Automated Scoring in Educational Assessment: A Survey of Short Answer Questions, Essays, and NLP Technologies

www.surveyx.cn

Abstract

In the evolving landscape of educational assessment, Automated Essay Scoring (AES) systems, powered by Natural Language Processing (NLP), have emerged as transformative tools, enabling efficient, objective, and scalable evaluation of student writing. This survey paper examines the integration of AES systems within modern educational frameworks, highlighting their significance in enhancing educational outcomes. The paper explores the development of AES systems, emphasizing the role of advanced machine learning models such as transformers and Large Language Models (LLMs) in improving scoring accuracy and feedback mechanisms. The survey also addresses the transition to computer-based testing environments, facilitated by AI technologies, which offers enhanced teaching effectiveness and scalable assessment solutions. Despite the advancements, challenges persist, including ethical concerns, biases in AI-driven evaluations, and the need for robust data and model frameworks. The paper advocates for continued research to refine AES technologies, ensuring their alignment with educational objectives and societal values. Future directions include integrating AI technologies for comprehensive assessment, enhancing data synthesis techniques, and developing ethical guidelines for AI use in education. By addressing these challenges, AES systems can significantly improve the quality and equity of educational assessments, fostering a more integrated and adaptive learning environment.

1 Introduction

1.1 Significance of Automated Scoring in Modern Education

Automated scoring systems have become essential tools in the contemporary educational landscape, offering significant advantages in scalability, efficiency, and objectivity. The shift towards digital learning, accelerated by challenges such as the COVID-19 pandemic, has increased the demand for effective assessment solutions [1]. Automated Essay Scoring (AES) systems leverage advanced algorithms and artificial intelligence to evaluate complex writing attributes, including argumentation and coherence, which are crucial for developing critical thinking skills.

The incorporation of Natural Language Processing (NLP) technologies into AES systems has markedly improved their ability to analyze linguistic features and provide consistent feedback, thus facilitating skill development in language learning and proficiency assessments [2]. This enhancement is particularly beneficial for non-native speakers, allowing for personalized learning pathways and demonstrating the educational value of technology in fostering adaptive learning environments [3].

Moreover, the introduction of AI-driven technologies in education necessitates innovative assessment methods capable of managing ambiguity and integrating domain-specific knowledge, enriching the overall educational experience [4]. Nevertheless, reliance on large language models (LLMs) for

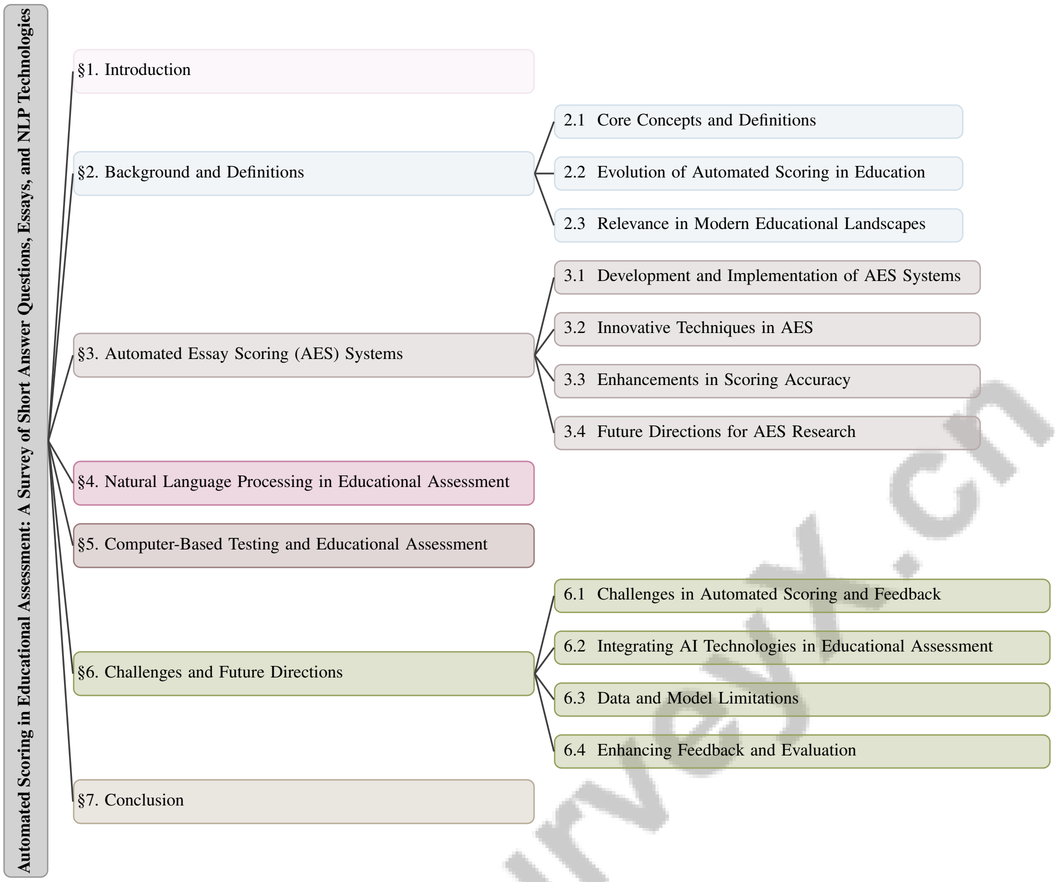

Automated Scoring in Educational Assessment: A Survey of Short Answer Questions, Essays, and NLP Technologies
§1. Introduction
§2. Background and Definitions
2.1 Core Concepts and Definitions
2.2 Evolution of Automated Scoring in Education
2.3 Relevance in Modern Educational Landscapes
§3. Automated Essay Scoring (AES) Systems
3.1 Development and Implementation of AES Systems
3.2 Innovative Techniques in AES
3.3 Enhancements in Scoring Accuracy
3.4 Future Directions for AES Research
§4. Natural Language Processing in Educational Assessment
§5. Computer-Based Testing and Educational Assessment
§6. Challenges and Future Directions
6.1 Challenges in Automated Scoring and Feedback
6.2 Integrating AI Technologies in Educational Assessment
6.3 Data and Model Limitations
6.4 Enhancing Feedback and Evaluation
§7. Conclusion

Figure 1: chapter structure

automated scoring presents challenges such as high computational costs and privacy concerns [5]. Addressing these issues is vital for the ethical implementation of automated scoring systems and enhancing AI literacy among students [6].

The survey emphasizes the importance of aligning evaluation methods with user needs and project goals, underscoring the relevance of automated scoring in modern educational assessments [7]. Ongoing research and development are crucial for refining these technologies to ensure their alignment with educational objectives and societal values [8]. The continuous evolution of these systems promises to further enhance their effectiveness and integration into education.

1.2 Role of Computer-Based Testing

The transition to computer-based testing represents a transformative shift in educational assessment, driven by advancements in artificial intelligence and machine learning. This change has been supported by the integration of AI technologies, which enhance teaching effectiveness and establish digital assessment formats as standard practice in contemporary education [9]. The growing reliance on AI tools, including Generative AI, in academic environments highlights the importance of equipping students with the skills necessary for effective interaction with these technologies [10].

Machine learning models, such as the fine-tuned Roberta Large Model, illustrate the potential of AI-driven systems to grade short answer questions with improved accuracy and consistency, addressing the biases and inconsistencies often associated with human grading. These advancements respond to the limitations of traditional scoring methods, which are frequently time-consuming and subjective, emphasizing the need for more efficient automated systems [11].

The integration of AI chatbots and language models, including ChatGPT, into student assessment practices has created new opportunities for interaction and feedback, particularly in essay writing. However, while LLMs can provide feedback, they often fall short in generating constructive criticism, necessitating the development of benchmarks for effectively evaluating LLM performance in educational contexts [12]. Consequently, the responsibility for delivering actionable feedback on essay argument strength remains with educators, underscoring the complementary role of AI in supporting, rather than replacing, human judgment [13].

1.3 Structure of the Survey

This survey provides a comprehensive examination of automated scoring systems in educational assessments, with a focus on the integration of Natural Language Processing (NLP) technologies and the evolution of computer-based testing environments. The paper is organized into several key sections, each addressing specific aspects of the topic.

Initially, the survey introduces automated scoring, highlighting its significance in modern education and the transformative impact of computer-based testing. This is followed by a background section defining core concepts such as short answer questions, essays, automated scoring, AES, NLP, educational assessment, and computer-based testing, while exploring the evolution and relevance of automated scoring in contemporary education.

Subsequent sections detail Automated Essay Scoring (AES) systems, discussing their development, implementation, and innovative techniques, including transformer models and pre-trained language models like BERT and GPT. The role of NLP in enhancing AES systems is examined, particularly its application in linguistic feature extraction and coherence assessment.

The survey also explores the transition to computer-based testing, detailing advantages such as increased accessibility and efficiency, alongside challenges like technical issues and the need for robust assessment frameworks. It considers broader implications for educational assessment practices and the evolving role of technology in learning environments [14, 15, 1]. Discussions include scalability, reliability, ethical considerations, and accessibility issues linked to automated scoring systems.

Finally, the survey identifies current challenges and potential future research directions, emphasizing the integration of AI technologies, addressing data and model limitations, and enhancing feedback mechanisms. The conclusion highlights the critical role of automated scoring systems in contemporary educational assessments, stressing the need for ongoing research and development to align these technologies with educational goals and societal values. This includes addressing limitations in evaluating complex writing tasks, such as essays, by improving assessments of content relevance, coherence, and persuasiveness. Leveraging advancements in artificial intelligence and machine learning can lead to more reliable and efficient evaluation tools that save educators time and provide meaningful feedback to students, ultimately fostering a more effective learning environment [13, 16, 15, 17].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts and Definitions

Foundational concepts in educational assessments are essential for implementing automated scoring systems effectively. Automated Essay Scoring (AES) employs machine learning algorithms to evaluate written content based on relevance, cohesion, and coherence, streamlining grading by providing immediate feedback and enhancing automation [18, 19]. AES often uses ordinal classification to assess essays according to proficiency levels defined by the Common European Framework of Reference (CEFR) [2].

Natural Language Processing (NLP) is crucial in AES, facilitating the extraction of linguistic features to assess text coherence and logical relationships [20]. NLP technologies enable the annotation and classification of argumentative elements, improving feedback accuracy and scoring reliability [21]. These capabilities are particularly relevant for complex questions requiring reasoning and decomposition [4].

Automated scoring extends to short answer questions, categorizing responses to provide insights into student comprehension [8]. Large Language Models (LLMs) enhance these systems by addressing challenges in semantic analysis and superficial scoring [6]. Key terms such as Grammatical Error Correction (GEC) and Automated Writing Evaluation (AWE) are vital for addressing bias and adver-

AI-generated, for reference only.

sarial attacks [8]. Automated Long Answer Grading (ALAG) and rubric-based systems emphasize analytic rubrics for evaluating lab reports and extended responses [5].

Ethical considerations in automated scoring frameworks are critical, especially concerning sensitive content like references to self-harm or violence in student essays [2]. Intrinsic and extrinsic evaluation methods are defined as crucial for assessing automated scoring systems in educational contexts [7]. These terms highlight the multifaceted nature of automated scoring and the transformative impact of AI and NLP technologies on educational assessments.

2.2 Evolution of Automated Scoring in Education

Automated scoring systems in education have evolved from basic e-assessment methods to advanced technologies capable of nuanced evaluations. Early systems relied on keyword matching and handcrafted features, which often failed to capture the complexity of student writing [1]. This led to the adoption of neural network models in AES, moving beyond traditional bag-of-words approaches that struggled with linguistic nuances [11].

The scope of automated scoring now includes various student outputs, such as long-form writing and lab reports. Benchmarks like RiceChem and the SweLL benchmark evaluate complex, fact-based responses and L2 Swedish writing, respectively. Resources such as ArgRewrite V.2 advance the study of revisions in argumentative writing, enhancing automated scoring capabilities [21].

Challenges persist in detecting essays generated by LLMs, as existing detectors struggle against adversarial attacks, necessitating more robust evaluation methods [6]. The evolution of Natural Language Generation (NLG) evaluation methods highlights the broader integration of automated scoring in educational settings [7]. The emergence of Automated Writing Evaluation (AWE) and Grammatical Error Correction (GEC) as distinct entities within automated scoring has not yet led to effective integration, indicating a need for further development [8]. Ongoing refinement of these systems is essential for overcoming traditional assessment limitations and ensuring reliable AI implementation in education. As automated scoring technologies evolve, they have the potential to significantly enhance the fairness and efficacy of educational assessments across diverse learning environments.

2.3 Relevance in Modern Educational Landscapes

Automated scoring systems are increasingly relevant in contemporary educational environments, addressing the limitations of traditional assessments through advanced technologies like NLP and AI. These systems provide scalable, efficient, and objective evaluations, crucial for meeting diverse modern educational needs [1]. The integration of LLMs exemplifies sophisticated applications, enhancing NLP capabilities and offering scalable solutions for personalized tutoring while maintaining effective pedagogical strategies [22].

Benchmarks such as SweLL and ArgRewrite V.2 demonstrate the applicability of automated scoring systems. The SweLL benchmark significantly contributes to L2 research, providing accessible data that enriches understanding of second language acquisition among researchers, teachers, and students [2]. ArgRewrite V.2 enhances NLP applications, especially in intelligent tutoring systems, by improving the analysis of argumentative writing and facilitating effective educational interventions [21].

Despite advancements, challenges like ethical considerations and biases in generative AI models persist, highlighting the need for responsible technology use and interdisciplinary collaboration to ensure equitable educational outcomes. Benchmarks designed to evaluate LLMs in scoring written essays underscore the importance of comparing their effectiveness with state-of-the-art models and analyzing their feedback capabilities, vital for refining these systems to better serve educational objectives [22].

3 Automated Essay Scoring (AES) Systems

3.1 Development and Implementation of AES Systems

Advancements in Automated Essay Scoring (AES) systems are crucial for modern educational assessments, leveraging machine learning and NLP technologies. As illustrated in Figure 2, this figure depicts the hierarchical categorization of these advancements, highlighting neural architectures, evaluation methods, and scoring mechanisms. End-to-end neural architectures, like VerAs, optimize LLM performance for complex tasks such as lab report scoring through prompt engineering [23, 22]. Transformer-based models dominate AES, with T-AES using transformer architectures for text classification, including features like politeness [24]. Interpretable systems such as IDLS-RFP enhance transparency in evaluating intricate writing tasks [25].

The integration of LLMs introduces novel evaluation methodologies, including assessments without predefined rubrics and pairwise comparisons, enhancing reliability and accuracy [26]. Benchmarks like CERD enrich datasets for AES model training by delineating rhetorical tasks [27]. Automated writing evaluation systems like eRevise employ NLP for rubric-based formative feedback, fostering targeted student development [28]. Content detection pipelines, such as MTLHealth, address sensitive topics in student essays using BERT models trained on mental health tasks [29].

Mitigating lexical biases and ensuring robust scoring mechanisms are essential in AES deployment. Techniques like Gradient Boosted Trees enhance reliability [30]. The OUTFOX framework improves LLM-generated text detectors through adversarial learning, reinforcing AES systems' resilience [6].

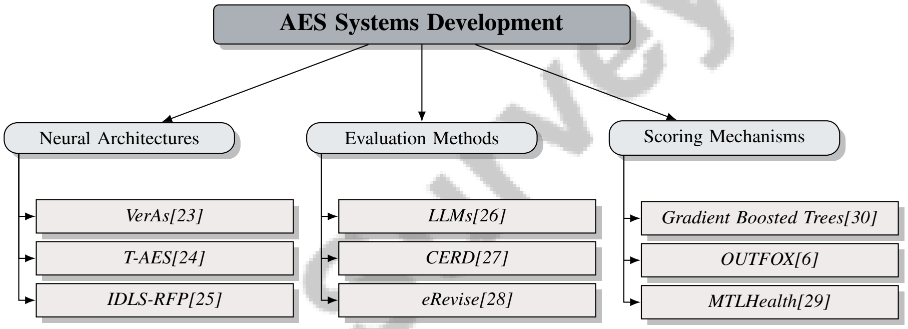


Figure 2: This figure depicts the hierarchical categorization of Automated Essay Scoring (AES) system advancements, highlighting neural architectures, evaluation methods, and scoring mechanisms.

3.2 Innovative Techniques in AES

The evolution of AES systems is propelled by advanced techniques enhancing multifaceted student output evaluation. Multi-dimensional scoring models assess vocabulary, grammar, and coherence, addressing biases across diverse demographics [31, 32, 33, 34]. Transformer models, RNNs, and LSTMs are pivotal in enhancing AES accuracy and efficiency.

Transformers have revolutionized NLP tasks, including AES, through robust text classification and feature extraction. BERT models exemplify the shift towards deep contextual embeddings for essay quality assessment [35]. The T-AES system illustrates transformers' utility in preserving word order and contextual meaning for accurate scoring [24].

RNNs and LSTMs offer powerful sequence processing capabilities for AES, accommodating text's sequential nature. The Multi-task BiLSTM model processes essays to predict grammatical errors and overall scores, highlighting these architectures' potential in multitask learning [36].

Innovative techniques include LLMs for implicit evaluations. The RMTS model integrates LLM-generated rationales with a fine-tuned scoring model, enhancing scoring by using rationales as inputs [37]. This underscores LLMs' significance in scalable feedback mechanisms.

These advancements illustrate AES's ongoing evolution, integrating state-of-the-art machine learning and NLP technologies. Innovations enhance student writing evaluation accuracy, challenging traditional paradigms and demonstrating smaller, fine-tuned transformer models' potential to outperform larger models. By leveraging transformer-based approaches accounting for long-term dependencies and word order, AES systems provide nuanced feedback and improve manual scoring efficiency, benefiting educators and learners [38, 24].

3.3 Enhancements in Scoring Accuracy

Recent AES advancements have improved scoring accuracy through data augmentation, transfer learning, and pre-trained models like BERT and GPT. Methodologies address educational assessments' intricate demands, enhancing precision and reliability. These include prompt-specific models, comprehensive datasets for diverse languages, and multi-dimensional scoring systems evaluating writing quality aspects. Robust evaluation frameworks assess AES models' fairness and generalizability, revealing insights into bias and scoring approach effectiveness, contributing to nuanced and equitable student essay assessment [31, 34, 38, 39, 33].

Data augmentation refines AES models, with techniques like Back-Translation Essay Augmentation (BTEA) enhancing training datasets by generating back-translated essays, improving robustness and generalization [11]. High inter-annotator agreement in CEFR level assignments within the SweLL dataset supports reliable learner performance assessments [2].

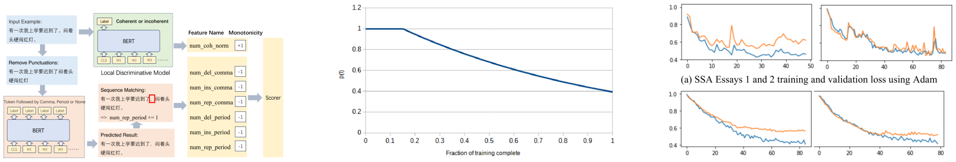
Transfer learning, particularly fine-tuning transformer models, has advanced AES capabilities. Models like Qwen1.5-7B and GPT-4 enhance scoring accuracy by understanding and generating rhetorical devices [27]. Specialized models like ELM align with academic needs, producing contextually appropriate outputs [5].

Pre-trained models like BERT and GPT enhance scoring accuracy, capturing low-level character n-grams and high-level semantic features, achieving above-human-level accuracy in essay scoring tasks [17]. Incorporating POS tags into LSA models improves grading accuracy by 5

Innovative approaches, like VerAs's dual-module architecture with a verifier and grader, significantly enhance scoring accuracy by improving relevant content identification in student submissions [23]. The ArgRewrite V.2 benchmark shows substantial improvements in model performance metrics, like F1-score and accuracy in predicting revision purposes, further underscoring AES performance advancements [21].

Challenges persist in aligning model-generated scores with human ratings. Metrics like Quadratic Weighted Kappa (QWK) evaluate agreement between model predictions and human evaluations, ensuring AES systems' reliability [22]. Systems like OUTFOX, adapting to adversarial examples, enhance detection performance against evolving evasion strategies [6]. Addressing these challenges is essential for aligning AES systems more closely with human evaluative standards, enhancing their educational utility.

The integration of data augmentation, transfer learning, and pre-trained models has substantially advanced AES systems' accuracy and efficacy. These advancements enhance AES's ability to deliver nuanced insights into student writing proficiency by incorporating diverse linguistic features, multi-dimensional scoring capabilities, and robust machine learning techniques. This evolution aligns with contemporary educational environments' dynamic requirements, enabling educators to provide timely, accurate feedback while addressing student writing complexities across contexts and demographics [31, 34, 24, 32, 33].



(a) Coherent vs. Incoherent Sentences Detection[40]

(b) The graph shows the probability of a binary event as a function of the fraction of training data completed.[41]

(c) SSA Essays 1 and 2 training and validation loss using Adam[16]

Figure 3: Examples of Enhancements in Scoring Accuracy

AI-generated, for reference only.

As shown in Figure 3, recent AES advancements have significantly improved scoring accuracy. Various methodologies and visual representations exemplify this development. One technique involves detecting coherent versus incoherent sentences using a local discriminative model, where a BERT model extracts features to predict sentence coherence, subsequently scored by a specialized scorer. Another approach examines the probability of a binary event relative to the fraction of training data completed, illustrating how probability evolves as training progresses. Additionally, the use of the Adam optimizer in training and validating SSA Essays datasets highlights the reduction in both training and validation losses over iterations. Collectively, these examples underscore the strides made in refining AES systems to achieve more accurate and reliable essay scoring [40, 41, 16].

3.4 Future Directions for AES Research

The future of AES research is set to explore innovative directions aimed at enhancing these systems' efficacy and adaptability in educational contexts. Integrating attention mechanisms into deep learning architectures is promising, significantly improving AES models' interpretative capabilities for nuanced and accurate essay evaluations [42]. This aligns with leveraging advanced neural network architectures to capture complex linguistic features, refining the scoring process.

Enhancing data synthesis techniques is another critical focus. Reasoning Distillation-Based Evaluation (RDBE) methodologies offer novel opportunities to extend AES capabilities beyond traditional essay scoring, potentially encompassing diverse evaluation tasks [43]. This expansion could facilitate developing versatile AES systems addressing diverse educational assessment needs.

The concept of computational essays, particularly within structured guidance for students, represents an intriguing direction for future AES research. By providing explicit frameworks for student writing, researchers can refine educational assessment methodologies, enhancing AES systems' pedagogical value [44]. This direction emphasizes aligning AES innovations with instructional goals, ensuring these technologies effectively support student learning and development.

These prospective research directions highlight AES technology's dynamic evolution, emphasizing integrating cutting-edge machine learning techniques and pedagogically informed frameworks. As AES systems evolve, they are poised to become integral to modern educational assessments, offering scalable, objective evaluations catering to contemporary learning environments' diverse needs. Recent studies underscore AES's potential to enhance grading efficiency and consistency, particularly in large classrooms, while addressing linguistic features influencing scoring accuracy. For instance, research on Arabic AES introduced a comprehensive dataset and employed advanced models like AraBERT, revealing promising grading accuracy and error analysis results. Additionally, investigations into AES models' fairness and generalizability highlight understanding biases affecting marginalized student groups, guiding the development of more effective and equitable assessment tools in education [31, 33, 32].

In recent years, the integration of Natural Language Processing (NLP) technologies into educational assessment has gained considerable attention. This integration not only streamlines the evaluation process but also enhances the quality of feedback provided to students. As illustrated in Figure 5, the figure illustrates the future directions for Automated Essay Scoring (AES) research, highlighting the integration of deep learning techniques, enhancement of data synthesis methods, and alignment with educational objectives. It emphasizes the potential of attention mechanisms and advanced neural architectures, the application of Reasoning Distillation-Based Evaluation (RDBE), and the role of computational essays in educational contexts. By examining these advancements, we can better understand the transformative potential of NLP in educational contexts.

4 Natural Language Processing in Educational Assessment

4.1 Role of NLP in Automated Essay Scoring

Natural Language Processing (NLP) technologies play a pivotal role in enhancing Automated Essay Scoring (AES) systems by extracting linguistic features and assessing coherence. Transformer models like BERT and XLNet have significantly increased AES precision by capturing complex contextual information and linguistic nuances [24, 45]. These models utilize attention mechanisms to improve scoring accuracy and coherence evaluation by extracting linguistic features across multiple text levels.

AI-generated, for reference only.

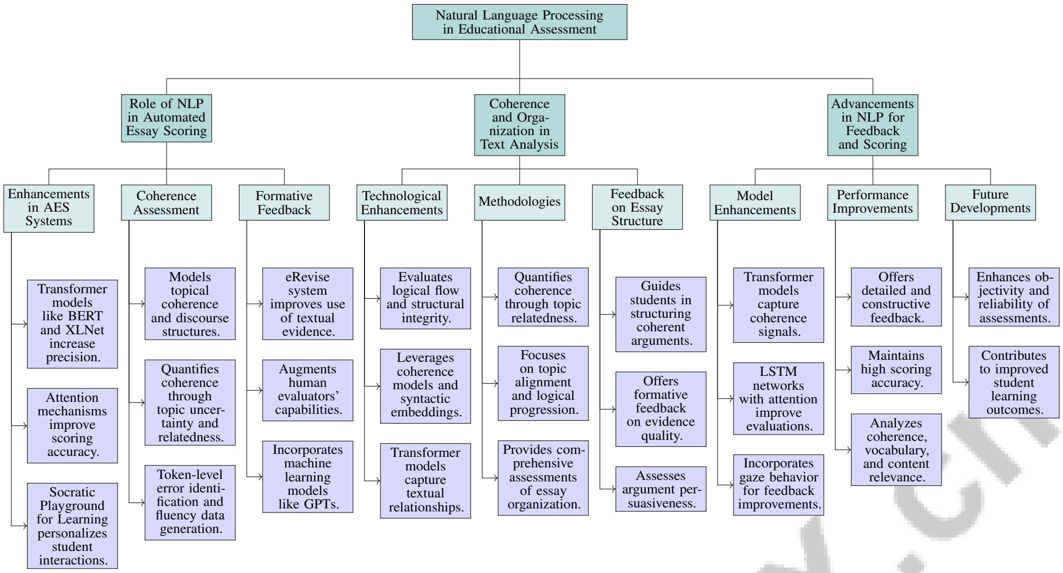

Natural Language Processing in Educational Assessment
Role of NLP in Automated Essay Scoring
Coherence and Organization in Text Analysis
Advancements in NLP for Feedback and Scoring
Enhancements in AES Systems
Coherence Assessment
Formative Feedback
Technological Enhancements
Methodologies
Feedback on Essay Structure
Model Enhancements
Performance Improvements
Future Developments
Transformer models like BERT and XLNet increase precision.
Attention mechanisms improve scoring accuracy.
Socratic Playground for Learning personalizes student interactions.
Models topical coherence and discourse structures.
Quantifies coherence through topic uncertainty and relatedness.
Token-level error identification and fluency data generation.
eRevise system improves use of textual evidence.
Augments human evaluators' capabilities.
Incorporates machine learning models like GPTs.
Evaluates logical flow and structural integrity.
Leverages coherence models and syntactic embeddings.
Transformer models capture textual relationships.
Quantifies coherence through topic relatedness.
Focuses on topic alignment and logical progression.
Provides comprehensive assessments of essay organization.
Guides students in structuring coherent arguments.
Offers formative feedback on evidence quality.
Assesses argument persuasiveness.
Transformer models capture coherence signals.
LSTM networks with attention improve evaluations.
Incorporates gaze behavior for feedback improvements.
Offers detailed and constructive feedback.
Maintains high scoring accuracy.
Analyzes coherence, vocabulary, and content relevance.
Enhances objectivity and reliability of assessments.
Contributes to improved student learning outcomes.

Figure 4: This figure illustrates the hierarchical structure of Natural Language Processing (NLP) applications in educational assessment, highlighting its role in Automated Essay Scoring (AES), coherence and organization analysis, and recent advancements for feedback and scoring. The diagram categorizes key enhancements, methodologies, and future developments in NLP technologies, emphasizing their contributions to improving educational assessments and student learning outcomes.

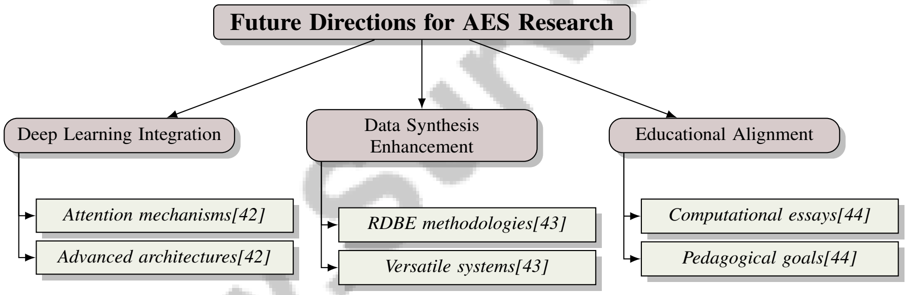


Figure 5: This figure illustrates the future directions for Automated Essay Scoring (AES) research, highlighting the integration of deep learning techniques, enhancement of data synthesis methods, and alignment with educational objectives. It emphasizes the potential of attention mechanisms and advanced neural architectures, the application of Reasoning Distillation-Based Evaluation (RDBE), and the role of computational essays in educational contexts.

Moreover, the Socratic Playground for Learning (SPL) leverages NLP to create adaptive tutoring dialogues that foster critical thinking by personalizing student interactions [46].

Coherence assessment, crucial for evaluating essays, benefits from NLP's capability to model topical coherence and discourse structures. A metric introduced by [47] quantifies coherence through topic uncertainty and relatedness within paragraphs using an unsupervised approach, complemented by models performing token-level error identification and generating fluency data through back-translation [48].

NLP technologies also enhance formative feedback in AES systems. For instance, the eRevise system employs NLP to provide feedback that improves students' use of textual evidence, aiding their writing development [28]. This highlights NLP's role in augmenting human evaluators' capabilities, streamlining feedback processes, and boosting grading efficiency.

As NLP technologies advance, they enhance the objectivity and reliability of educational assessments, incorporating machine learning models like generative pre-trained transformers (GPTs) to effectively evaluate both human and AI-generated texts. This evolution aligns with contemporary learning environments' dynamic needs, with innovative methods like the eRevise platform empowering students to refine their writing skills through targeted feedback on evidence usage. Additionally, advanced neural network techniques facilitate the extraction of topical components from source texts, enriching the assessment process and providing educators with valuable insights into student performance [28, 16, 15, 49].

4.2 Coherence and Organization in Text Analysis

Assessing coherence and organization in student essays is crucial for AES systems, significantly enhanced by NLP technologies. These technologies enable AES systems to evaluate logical flow and structural integrity of written work. By leveraging coherence models, syntactic embeddings, and topical component extraction, AES systems provide nuanced evaluations that inform users about their understanding of the material, contributing to learning analytics and guiding effective revisions [50, 49, 51].

Key advancements include transformer models that capture intricate relationships between textual elements. Vanilla, hierarchical, multi-task learning, and fact-aware transformers create a robust framework for evaluating coherence by analyzing topic expression and interrelation within text [52]. These models utilize attention mechanisms to discern essays' hierarchical structure, ensuring both local and global coherence are considered.

Methods that quantify coherence by examining topical relationships within paragraphs enhance AES capabilities. These approaches evaluate coherence by assessing topic relatedness and uncertainty, offering an unsupervised methodology for discourse structure analysis [47]. By focusing on topic alignment and logical progression, NLP technologies enable AES systems to deliver comprehensive assessments of essay organization.

Advanced NLP models in AES systems improve coherence assessment accuracy and support detailed feedback on essay structure. This feedback guides students in enhancing writing skills, particularly in structuring coherent arguments. Research indicates targeted feedback on evidence and reasoning in argumentative writing promotes effective revisions. Tools like eRevise utilize NLP to offer formative feedback, improving text evidence quality in student essays. Additionally, automated systems assess argument persuasiveness, providing actionable insights that complement traditional teacher feedback, fostering a collaborative approach to refining argumentative writing [28, 53, 13]. As NLP technologies evolve, they will further enhance AES systems' ability to assess coherence and organization, contributing to more effective educational assessments and improved student learning outcomes.

4.3 Advancements in NLP for Feedback and Scoring

Recent advancements in NLP have significantly enhanced AES systems' capacity to provide nuanced feedback and improve scoring accuracy. Transformer-based models have achieved state-of-the-art results in coherence assessment tasks, demonstrating their effectiveness in capturing coherence signals [52]. These models employ attention mechanisms to focus on salient information, enhancing long text understanding and performance across various tasks [54].

The integration of Long Short-Term Memory (LSTM) networks with attention mechanisms has further advanced feedback and scoring. These architectures enable AES systems to dynamically focus on relevant sections of student essays, leading to more precise evaluations and targeted feedback [41]. This capability ensures feedback is relevant and actionable, supporting students in refining their writing skills.

Incorporating gaze behavior into NLP tasks has shown potential for enhancing feedback mechanisms. By analyzing reader engagement with text, AES systems can mimic human evaluative processes, improving scoring accuracy and feedback quality [55]. This approach emphasizes the importance of integrating cognitive insights into NLP models to develop more effective educational assessment tools.

Advancements in NLP technologies have significantly bolstered AES systems' performance, enabling them to offer detailed and constructive feedback while maintaining high scoring accuracy. By leveraging transformer-based models capable of evaluating complex linguistic features and long-term dependencies in text, these systems can analyze writing aspects, including coherence, vocabulary, and content relevance. Consequently, AES systems now provide timely and consistent evaluations, facilitating more effective learning and assessment in educational contexts [31, 38, 24, 56, 32]. As NLP evolves, it will further enhance the objectivity and reliability of educational assessments, ultimately contributing to improved student learning outcomes.

5 Computer-Based Testing and Educational Assessment

5.1 Shift Towards Computer-Based Testing

The transition to computer-based testing marks a significant evolution in educational assessment, driven by technological advancements that enhance both efficiency and effectiveness. Systems like WebGPT exemplify this shift, utilizing real-time web searches and human feedback to improve scoring accuracy [57]. By leveraging the vast resources of the internet, these systems provide immediate, contextually relevant evaluations. Automation in test question imports, as seen in platforms like Moodle, further highlights the efficiency of computer-based testing [58]. These systems streamline diverse question formats, reducing educators' administrative load and facilitating frequent assessments, particularly in large-scale environments where traditional methods are impractical.

As illustrated in Figure 6, the figure highlights the key aspects of the shift towards computer-based testing, showcasing the efficiency improvements and the challenges faced, such as scoring accuracy and reliance on stable internet connectivity. It underscores the integral role of systems like WebGPT and Moodle in enhancing efficiency while addressing the challenges that accompany this transition. However, challenges persist, especially concerning the scoring accuracy of open-ended questions and reliance on stable internet connectivity [59]. Limited internet access can undermine the effectiveness of computer-based assessments, necessitating robust offline solutions. Additionally, ensuring accurate automated scoring for subjective responses remains a critical area for ongoing research and development.

This transition reshapes educational assessment by providing scalable, efficient, and flexible solutions to traditional challenges such as increasing student-to-teacher ratios and the labor-intensive nature of manual evaluations. The development of automated scoring systems for essays and short answers, utilizing advanced machine learning techniques, enhances grading reliability and accuracy, despite ongoing challenges in evaluating complex writing elements like content relevance and coherence [16, 15, 17]. As these technologies evolve, they promise to improve the quality and accessibility of educational assessments, aligning with modern educational needs.

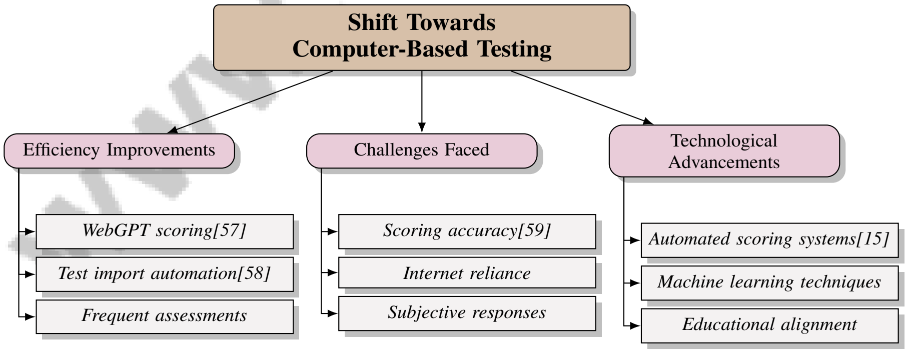


Figure 6: This figure illustrates the key aspects of the shift towards computer-based testing, highlighting efficiency improvements, challenges faced, and technological advancements. It underscores the role of systems like WebGPT and Moodle in enhancing efficiency, while also addressing challenges such as scoring accuracy and internet reliance. Furthermore, it emphasizes the impact of technological advancements like automated scoring systems and machine learning techniques in aligning educational assessments with modern needs.

5.2 Benefits of Automated Scoring Systems

Automated scoring systems offer substantial advantages in educational assessments, significantly boosting grading efficiency, objectivity, and personalization. A primary benefit is the provision of immediate feedback, which facilitates timely improvements in student learning outcomes [1]. This immediacy aids comprehension and reduces grading time for educators, allowing them to focus on instructional activities. Systems like eRevise illustrate the benefits of personalized feedback, helping students improve their writing skills through tailored guidance [28]. Integrated systems further support this personalization by offering holistic assessments and grammatical corrections, fostering comprehensive writing skill development [8].

The efficiency of automated assessments is underscored by their ability to leverage diverse datasets and advanced deep learning techniques, as demonstrated by systems like MTLHealth, which achieve high performance across multiple metrics, ensuring reliable evaluation outcomes [29]. Additionally, the cost-effectiveness of e-assessment solutions benefits educational institutions by reducing the resources required for traditional grading methods [1]. Automated scoring systems enhance feedback speed and accuracy, providing high-quality evaluations particularly beneficial for novice graders [19]. By minimizing human intervention, these systems maintain consistent standards across student responses, improving the reliability of educational assessments. The integration of advanced technologies in automated scoring systems continues to promise scalable and effective solutions that align with the evolving needs of modern educational environments.

5.3 Scalability and Reliability Issues

Scalability and reliability are critical challenges in deploying automated scoring systems across diverse educational settings. As institutions increasingly adopt these solutions, the need for systems capable of handling large data volumes without compromising accuracy becomes paramount. Advanced machine learning techniques, as seen in the MTLHealth system, demonstrate potential for high-performance outcomes across multiple evaluation metrics [29]. However, maintaining consistent performance at scale is an ongoing concern. Scalability issues are particularly pronounced with diverse linguistic inputs and varied assessment types. The ability of automated essay scoring (AES) systems to generalize across different languages and educational contexts is often limited by the availability and diversity of training datasets [2]. This limitation necessitates the development of comprehensive datasets and the implementation of transfer learning techniques to enhance system adaptability [27].

Reliability is challenged by inherent variability in student responses and potential biases in automated evaluations. Techniques such as adversarial learning, demonstrated by the OUTFOX framework, are crucial for improving the robustness of scoring systems against manipulative inputs [6]. Additionally, aligning automated scoring with human evaluators is essential for maintaining trust in these systems. Metrics like Quadratic Weighted Kappa (QWK) are instrumental in assessing the alignment between automated and human scores, ensuring reliable evaluations [22].

5.4 Ethical and Accessibility Concerns

The integration of automated scoring systems in educational assessment raises significant ethical and accessibility concerns. A primary ethical issue is the potential for bias in AI-driven evaluations, which can lead to unfair assessments of student work. Such bias often stems from training data that inadequately represents the diversity of student populations [60]. Ensuring the inclusivity and fairness of automated scoring systems requires the development of diverse datasets and the implementation of bias detection and mitigation strategies.

Accessibility is another critical concern, as ensuring that all students, regardless of technological proficiency or resource access, can benefit equitably from automated scoring systems is vital. The digital divide, characterized by disparities in access to technology and the internet, poses substantial obstacles to the widespread adoption of advanced systems, such as automated essay scoring and large language models (LLMs) like ChatGPT, potentially exacerbating educational inequities and limiting the advantages of technological advancements in learning environments [61, 15, 62, 33]. Addressing this issue involves providing adequate technological infrastructure and support to ensure full participation in computer-based assessments.

The transparency of AI algorithms in automated scoring systems is crucial for fostering trust and accountability among users, enabling stakeholders to understand how these algorithms assess and evaluate textual data, particularly in light of findings highlighting potential biases and discrepancies in scoring human versus AI-generated content [15, 17, 33]. Educators and students must comprehend the operational mechanisms of these systems and the criteria used for evaluation. Enhancing transparency can be achieved through developing interpretable models that provide insights into AI decision-making processes, fostering greater acceptance and confidence in their use.

The ethical implementation of automated scoring systems necessitates continuous dialogue among diverse stakeholders—educators, students, policymakers, and technologists—to ensure that these technologies enhance assessment efficiency while upholding educational values and objectives. This dialogue is essential, as current automated essay scoring tools often prioritize grammar and organization over critical elements like content relevance and persuasiveness, which are vital for meaningful feedback. By fostering ongoing conversations, stakeholders can work towards developing scoring systems more aligned with pedagogical goals, capable of providing actionable insights that complement traditional teaching methods [13, 16, 17, 63]. Addressing these ethical and accessibility concerns will facilitate the effective integration of automated scoring systems into educational frameworks, enhancing their potential to support equitable and inclusive learning environments.

6 Challenges and Future Directions

6.1 Challenges in Automated Scoring and Feedback

Automated Essay Scoring (AES) systems confront several challenges that affect their reliability and effectiveness in educational settings. A major issue is the requirement for extensive data and computational resources to train large language models (LLMs), which may be prohibitive for smaller organizations and raise ethical concerns about data usage [5]. The computational complexity inherent in models using Latent Semantic Analysis (LSA) further complicates scalability and accessibility in large document collections [20].

The ambiguity and variability in open-ended essay texts present significant hurdles for developing effective machine learning models [18]. Often, benchmarks focus on single drafts, overlooking the iterative nature of writing and revision [21]. Moreover, the gap between Automated Writing Evaluation (AWE) and Grammatical Error Correction (GEC) systems limits comprehensive feedback addressing both writing quality and grammatical accuracy [8].

Biases in automated scoring systems pose critical concerns, potentially leading to skewed scoring outcomes. Current evaluation methods may not accurately reflect text quality, mirroring human scoring biases [7]. The performance of detectors diminishes significantly when LLM-generated texts are paraphrased, rendering them vulnerable to evasion strategies [6].

Additionally, the lack of technical infrastructure and technological familiarity among students and educators presents barriers to effective AES implementation [1]. Addressing these challenges requires comprehensive training and support for stakeholders to engage effectively with automated scoring technologies.

To address these issues, it is crucial to establish robust feedback mechanisms, actively reduce biases in scoring algorithms, and enhance system interpretability. Such improvements will ensure that AES evaluations are accurate and fair, capable of handling the complexities of language and diverse student demographics, while mitigating oversensitivity and overstability issues that may lead to misleading assessments [39, 56, 33].

6.2 Integrating AI Technologies in Educational Assessment

The integration of AI technologies into educational assessment significantly advances the capabilities and reliability of automated scoring systems. Generative AI models, particularly Large Language Models (LLMs), enhance the accuracy and efficiency of these systems by improving the processing and evaluation of complex linguistic features, thus providing a deeper understanding of student responses [35].

Innovative methodologies, such as Topic-to-Essay Generation via Knowledge Enhancement (TEGKE), emphasize knowledge transfer and adversarial training to generate coherent and contextually relevant essay outputs, bridging information gaps and enhancing scoring accuracy [64]. Furthermore, autoregressive models like the Autoregressive Essay Scoring (ArTS) framework capture trait dependencies within a single model, streamlining the scoring process while maintaining high levels of accuracy and reliability [65].

AI-generated, for reference only.

Deep learning architectures, such as DeLAES, effectively capture semantic features and contextual information, enhancing essay evaluation precision compared to traditional methods [42]. Reasoning distillation-based evaluation (RDBE) methodologies further contribute to the interpretability and performance of scoring systems, offering insights into response quality [43].

Structured workshops, like the Prompt Engineering Workshop, aim to enhance students' capabilities in interacting with AI technologies, equipping them with skills to maximize the benefits of AI in educational assessment [10]. Ethical considerations and interdisciplinary approaches are critical for the effective integration of AI technologies into educational assessment. Future research should focus on developing ethical guidelines and exploring the implications of AI on teaching and learning dynamics to support equitable and inclusive educational environments.

6.3 Data and Model Limitations

The effectiveness of automated scoring systems is often constrained by data quality and model robustness. In the context of LLMs, performance is particularly sensitive to data quality variations, leading to inconsistencies in scoring outcomes [35]. This underscores the need for high-quality, diverse datasets that represent the range of student responses and linguistic features in educational settings.

The VerAs system illustrates challenges related to model generalizability; while effective in scoring complex outputs, it requires retraining for new datasets, limiting its applicability across diverse educational contexts [23]. This emphasizes the importance of developing models capable of generalizing effectively without significant performance degradation.

Neural multi-task learning approaches have shown promise in enhancing AES but may not significantly benefit related tasks, such as Grammatical Error Detection (GED), from overall essay scores [66]. This highlights the necessity for models that address multiple facets of writing evaluation, providing comprehensive assessments of both content quality and linguistic accuracy.

Addressing data quality and model architecture limitations is essential for enhancing the reliability and scalability of automated scoring systems. Advancements in machine learning, particularly transformer-based models, reveal varying performance in assessing human versus GPT-generated text, emphasizing the need for optimized training data to improve effectiveness across different paradigms [15, 67]. Future research should focus on developing robust models that maintain performance across diverse datasets and educational contexts to effectively support assessment practices.

6.4 Enhancing Feedback and Evaluation

Benchmark	Size	Domain	Task Format	Metric
AI-PHE[68]	50	Physics	Essay Writing	Average Mark, Plagiarism Score
AES-Benchmark[69]	17,000	Automated Essay Scoring	Multi-label Classification	Accuracy
AES-Bench[32]	12,100	Linguistic Assessment	Essay Scoring	Pearson correlation, MAE
ArguGPT[70]	8,153	Argumentative Writing	Essay Generation	Accuracy, F1-score
AES-Benchmark[67]	50,000	Automated Essay Scoring	Essay Scoring	QWK, SMD
DREsS[71]	49,979	Efl Writing	Automated Essay Scoring	QWK
ASAP-Hindi[72]	1,909	Automated Essay Scoring	Essay Scoring	Quadratic Weighted Kappa
AES[38]	7,000	Essay Scoring	Scoring	QWK, Acc

Table 1: This table provides a comprehensive overview of various benchmarks used in Automated Essay Scoring (AES) systems, detailing their size, domain, task format, and evaluation metrics. These benchmarks are crucial for assessing the performance and reliability of AES systems in different contexts, facilitating advancements in educational assessment methodologies.

Enhancing feedback mechanisms and evaluation processes in Automated Essay Scoring (AES) systems is crucial for improving educational outcomes and ensuring assessment reliability. Leveraging advanced language models to provide detailed, context-aware insights into student writing can enrich instructional feedback, reflecting the complexity of student responses [8]. Multi-task learning techniques can validate predictions with additional annotated data, enhancing the robustness of feedback mechanisms [21].

The Multi-Trait Scoring (MTS) approach offers significant benefits, including improved scoring consistency and adaptability to new prompts without retraining. This aligns with human evaluations by focusing on trait-specific assessments, addressing adversarial responses, and enhancing the fairness of automated scoring systems [18]. Future research should explore integrating features beyond string kernels to improve identification accuracy, directing efforts to enhance feedback and evaluation in natural language inference (NLI) systems [20].

Integrating human-AI collaboration in automated scoring systems can enhance feedback mechanisms and evaluation processes, ensuring more accurate and contextually relevant assessments [21]. This collaborative approach can diversify feedback and ensure consistency between trait and overall scores. Incorporating affective computing may also improve responsiveness to student emotions, enhancing pedagogical strategies [8].

Real-time analytics provided to educators through learning analytics dashboards can significantly enhance their ability to evaluate student performance, improving feedback mechanisms [1]. Future research should focus on developing more explainable feedback models and exploring personalized LLM agents that adapt to individual learner needs, ensuring AES systems respond effectively to unique student challenges [7].

Innovative frameworks such as OUTFOX can improve the detection of LLM-generated text and can be adapted to enhance feedback mechanisms in AES systems, ensuring accurate evaluations and addressing potential biases [6]. By developing ethical AI usage guidelines and enhancing AI literacy among educators, AES systems can be implemented responsibly, ultimately supporting equitable and inclusive educational practices.

The strategies discussed highlight the potential of Automated Essay Scoring (AES) systems to deliver effective, personalized feedback tailored to individual learners. By leveraging advanced machine learning techniques, such as prompt-specific models and large language models, AES can enhance educational outcomes through timely, consistent, and contextually relevant evaluations. Addressing inherent biases and utilizing comprehensive datasets that consider diverse demographic factors will further improve reliability across various learning environments, fostering personalized learning experiences and contributing to the advancement of educational assessment methodologies [31, 32, 33, 73]. Table 1 presents a detailed summary of representative benchmarks utilized in Automated Essay Scoring (AES) systems, highlighting their significance in enhancing feedback mechanisms and evaluation processes.

7 Conclusion

Automated scoring systems represent a significant advancement in educational assessment, offering solutions that are scalable, efficient, and objective, effectively addressing the limitations of traditional evaluation methods. The integration of sophisticated AI technologies, particularly Automated Essay Scoring (AES) systems, not only enhances feedback mechanisms but also positively impacts student learning outcomes. The RMTS framework exemplifies this by integrating LLM-generated rationales to improve the accuracy and reliability of multi-trait essay scoring.

Key resources such as the school student essay corpus are instrumental in examining the link between argumentative structure and essay quality, thereby supporting ongoing research in writing assistance. The Reasoning Distillation-Based Evaluation (RDBE) methodology further underscores the importance of advanced models in enhancing automated scoring, achieving state-of-the-art performance in AES.

Future research directions should focus on developing more sophisticated models and features to further improve AES performance. The integration of Automated Writing Evaluation (AWE) and Grammatical Error Correction (GEC) systems has demonstrated potential in advancing the writing skills of second language learners, highlighting their value in language education.

The ongoing advancements in AES and related technologies mark a transformative shift in educational assessment practices, contributing to the enhancement of educational quality and equity through personalized and reliable evaluations. As these systems continue to develop, they hold the promise of significantly enhancing learner engagement and critical thinking, aligning closely with educational objectives and societal values.

References

[1] Nuha Alruwais, Gary Wills, and Mike Wald. Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, 8(1):34–37, 2018.

[2] Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. Swell on the rise: Swedish learner language corpus for european reference level studies, 2016.

[3] Olga Tapalova and Nadezhda Zhiyenbayeva. Artificial intelligence in education: Aied for personalised learning pathways. *Electronic Journal of e-Learning*, 20(5):639–653, 2022.

[4] Ehsan Latif, Yifan Zhou, Shuchen Guo, Yizhu Gao, Lehong Shi, Matthew Nayaaba, Gyeonggeon Lee, Liang Zhang, Arne Bewersdorff, Luyang Fang, Xiantong Yang, Huaqin Zhao, Hanqi Jiang, Haoran Lu, Jiaxi Li, Jichao Yu, Weihang You, Zhengliang Liu, Vincent Shung Liu, Hui Wang, Zihao Wu, Jin Lu, Fei Dou, Ping Ma, Ninghao Liu, Tianming Liu, and Xiaoming Zhai. A systematic assessment of openai o1-preview for higher order thinking in education, 2024.

[5] João Gonçalves, Nick Jelicic, Michele Murgia, and Evert Stamhuis. The advantages of context specific language models: the case of the erasmian language model, 2024.

[6] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples, 2024.

[7] Emiel van Miltenburg. Evaluating nlg systems: A brief introduction, 2023.

[8] Izia Xiaoxiao Wang, Xihan Wu, Edith Coates, Min Zeng, Jiexin Kuang, Siliang Liu, Mengyang Qiu, and Jungyeul Park. Neural automated writing evaluation with corrective feedback, 2024.

[9] Jiahui Huang, Salmiza Saleh, and Yufei Liu. A review on artificial intelligence in education. *Academic Journal of Interdisciplinary Studies*, 10(3), 2021.

[10] David James Woo, Deliang Wang, Tim Yung, and Kai Guo. Effects of a prompt engineering intervention on undergraduate students' ai self-efficacy, ai knowledge and prompt engineering ability: A mixed methods study, 2024.

[11] You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. Improving performance of automated essay scoring by using back-translation essays and adjusted scores, 2022.

[12] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction, 2024.

[13] Yann Hicke, Tonghua Tian, Karan Jha, and Choong Hee Kim. Automated essay scoring in argumentative writing: Deberteachingassistant, 2023.

[14] Azam Rabiee, Alok Goel, Johnson D'Souza, and Saurabh Khanwalkar. Question-type identification for academic questions in online learning platform, 2022.

[15] Marialena Bevilacqua, Kezia Oketch, Ruiyang Qin, Will Stamey, Xinyuan Zhang, Yi Gan, Kai Yang, and Ahmed Abbasi. When automated assessment meets automated content generation: Examining text quality in the era of gpts, 2023.

[16] Mike Hardy. Toward educator-focused automated scoring systems for reading and writing, 2021.

[17] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527, 2022.

[18] Akriti Chadda, Kelly Song, Raman Chandrasekar, and Ian Gorton. Engineering an intelligent essay scoring and feedback system: An experience report, 2021.

[19] Firuz Kamalov, David Santandreu Calonge, and Ikhlaas Gurrib. New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability*, 15(16):12451, 2023.

[20] Tuomo Kakkonen, Niko Myller, and Erkki Sutinen. Applying part-of-seech enhanced lsa to automatic essay grading, 2006.

[21] Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. Argrewrite v.2: an annotated argumentative revisions corpus, 2022.

[22] Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. Can large language models automatically score proficiency of written essays?, 2024.

[23] Berk Atil, Mahsa Sheikhi Karizaki, and Rebecca J. Passonneau. Veras: Verify then assess stem lab reports, 2024.

[24] Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. Automated essay scoring using transformer models, 2021.

[25] Subhadip Maji, Anudeep Srivatsav Appe, Raghav Bali, Veera Raghavendra Chikka, Arijit Ghosh Chowdhury, and Vamsi M Bhandaru. An interpretable deep learning system for automatically scoring request for proposals, 2020.

[26] Toru Ishida, Tongxi Liu, Hailong Wang, and William K. Cheung. Large language models as partners in student essay evaluation, 2024.

[27] Nuowei Liu, Xinhao Chen, Hongyi Wu, Changzhi Sun, Man Lan, Yuanbin Wu, Xiaopeng Bai, Shaoguang Mao, and Yan Xia. Cerd: A comprehensive chinese rhetoric dataset for rhetorical understanding and generation in essays, 2024.

[28] Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, Lindsay Clare Matsumura, Emily Howe, and Rafael Quintana. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing, 2019.

[29] Joseph Valencia and Erin Yao. Mtlhealth: A deep learning system for detecting disturbing content in student essays, 2021.

[30] Georgios Balikas. Lexical bias in essay level prediction, 2018.

[31] Rayed Ghazawi and Edwin Simpson. Automated essay scoring in arabic: a dataset and analysis of a bert-based system, 2024.

[32] Sowmya Vajjala. Automated assessment of non-native learner essays: Investigating the role of linguistic features, 2016.

[33] Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability, 2024.

[34] Kun Sun and Rong Wang. Automatic essay multi-dimensional scoring with fine-tuning and multiple regression, 2024.

[35] Saurabh Pahune and Manoj Chandrasekharan. Several categories of large language models (llms): A short survey, 2023.

[36] Li-Hsin Chang, Iiro Rastas, Sampo Pyysalo, and Filip Ginter. Deep learning for sentence clustering in essay grading support, 2021.

[37] SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms, 2025.

[38] Christopher M Ormerod, Akanksha Malhotra, and Amir Jafari. Automated essay scoring using efficient transformer-based language models, 2021.

[39] Anubha Kabra, Mehar Bhatia, Yaman Kumar, Junyi Jessy Li, and Rajiv Ratn Shah. Evaluation toolkit for robustness testing of automatic essay scoring systems, 2021.

[40] Chen Zheng, Huan Zhang, Yan Zhao, and Yuxuan Lai. Improving the generalization ability in essay coherence evaluation through monotonic constraints, 2023.

AI-generated, for reference only.

[41] Oscar Morris. The effectiveness of a dynamic loss function in neural network based automated essay scoring, 2023.

[42] Tsegaye Misikir Tashu, Chandresh Kumar Maurya, and Tomas Horvath. Deep learning architecture for automatic essay scoring, 2022.

[43] Ali Ghiasvand Mohammadkhani. Rdbe: Reasoning distillation-based evaluation enhances automatic essay scoring, 2024.

[44] Tor Ole B. Odden and John Burk. Computational essays in the physics classroom, 2020.

[45] Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. Language models and automated essay scoring, 2019.

[46] Liang Zhang, Jionghao Lin, Ziyi Kuang, Sheng Xu, and Xiangen Hu. Spl: A socratic playground for learning powered by large language model, 2024.

[47] Disha Shrivastava, Abhijit Mishra, and Karthik Sankaranarayanan. Modeling topical coherence in discourse without supervision, 2018.

[48] Jingshen Zhang, Xiangyu Yang, Xinkai Su, Xinglu Chen, Tianyou Huang, and Xinying Qiu. System report for ccl24-eval task 7: Multi-error modeling and fluency-targeted pre-training for chinese essay evaluation, 2024.

[49] Haoran Zhang and Diane Litman. Automated topical component extraction using neural network attention scores from source-based essay scoring, 2020.

[50] Bagiya Lakshmi S, Sanjjushri Varshini R, Rohith Mahadevan, and Raja CSP Raman. Comparative study and framework for automated summariser evaluation: Langchain and hybrid algorithms, 2023.

[51] Xinying Qiu, Shuxuan Liao, Jiajun Xie, and Jian-Yun Nie. Tapping the potential of coherence and syntactic features in neural models for automatic essay scoring, 2022.

[52] Tushar Abhishek, Daksh Rawat, Manish Gupta, and Vasudeva Varma. Transformer models for text coherence assessment, 2022.

[53] Tazin Afrin, Elaine Wang, Diane Litman, Lindsay C. Matsumura, and Richard Correnti. Annotation and classification of evidence and reasoning revisions in argumentative writing, 2021.

[54] Yan Liu and Yazheng Yang. Enhance long text understanding via distilled gist detector from abstractive summarization, 2021.

[55] Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. A survey on using gaze behaviour for natural language processing, 2022.

[56] Yaman Kumar Singla, Swapnil Parekh, Somesh Singh, Junyi Jessy Li, Rajiv Ratn Shah, and Changyou Chen. Aes systems are both overstable and oversensitive: Explaining why and proposing defenses, 2021.

[57] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[58] Iryna S. Mintii, Svitlana V. Shokaliuk, Tetiana A. Vakaliuk, Mykhailo M. Mintii, and Vladimir N. Soloviev. Import test questions into moodle lms, 2020.

[59] ABPIRS Sari, Dwi Iswahyuni, Sri Rejeki, and Sutanto Sutanto. Google forms as an efl assessment tool: Positive features and limitations. 2020.

[60] Catalin Vrabie. Artificial intelligence from idea to implementation. how can ai reshape the education landscape?, 2024.

[61] Zijie Zeng, Lele Sha, Yuheng Li, Kaixun Yang, Dragan Gašević, and Guanliang Chen. Towards automatic boundary detection for human-ai collaborative hybrid essay in education, 2023.

[62] Andrew Jelson and Sang Won Lee. An empirical study to understand how students use chatgpt for writing essays and how it affects their ownership, 2024.

[63] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review, 2023.

[64] Zhiyue Liu, Jiahai Wang, and Zhenghong Li. Topic-to-essay generation with comprehensive knowledge enhancement, 2021.

[65] Heejin Do, Yunsu Kim, and Gary Geunbae Lee. Autoregressive score generation for multi-trait essay scoring, 2024.

[66] Ronan Cummins and Marek Rei. Neural multi-task learning in automated assessment, 2018.

[67] Christopher Ormerod, Amir Jafari, Susan Lottridge, Milan Patel, Amy Harris, and Paul van Wamelen. The effects of data size on automated essay scoring engines, 2021.

[68] Will Yeadon, Oto-Obong Inyang, Arin Mizouri, Alex Peach, and Craig Testrow. The death of the short-form physics essay in the coming ai revolution, 2022.

[69] Kshitij Gupta. Data augmentation for automated essay scoring using transformer models, 2023.

[70] Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models, 2023.

[71] Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. Dress: Dataset for rubric-based essay scoring on efl writing, 2024.

[72] Shubhankar Singh, Anirudh Pupneja, Shivaansh Mital, Cheril Shah, Manish Bawkar, Lakshman Prasad Gupta, Ajit Kumar, Yaman Kumar, Rushali Gupta, and Rajiv Ratn Shah. H-aes: Towards automated essay scoring for hindi, 2023.

[73] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. Exploring llm prompting strategies for joint essay scoring and feedback generation, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.