hand, multiple data sources and datasets with different perspectives can be included in the training process of the model to increase data diversity [22]. While this review includes studies that utilize non-English datasets, the dominance of English remains a significant limitation. Approximately two-thirds of the studies reviewed rely exclusively on English-language data, which may constrain the applicability of LLMPAA in multilingual or culturally diverse educational contexts. This overrepresentation of English may inadvertently bias model development and evaluation, leading to reduced fairness for students using other languages or dialects. Although some studies—such as those by Li and Liu [33] and Firoozi et al. [22]—address these concerns, broader and more systematic inclusion of low-resource and typologically diverse languages is needed. Future research should focus on evaluating LLMPAA models across varied linguistic and cultural settings to ensure their effectiveness and equity in global education systems. Furthermore, methods such as oversampling, under sampling or weighted loss functions can be preferred to avoid performance deviations due to unbalanced distribution of data [36,54]. Thus, the LLMPAA may exhibit higher agreement, consistency, and reliability across different raters. In addition to these, Song and colleagues [47] suggested that various methods can increase consistency. Techniques such as fine-tuning and prompt engineering were found to make models more consistent and predictable. A prompt is a specific instruction or query provided to direct the behavior of the language model and produce the desired outputs [75]. Prompts are important for controlling an LLM's output, so they should be crafted carefully [76]. Some studies found that providing prompts increased model consistency. For example, Li and Liu found that GPT-4, which was guided using prompts, showed more consistent results than models without prompts [33].

Existing research has shown that LLMs can accelerate the assessment process [46], reduce the workload of educators [45], and even cut costs [20]. However, these benefits must be weighed against the significant computational and data resources required. Additionally, LLMs can achieve high accuracy in evaluation and provide consistent results [1,19]. It is also stated that it has the potential to provide personalized [41], fast, and detailed [25] feedback [28] as well as large-scale evaluation in a short time. Bui and Barot note that such feedback can support the writing process. However, interpreting LLM performance solely by overall agreement can be misleading [17]. For example, one study reported low alignment with human raters, and other research found systematic scoring differences (e.g., GPT-4 giving lower scores and other GPT models giving higher grades than humans) [24]. Thus, LLMPAA results cannot be assumed to always match human scoring, and the potential to reduce workload should be viewed with caution. These mixed findings highlight the need for continued human oversight when using LLMs for assessment.

A variety of metrics have been used to evaluate model performance, assess agreement between LLMs and human raters, and measure consistency and reliability among different evaluators (both human and machine). On the other hand, it is stated that some of these metrics, such as ACC, should be evaluated together with others like F1-score, precision, recall, r, and QWK, as ACC alone is not sufficient and can be misleading in some datasets [77]. Although ACC is commonly used to assess LLMPAA performance, it often proves insufficient when applied in isolation. In datasets where one class disproportionately dominates, a model may achieve high accuracy simply by predicting the majority class, while failing to detect less frequent but important cases, thus misrepresenting its actual performance [77]. Therefore, incorporating complementary metrics such as precision, recall, and F1-score is essential. Precision captures the ability to avoid false positives, recall reflects the model's capacity to identify all relevant instances, and F1-score provides a harmonic mean between the two [78]. In educational assessment—particularly when evaluating partial understanding or nuanced responses—these metrics provide a more comprehensive evaluation of model performance than ACC alone. Additionally, QWK has become one of the most widely adopted metrics in LLMPAA research because it accounts for the ordinal nature of scoring and proportionally penalizes disagreements, offering a robust measure of alignment between human and machine scores [20,79]. For instance, Clauser et al. reported a QWK score of 0.99 when evaluating short-answer medical exam responses, demonstrating that LLMPAA can achieve near-human accuracy even in specialized domains [20]. While some studies reported low levels of agreement (0.30–0.50), the majority demonstrated moderate-to-excellent alignment between LLMPAA and human raters. Therefore, these results suggest that LLMs largely align with human ratings and can improve rater consistency. These findings support the potential of LLMs to automate assessment processes and reduce educators' workload. In addition, Grévisse argues that LLMs perform similarly to human raters and can be used without requiring technical expertise [23]. Urrutia and Araya [2] show that LLMs perform strongly even with a small number of training examples. Similarly, Xue and colleagues show that their proposed model achieves effective results with less labeled data [55]. In summary, the fact that LLMs provide consistent results with human raters reveals the potential of LLMs as an alternative to human evaluation. However, despite the consistent results, different problems that arise are also noteworthy. For example, Quah and colleagues show that GPT-4 correlates strongly with human ratings and provides high reliability, but they also note that GPT-4 gives lower scores than human ratings and fails to penalize inaccurate or irrelevant content [41]. Similarly, it has been determined that ChatGPT tends to give higher grades even though it shows high consistency compared to humans [25,28]. Grevisse [23], in his study comparing GPT-4 and Gemini with human raters, found that GPT-4 gave lower

grades than [38] humans, while there was no significant difference between Gemini and human raters. Likewise, Jackaria and colleagues concluded that the scoring of LLMs is not linearly related to the scoring of human raters [26]. Pack and colleagues state that LLMs make errors during grading and give different outputs for the same input [40]. Arici and colleagues found that in some cases the model had difficulty in accurately evaluating the answers to open-ended questions [13]. In addition, Mizumoto and Eguchi found that there are some variations in the grading process because LLMs are not fully compatible with human ratings [38]. In addition, it is stated that the accuracy rates of LLMs may not be sufficient in high-risk exams [36]; reliability issues arise in the evaluation of symbolic and numerical calculations [28]; the consistency of LLM evaluations varies depending on the quality of the key [23]; and due to the probabilistic nature of LLMs, different results can be obtained in different runs [29]. Therefore, it is suggested that these systems require further development and fine-tuning to increase the level of consistency [39].

Based on the reviewed studies, various issues have been identified regarding the use of LLMPAA in the evaluation process. One of these is the risk of reflecting biases. Evaluations conducted in different languages and cultural contexts have revealed that the model may behave in a biased manner [25,39], which can lead to injustices in the evaluation processes. In this context, the diversity and representation of training data is a critical factor in enhancing the fairness of the model [18,20]. Furthermore, the necessity for consistent evaluations based on the same criteria across different periods and texts necessitates the transparency and explainability of the model's decision-making processes. Another issue is hallucination. LLMPAA stands out for achieving results close to human performance in evaluation tasks [18,20]. However, the question of whether it can fully replicate human thought and creativity remains a topic of debate [80]. In this context, the necessity of human oversight in LLMPAA is emphasized [32].

Privacy and security are important issues to be considered in LLMPAA applications [68]. Protecting the privacy of the datasets used and the evaluation results obtained, as well as ensuring security against malicious use, is critical for the sustainability of the application. First, it is crucial to use encryption methods to store and transmit data securely [68]. Furthermore, strong authentication and authorization mechanisms should be implemented to prevent unauthorized access. In addition, techniques such as differential privacy and federated learning can be used to protect the model against adversarial attacks and data leakage. Anonymization methods should also be applied to protect user data. Furthermore, the continuous monitoring of the system and regular assessment of vulnerabilities are important to prevent malicious use. Finally, setting data protection policies in line with ethical rules and regulatory frameworks and adoption of these policies by all stakeholders will contribute to the sustainable provision of security and privacy in LLMPAA applications.

In addition to the technical and practical challenges discussed, the ethical dimension of LLMPAA warrants closer attention. While issues such as bias, transparency, and privacy are addressed throughout this review, aligning these with globally recognized ethical frameworks is crucial for responsible deployment. For example, the OECD Principles on Artificial Intelligence emphasize values such as fairness, transparency, and accountability in AI systems [81]. Similarly, the EU AI Act classifies AI applications in education as "high-risk", requiring rigorous ethical oversight and governance [43]. Furthermore, the IEEE Ethically Aligned Design framework advocates for human-centered values, explainability, and inclusive design [82]. Incorporating such principles into LLMPAA development ensures that these technologies not only perform effectively, but also uphold fundamental rights, protect vulnerable groups, and foster trust in educational contexts [11,67]. Therefore, future LLMPAA research and implementation should systematically integrate ethical

assessments and reference these frameworks to guide design, deployment, and evaluation practices [47,68].

Finally, LLMs may tend to perform according to the datasets on which they were trained. This can result in a decline in their overall performance across different contexts or tasks [8]. Analyses have revealed limitations in these studies, such as variability in the reliability of LLMs depending on context and their occasional ability to make erroneous evaluations [40]. Additionally, there is a necessity for retraining the model to address new questions [14]. To ensure that LLMs demonstrate more reliable and consistent performance in different contexts, various strategies can be developed. First, using diverse and balanced datasets during the model's training process can enhance its generalization ability by preventing overfitting to a specific context [83]. Additionally, techniques such as transfer learning and fine-tuning can optimize the model for new tasks or contexts [84]. Furthermore, methods like reinforcement learning with human feedback (RLHF) can be employed to improve the accuracy of LLM outputs [85]. Producing a confidence score based on uncertainty measurements and seeking human oversight when necessary can also be an approach to enhance reliability by reducing erroneous evaluations in different contexts [86]. Lastly, modular architectures or hybrid models can be used to improve the overall performance of LLMs. For instance, integrating sub-models specialized in specific contexts within a broader system can facilitate the model's adaptation to various writing tasks [73,87]. This way, both the ability to generalize can be increased and the risks of erroneous evaluations minimized.

5. Conclusions

This study demonstrates that LLMPAA can significantly reduce educator workload and accelerate assessment processes. Their ability to analyze large volumes of text in a short period enables real-time monitoring of student progress and facilitates fast, personalized feedback, offering a major advantage for adaptive learning environments [18,20,39,88]. However, claims regarding cost reduction must be evaluated cautiously, as deploying and fine-tuning these systems requires substantial investments in computing infrastructure and technical expertise. Despite these benefits, several critical limitations persist. Most notably, algorithmic biases continue to threaten fair assessment, which can disadvantage students from diverse linguistic and cultural backgrounds [10]. Issues such as performance inconsistency, limited explainability, and hallucinated or incorrect feedback weaken the reliability of LLMPAA, particularly in high-stakes or creative tasks. Furthermore, the predominance of English-language datasets in existing research limits the generalizability of these systems across multilingual and culturally varied contexts. Unresolved issues include the impact of LLMPAA on student motivation, creativity, and perceptions of fairness in evaluation—areas that remain underexplored in empirical research [59]. In addition, privacy and data security concerns persist, especially when commercial LLMs process student data on external servers. Future research should address these limitations through cross-cultural validation studies, investigate the pedagogical and emotional effects of automated feedback, and establish robust governance models for ethical implementation [11]. In addition to these challenges, the rapidly evolving landscape of LLMs presents a temporal limitation for current findings. Since the data collection phase of this review, newer and more advanced models—such as Grok 3 and DeepSeek-V3—have emerged. Grok 3, developed by xAI, was trained on the Colossus supercluster with ten times the computation of previous models, demonstrating significant improvements in reasoning, mathematics, coding, and instruction-following tasks [89]. DeepSeek-V3, on the other hand, is a mixture-of-experts model with 671 billion total parameters, utilizing innovative architectural features like multi-head latent attention and a multi-token prediction training objective, leading to

enhanced performance in multilingual tasks, particularly in languages such as Chinese [90]. The emergence of such models suggests that some of the models analyzed in this review may already be outpaced. Therefore, continuous re-evaluation and benchmarking against the most current LLMs are essential to maintain the validity and applicability of LLMPAA. Future research should integrate these recent models and assess their effectiveness across a broader range of linguistic and cultural contexts to ensure sustained relevance and fairness. The adoption of hybrid approaches—such as RLHF—and the development of interpretable model designs may offer promising pathways to preserve the scalability of LLMPAA while mitigating associated risks.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AES	Automated essay scoring
ASAG	Automated short answer grading
BERT	Bidirectional encoder representations from transformers
ICC	Intraclass correlation coefficient
LLMPAA	Large Language Model-Powered Automated Assessment
LWK	Linear weighted kappa
ML	Machine learning
NLP	Natural language processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RLHF	Reinforcement learning with human feedback
QWK	Quadratic weighted kappa

References

1. Ikiss, S.; Daoudi, N.; Abourezq, M.; Bellafkih, M. Improving Automatic Short Answer Scoring Task Through a Hybrid Deep Learning Framework. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 1066–1073. [CrossRef]
2. Urrutia, F.; Araya, R. Who's the Best Detective? Large Language Models vs. Traditional Machine Learning in Detecting Incoherent Fourth Grade Math Answers. *J. Educ. Comput. Res.* **2024**, *61*, 187–218. [CrossRef]
3. Süzen, N.; Gorban, A.N.; Levesley, J.; Mirkes, E.M. Automatic Short Answer Grading and Feedback Using Text Mining Methods. *Procedia Comput. Sci.* **2020**, *169*, 726–743. [CrossRef]
4. Yan, Y.; Liu, H. Ethical Framework for AI Education Based on Large Language Models. *Educ. Inf. Technol.* **2024**, 1–19. [CrossRef]
5. Dikli, S.; Russell, M. An Overview of Automated Scoring of Essays. *J. Technol. Learn. Assess.* **2006**, *5*, 1–35.
6. Firoozi, T.; Mohammadi, H.; Gierl, M.J. Using Active Learning Methods to Strategically Select Essays for Automated Scoring. *Educ. Meas. Issues Pract.* **2023**, *42*, 34–43. [CrossRef]
7. Burrows, S.; Gurevych, I.; Stein, B. The Eras and Trends of Automatic Short Answer Grading. *Int. J. Artif. Intell. Educ.* **2015**, *25*, 60–117. [CrossRef]
8. Henkel, O.; Hills, L.; Roberts, B.; Mcgrane, J. Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset. *Int. J. Artif. Intell. Educ.* **2024**, 1–26. [CrossRef]
9. Moher, D.; Shamseer, L.; Clarke, M.; Ghersi, D.; Liberati, A.; Petticrew, M.; Shekelle, P.; Stewart, L.A.; Estarli, M.; Barrera, E.S.A.; et al. Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement. *Rev. Esp. Nutr. Humana Diet.* **2016**, *20*, 148–160. [CrossRef]
10. Mendonça, P.C.; Quintal, F.; Mendonça, F. Evaluating LLMs for Automated Scoring in Formative Assessments. *Appl. Sci.* **2025**, *15*, 2787. [CrossRef]

11. Thurzo, A. Provable AI Ethics and Explainability in Medical and Educational AI Agents: Trustworthy Ethical Firewall. *Electronics* **2025**, *14*, 1294. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
13. Arici, N.; Gerevini, A.E.; Olivato, M.; Putelli, L.; Sigalini, L.; Serina, I. Real-World Implementation and Integration of an Automatic Scoring System for Workplace Safety Courses in Italian. *Future Internet* **2023**, *15*, 268. [CrossRef]
14. Azhari, A.; Santoso, A.; Agung, A.; Ratna, P.; Prestiliano, J. Optimization of AES Using BERT and BiLSTM for Grading the Online Exams. *Int. J. Intell. Eng. Syst.* **2024**, *17*, 395–411. [CrossRef]
15. Beseiso, M.; Alzahrani, S. An Empirical Analysis of BERT Embedding for Automated Essay Scoring. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 204–210. [CrossRef]
16. Bonthu, S.; Sree, S.R.; Prasad, M.H.M.K. SPRAG: Building and Benchmarking a Short Programming-Related Answer Grading Dataset. *Int. J. Data Sci. Anal.* **2024**, 1–13. [CrossRef]
17. Bui, N.M.; Barrot, J.S. ChatGPT as an Automated Essay Scoring Tool in the Writing Classrooms: How It Compares with Human Scoring. *Educ. Inf. Technol.* **2024**, *30*, 2041–2058. [CrossRef]
18. Chassab, R.H.; Zakaria, L.Q.; Tiun, S. An Optimized BERT Fine-Tuned Model Using an Artificial Bee Colony Algorithm for Automatic Essay Score Prediction. *PeerJ Comput. Sci.* **2024**, *10*, e2191. [CrossRef]
19. Cho, M.; Huang, J.X.; Kwon, O.W. Dual-Scale BERT Using Multi-Trait Representations for Holistic and Trait-Specific Essay Grading. *ETRI J.* **2024**, *46*, 82–95. [CrossRef]
20. Clauser, B.E.; Yaneva, V.; Baldwin, P.; An Ha, L.; Mee, J. Automated Scoring of Short-Answer Questions: A Progress Report. *Appl. Meas. Educ.* **2024**, *37*, 209–224. [CrossRef]
21. Fernández, A.A.; López-Torres, M.; Fernández, J.J.; Vázquez-García, D. ChatGPT as an Instructor's Assistant for Generating and Scoring Exams. *J. Chem. Educ.* **2024**, *101*, 3788. [CrossRef]
22. Firoozi, T.; Mohammadi, H.; Gierl, M.J. Using Automated Procedures to Score Educational Essays Written in Three Languages. *J. Educ. Meas.* **2025**, *62*, 33–56. [CrossRef]
23. Grévisse, C. LLM-Based Automatic Short Answer Grading in Undergraduate Medical Education. *BMC Med. Educ.* **2024**, *24*, 1060. [CrossRef]
24. Huang, F.; Sun, X.; Mei, A.; Wang, Y.; Ding, H.; Zhu, T. LLM Plus Machine Learning Outperform Expert Rating to Predict Life Satisfaction from Self-Statement Text. *IEEE Trans. Comput. Soc. Syst.* **2024**. [CrossRef]
25. Ivanovic, I. Can AI-Assisted Essay Assessment Support Teachers? A Cross-Sectional Mixed-Methods Research Conducted at the University of Montenegro. *Ann. Istrian Mediterr. Stud.* **2023**, *33*, 571–590.
26. Jackaria, P.M.; Hajan, B.H.; Mastul, A.R.H. A Comparative Analysis of the Rating of College Students' Essays by ChatGPT versus Human Raters. *Int. J. Learn. Teach. Educ. Res.* **2024**, *23*, 478–492. [CrossRef]
27. Kaya, M.; Cicekli, I. A Hybrid Approach for Automated Short Answer Grading. *IEEE Access* **2024**, *12*, 96332–96341. [CrossRef]
28. Kortemeyer, G. Toward AI Grading of Student Problem Solutions in Introductory Physics: A Feasibility Study. *Phys. Rev. Phys. Educ. Res.* **2023**, *19*, 020163. [CrossRef]
29. Kortemeyer, G. Performance of the Pre-Trained Large Language Model GPT-4 on Automated Short Answer Grading. *Discov. Artif. Intell.* **2024**, *4*, 47. [CrossRef]
30. Kusumaningrum, R.; Kadarisman, K.; Endah, S.N.; Sasongko, P.S.; Khadijah, K.; Sutikno, S.; Rismiyati, R.; Afriani, A. Automated Essay Scoring Using Convolutional Neural Network Long Short-Term Memory with Mean of Question-Answer Encoding. *ICIC Express Lett.* **2024**, *18*, 785–792. [CrossRef]
31. Lee, J.H.; Park, J.S.; Shon, J.G. A BERT-Based Automatic Scoring Model of Korean Language Learners' Essay. *J. Inf. Process. Syst.* **2022**, *18*, 282–291. [CrossRef]
32. Li, F.; Xi, X.; Cui, Z.; Li, D.; Zeng, W. Automatic Essay Scoring Method Based on Multi-Scale Features. *Appl. Sci.* **2023**, *13*, 6775. [CrossRef]
33. Li, W.; Liu, H. Applying Large Language Models for Automated Essay Scoring for Non-Native Japanese. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1–15. [CrossRef]
34. Li, Z.; Huang, Q.; Liu, J. ChatGPT Analysis of Strengths and Weaknesses in English Writing and Their Implications. *Appl. Math. Nonlinear Sci.* **2024**, *9*, 1–15. [CrossRef]
35. Machhout, R.A.; Ben Othmane Zribi, C. Enhanced BERT Approach to Score Arabic Essay's Relevance to the Prompt. *Commun. IBIMA* **2024**, *2024*, 176992. [CrossRef]

36. Mardini, G.I.D.; Quintero, M.C.G.; Viloria, N.C.A.; Percybrooks, B.W.S.; Robles, N.H.S.; Villalba, R.K. A Deep-Learning-Based Grading System (ASAG) for Reading Comprehension Assessment by Using Aphorisms as Open-Answer-Questions. *Educ. Inf. Technol.* **2024**, *29*, 4565–4590. [CrossRef]
37. Meccawy, M.; Bayazed, A.A.; Al-Abdullah, B.; Algamdi, H. Automatic Essay Scoring for Arabic Short Answer Questions Using Text Mining Techniques. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 768–775. [CrossRef]
38. Mizumoto, A.; Eguchi, M. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. [CrossRef]
39. Organisciak, P.; Acar, S.; Dumas, D.; Berthiaume, K. Beyond Semantic Distance: Automated Scoring of Divergent Thinking Greatly Improves with Large Language Models. *Think. Ski. Creat.* **2023**, *49*, 101356. [CrossRef]
40. Pack, A.; Barrett, A.; Escalante, J. Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100234. [CrossRef]
41. Quah, B.; Zheng, L.; Sng, T.J.H.; Yong, C.W.; Islam, I. Reliability of ChatGPT in Automated Essay Scoring for Dental Undergraduate Examinations. *BMC Med. Educ.* **2024**, *24*, 962. [CrossRef]
42. Ramesh, D.; Sanampudi, S.K. A Multitask Learning System for Trait-Based Automated Short Answer Scoring. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 454–460. [CrossRef]
43. Ramesh, D.; Sanampudi, S.K. Coherence-Based Automatic Short Answer Scoring Using Sentence Embedding. *Eur. J. Educ.* **2024**, *59*, e12684. [CrossRef]
44. Salim, H.R.; De, C.; Pratamaputra, N.D.; Suhartono, D. Indonesian Automatic Short Answer Grading System. *Bull. Electr. Eng. Inform.* **2022**, *11*, 1586–1603. [CrossRef]
45. Shamim, M.S.; Zaidi, S.J.A.; Rehman, A. The Revival of Essay-Type Questions in Medical Education: Harnessing Artificial Intelligence and Machine Learning. *J. Coll. Physicians Surg. Pak.* **2024**, *34*, 595–599. [CrossRef] [PubMed]
46. Sharma, A.; Jayagopi, D.B. Modeling Essay Grading with Pre-Trained BERT Features. *Appl. Intell.* **2024**, *54*, 4979–4993. [CrossRef]
47. Song, Y.; Zhu, Q.; Wang, H.; Zheng, Q. Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE Trans. Learn. Technol.* **2024**, *17*, 1920–1930. [CrossRef]
48. Tang, X.; Chen, H.; Lin, D.; Li, K. Harnessing LLMs for Multi-Dimensional Writing Assessment: Reliability and Alignment with Human Judgments. *Heliyon* **2024**, *10*, e34262. [CrossRef]
49. Tate, T.P.; Steiss, J.; Bailey, D.; Graham, S.; Moon, Y.; Ritchie, D.; Tseng, W.; Warschauer, M. Can AI Provide Useful Holistic Essay Scoring? *Comput. Educ. Artif. Intell.* **2024**, *7*, 100255. [CrossRef]
50. Wang, Q.; Gayed, J.M. Effectiveness of Large Language Models in Automated Evaluation of Argumentative Essays: Finetuning vs. Zero-Shot Prompting. *Comput. Assist. Lang. Learn.* **2024**. [CrossRef]
51. Wijanto, M.C.; Yong, H.S. Combining Balancing Dataset and SentenceTransformers to Improve Short Answer Grading Performance. *Appl. Sci.* **2024**, *14*, 4532. [CrossRef]
52. Wijaya, M.C. Automatic Short Answer Grading System in Indonesian Language Using BERT Machine Learning. *Rev. D'intelligence Artif.* **2021**, *35*, 503–509. [CrossRef]
53. Wu, Y.; Henriksson, A.; Nouri, J.; Duneld, M.; Li, X. Beyond Benchmarks: Spotting Key Topical Sentences While Improving Automated Essay Scoring Performance with Topic-Aware BERT. *Electronics* **2022**, *12*, 150. [CrossRef]
54. Zhu, X.; Wu, H.; Zhang, L. Automatic Short-Answer Grading via BERT-Based Deep Neural Networks. *IEEE Trans. Learn. Technol.* **2022**, *15*, 364–375. [CrossRef]
55. Xue, J.; Tang, X.; Zheng, L. A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring. *IEEE Access* **2021**, *9*, 125403–125415. [CrossRef]
56. Yamashita, T. An Application of Many-Facet Rasch Measurement to Evaluate Automated Essay Scoring: A Case of ChatGPT-4.0. *Res. Methods Appl. Linguist.* **2024**, *3*, 100133. [CrossRef]
57. Yavuz, F.; Çelik, Ö.; Yavaş Çelik, G. Utilizing Large Language Models for EFL Essay Grading: An Examination of Reliability and Validity in Rubric-Based Assessments. *Br. J. Educ. Technol.* **2024**, *56*, 150–166. [CrossRef]
58. Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; Zou, J. GPT Detectors Are Biased against Non-Native English Writers. *Patterns* **2023**, *4*, 100799. [CrossRef]
59. Lin, H.; Chen, Q. Artificial Intelligence (AI) -Integrated Educational Applications and College Students' Creativity and Academic Emotions: Students and Teachers' Perceptions and Attitudes. *BMC Psychol.* **2024**, *12*, 487. [CrossRef]
60. Hackl, V.; Müller, A.E.; Granitzer, M.; Sailer, M. Is GPT-4 a Reliable Rater? Evaluating Consistency in GPT-4's Text Ratings. *Front. Educ.* **2023**, *8*, 1272229. [CrossRef]
61. Kosinski, M. Evaluating Large Language Models in Theory of Mind Tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *121*, e2405460121. [CrossRef]
62. Binz, M.; Schulz, E. Using Cognitive Psychology to Understand GPT-3. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2218523120. [CrossRef]

63. Masters, K. Medical Teacher's First ChatGPT's Referencing Hallucinations: Lessons for Editors, Reviewers, and Teachers. *Med. Teach.* **2023**, *45*, 673–675. [CrossRef] [PubMed]
64. Sallam, M.; Salim, N.A.; Barakat, M.; Al-Tammemi, A.B. ChatGPT Applications in Medical, Dental, Pharmacy, and Public Health Education: A Descriptive Study Highlighting the Advantages and Limitations. *Narra J.* **2023**, *3*, e103. [CrossRef] [PubMed]
65. Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.A.; Sheikh, J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef] [PubMed]
66. Deng, J.; Lin, Y. The Benefits and Challenges of ChatGPT: An Overview. *Front. Comput. Intell. Syst.* **2022**, *2*, 81–83. [CrossRef]
67. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
68. Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; Zhang, Y. A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. *High-Confid. Comput.* **2024**, *4*, 100211. [CrossRef]
69. Gao, R.; Merzdorf, H.E.; Anwar, S.; Hipwell, M.C.; Srinivasa, A.R. Automatic Assessment of Text-Based Responses in Post-Secondary Education: A Systematic Review. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100206. [CrossRef]
70. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process Syst.* **2020**, *33*, 1877–1901.
71. Polverini, G.; Gregorcic, B. Performance of ChatGPT on the Test of Understanding Graphs in Kinematics. *Phys. Rev. Phys. Educ. Res.* **2024**, *20*, 010109. [CrossRef]
72. Kaufmann, T.; Weng, P.; Kunshan, D.; Bengs, V.; Hüllermeier, E. A Survey of Reinforcement Learning from Human Feedback. *arXiv* **2023**, arXiv:2312.14925.
73. Atkinson, J.; Palma, D. An LLM-Based Hybrid Approach for Enhanced Automated Essay Scoring. *Sci. Rep.* **2025**, *15*, 14551. [CrossRef] [PubMed]
74. Introducing Claude 3.5 Sonnet\Anthropic. Available online: https://www.anthropic.com/news/claude-3-5-sonnet (accessed on 8 May 2025).
75. Giray, L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Ann. Biomed. Eng.* **2023**, *51*, 2629–2633. [CrossRef] [PubMed]
76. Nazir, A.; Wang, Z. A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. *Meta Radiol.* **2023**, *1*, 100022. [CrossRef]
77. Provost, F.; Fawcett, T. Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Big Data* **2013**, *1*, 51–59. [CrossRef]
78. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *arXiv* **2020**, arXiv:2010.16061.
79. Cohen, J. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef]
80. Mohammad, A.F.; Clark, B.; Agarwal, R.; Summers, S. LLM/GPT Generative AI and Artificial General Intelligence (AGI): The Next Frontier. In Proceedings of the 2023 Congress in Computer Science, Computer Engineering, and Applied Computing, CSCE 2023, Las Vegas, NV, USA,, 24–27 July 2023; pp. 413–417. [CrossRef]
81. AI Principles | OECD. Available online: https://www.oecd.org/en/topics/ai-principles.html (accessed on 7 May 2025).
82. Shahriari, K.; Shahriari, M. IEEE Standard Review—Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. In Proceedings of the IHTC 2017—IEEE Canada International Humanitarian Technology Conference 2017, Toronto, ON, Canada, 21–22 July 2017; pp. 197–201. [CrossRef]
83. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1.
84. Howard, J.; Ruder, S. Universal Language Model Fine-Tuning for Text Classification. In Proceedings of the ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 328–339. [CrossRef]
85. Christiano, P.F.; Leike, J.; Brown, T.B.; Martic, M.; Legg, S.; Amodei, D. Deep Reinforcement Learning from Human Preferences. *Adv. Neural Inf. Process Syst.* **2017**, *2017*, 4300–4308.
86. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York, NY, USA, 19–24 June 2016; Volume 3, pp. 1651–1660.
87. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2019**, *21*, 1–67.
88. Goslen, A.; Kim, Y.J.; Rowe, J.; Lester, J. LLM-Based Student Plan Generation for Adaptive Scaffolding in Game-Based Learning Environments. *Int. J. Artif. Intell. Educ.* **2024**, 1–26. [CrossRef]

89. Grok 3 Beta—The Age of Reasoning Agents | XAI. Available online: https://x.ai/news/grok-3 (accessed on 8 May 2025).
90. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. Deepseek-v3 technical report. *arXiv* **2024**, arXiv:2412.19437.