SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Automated Essay Scoring In Educational AI: A Review Of Fairness, Bias, And Generalization

Pramod Dharmadhikari[1], Santosh Borde[2], Navnath Kale[3], Sonali Rangdale[4]

[1] PhD Scholar, School of Computer Science Engineering, Ajinkya DY Patil University, Pune, India pramod.dharmadhikari@adypu.edu.in
[2] Professor, School of Computer Science Engineering, Ajinkya DY Patil University, Pune, India, sprao.borde@adypu.edu.in
[3] Sr. Assistant Professor, Department of Computer Engineering, MIT Academy of Engineering, Alandi, Pune, navnath.kale@mitaoe.ac.in
[4] Dept. of Computer Science Engineering, G H Raisoni skill Tech University, Pune, India, sonalirangdale1278@gmail.com

Abstract

The automated essay scoring (AES) is a key use case of educational AI, and there has been an increasing adoption of this in the fields of formative support, large-scale assessment, and feedback-heavy writing instruction. While transformer-based models and large language models (LLMs) have shown potential for enhancing the technical properties of AES, they have also raised issues of fairness, bias, and generalization across prompts, populations, and educational settings. This review explores the shift from feature-based and hybrid models to neural and LLM-based systems and the importance of correctly deploying models that are trustworthy. The article is organized through the lens of literature review using PRISMA, which reviews studies on the basics of AES, benchmark corpora, fairness-centered analyses, transfer-learning studies, and newer studies of LLLM assessments. The review reveals the influence of benchmark datasets on methodological development, but also the limitation on the type of validity and equity claims that can be made. It also identifies a range of fairness outcomes that differ across model families, fairness measures, the composition of the training data and across various demographic subgroups, and transfer and domain-shift studies show that fairness limits are enduring across prompts and students. While the possibilities of scalability and richer contextual models are new, there are also concerns such as calibration issues, representation bias, and opacity and dependence on human-like yet unstable judgements in LLM-based AES. Significant areas of weakness in the following areas remain to be addressed: fairness portability, multimodal assessment, intersectional evaluation, and benchmark design for real-world educational settings. Future AES research should combine fairness-aware modeling, strong evaluation, and educational validity into a unified framework to enable systems that are accurate, fair, transparent, and trustworthy.

Keyword: Automated Essay Scoring, Educational AI, Fairness, Bias, Generalization, Large Language Models, Transfer Learning

Introduction

Writing is a key component of educational assessment as both a language skill and a reflection of reasoning, content knowledge, and the ability to communicate ideas within the context of task demands (Williamson et al., 2012). However, essay scoring at scale is expensive, time-consuming, and subject to rater variability between essays and across prompts, contexts, and populations (Attali & Burstein, 2006; Williamson et al., 2012). To overcome that tension, and to be more efficient and score more consistently, came Automated Essay Scoring (AES). However, AES is not just a technical solution, it's a learning AI challenge where validity, equity and trust are inextricably bound up in predictive performance over time.

The field has shifted from feature-based system to neural, transformer-based, and Large Language Model (LLM)-based approaches. Early operational systems, like e-rater, followed an approach of engineering linguistic and discourse features in defensible ways, to approximate human scoring (Attali & Burstein, 2006). Then benchmark datasets facilitated comparative research, particularly in the context of the Automated Student Assessment Prize (ASAP) which introduced standardization of the assessment process and the idea of performance by prompt became a research priority (Shermis, 2014). Recent systems using pre-trained language models have increased the capacity of the representations and decreased the need for hand-designed features, but also have raised concerns about their opacity, calibration, and portability to different contexts (Devlin et al., 2019).

These technical changes are important because AES systems are not working in a neutral environment. Small systematic differences in performance between groups can be educationally significant when scores are used to place, provide feedback or hold students accountable. Previous research indicates that fairness is not a characteristic of a model, but a relationship between the data, scoring labels, model family, and evaluation metric . In addition, it has been found that fairness issues can be learned and reproduced by automated systems: that is, that fairness issues can lie in the data used to produce scores, not just in the scores themselves. More recent findings from AES have also indicated that the use of balanced training data can help to minimize some of the subgroup differences, although not all ().

The parallel concern with generalization is also raised. The high success rate of a good AES system on a given prompt or corpus can be lost when the system is transferred to another one (Phandi et al., 2015). This equates to a lack of educational robustness unless benchmark performance is used in conjunction with measures of attainment. In a narrow domain, a model can be correct, but it can misbehave when essay length or discourse structure, as well as the demographics or scoring practice, vary. Therefore, issues of generalization and fairness are not discrete technical add-on questions, but are intertwined issues of validity.

This review therefore examines AES through four connected lenses: model evolution, fairness and bias, generalization and transfer, and the implications of LLM-based scoring for trustworthy educational AI. **Figure 1** organizes that logic visually by showing how benchmark corpora, model families, fairness analysis, and domain shift interact within the review framework.

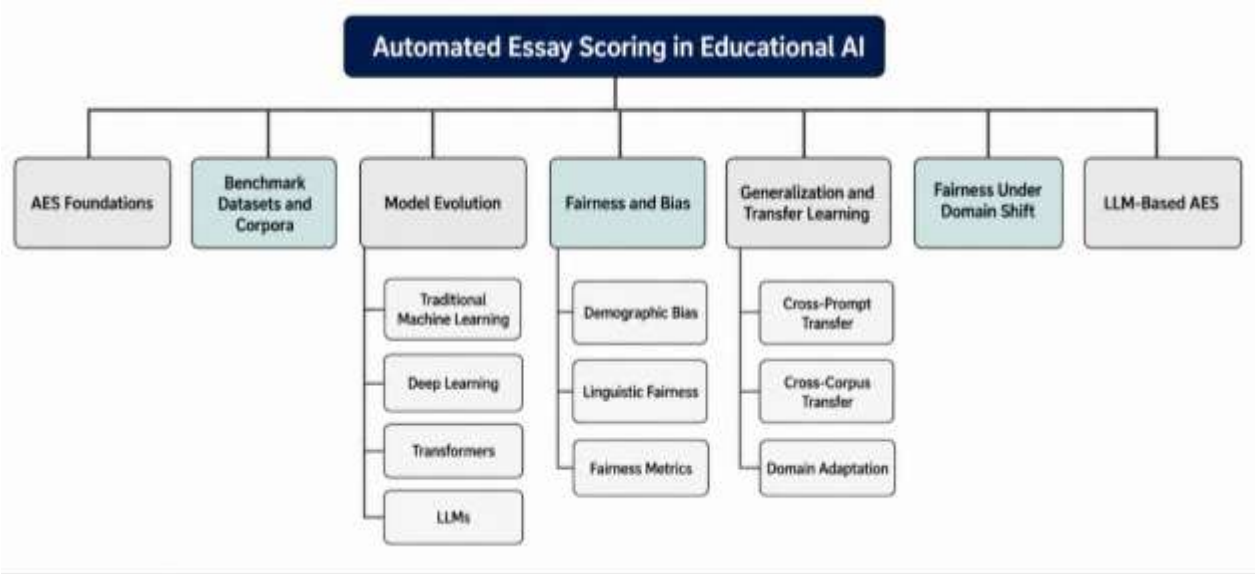

Figure 1 The review taxonomy used to organize the manuscript's thematic structure

The figure should be interpreted as a conceptual roadmap rather than a chronology: it highlights how choices in dataset design shape both model development and the kinds of fairness claims that can legitimately be made. That framing is important for ET&S because it links assessment technology to educational validity instead of treating accuracy as the only outcome of interest.

Table 1 summarizes the major thematic areas covered in the review and shows how each section maps to the central research questions

Table 1. Major Themes Covered in the Review

Theme	Description	Key Topics Covered	Related Research Questions
AES Foundations	Traces the historical and conceptual development of AES from early operational systems to modern AI-based approaches.	Early feature-based scoring, validity, reliability, human-machine agreement, operational deployment.	RQ1

Benchmark Datasets and Corpora	Reviews the datasets that structure AES research and shape evaluation claims.	ASAP, ASAP 2.0, PERSUADE, corpus design, rubric structure, demographic metadata.	RQ1, RQ2
Traditional Machine Learning AES	Summarizes feature-engineered AES systems and their role in establishing early scoring baselines.	Hand-crafted features, regression and tree-based models, interpretability, rubric alignment.	RQ1
Deep Learning AES	Reviews neural AES models that reduced feature engineering and improved contextual representation.	CNNs, RNNs, attention mechanisms, hybrid systems, performance trade-offs.	RQ1
Transformer-Based AES	Examines AES systems built on transformer architectures and their impact on scoring, transfer, and efficiency.	BERT, Longformer, fine-tuning, long-context modeling, computational cost.	RQ1, RQ3
LLM-Based AES	Focuses on prompted and finetuned large language model approaches to essay scoring.	Prompting, finetuning, representation bias, explainability, reliability, hallucination risk.	RQ1, RQ4
Fairness and Bias	Synthesizes evidence on subgroup disparities, measurement bias, and fairness evaluation methods.	Racial/ethnic bias, gender bias, SES effects, ELL effects, label bias, mitigation.	RQ2, RQ4
Generalization	Addresses how well AES models perform across prompts, tasks, and datasets.	Cross-prompt transfer, prompt independence, performance under distribution shift.	RQ3
Transfer Learning	Reviews methods designed to improve cross-domain performance in AES.	Domain adaptation, representation reuse, prompt-independent scoring, sparse-label transfer.	RQ3
Domain Shift	Considers how changes in prompt, corpus, or population affect AES reliability and validity.	Cross-corpus evaluation, demographic shift, institutional shift, annotation mismatch.	RQ3, RQ4
Fairness Portability	Introduces the concern that fairness may not transfer across datasets, demographics, or contexts.	Cross-corpus fairness, subgroup stability, fairness under shift.	RQ4
Future Research Directions	Consolidates the agenda emerging from identified gaps in fairness, transfer, and LLM-based AES.	Fairness-by-design, multilingual datasets, explainable LLMs, governance, auditing.	RQ5

The table is useful because it creates transparency in the structure of the review; AES foundations, benchmark corpora, model evolution, and fairness, transfer, domain shift, and future directions are all considered interconnected problems, rather than separate topics. That structure facilitates a synthesis that benefits researchers, developers, educators and policy audiences who require information regarding the technical performance and educational outcomes.

It examines five questions: Has AES progressed technically? What are the benchmark datasets that have influenced the field? What is known about fairness and bias in the various families of models and learner groups? Does AES transfer well to other prompts and domains? What does the use of LLMs for AES suggest for trustworthy uses in education? In combination, these questions define AES as a mature, but still developing field of educational AI. The main conclusion of the review is that the future needs to be built on the integration of the predictive component and the evaluation component with concerns of fairness, the cross domain robustness and more robust educational validity arguments.

2. Review Methodology

The design of the present review was systematic literature review, focusing on the field of automated essay scoring (AES) which is shaped by the technical performance, fairness and generalization of the educational AI. This is because research into AES is spread out across the fields of educational measurement, natural language processing and educational technology, and the studies vary significantly in the datasets used, the prompts they explored, the families of models used, and the criteria used to assess the models. Therefore, the review followed the PRISMA principles for identifying, screening, eligibility assessment, and inclusion of studies (Moher et al., 2009; Page et al., 2021). This also fits with the recommendations for review methodology to include explicit procedures where the evidence base is heterogeneous (Tranfield et al., 2003, Xiao & Watson, 2019).

The search strategy was formulated to include fundamental AES research and more recent studies on fairness, bias, transfer learning, domain shift and LLM. We did a search in Scopus, Web of Science, IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar which covered the interdisciplinary publication pattern of AES research. The terms used for each search were the core constructs of "automated essay scoring" and "AES" plus concept terms: "fairness," "bias," "generalization," "transfer learning," "domain shift," "educational AI," and "large language models." It was necessary to use this wider search logic as important AES studies are found in assessment, NLP, learning analytics and AI-in-education research streams, but not in a single one.

The inclusion criteria aimed at studies that had a direct contribution to AES theory, method, dataset or evaluation. The following types of works were accepted: peer-reviewed journal articles, conference papers, and benchmark or corpus papers that addressed one or more of the review's focus areas: AES foundations, benchmark datasets, fairness and bias, transfer and generalization, or LLM based scoring. Studies that were included as foundational, even in older years, were still considered important for developing later in the operational scoring and validity arguments. Exclusion criteria eliminated articles that were not in English, non-peer-reviewed articles, and articles that were not directly related to AES, such as broader articles on education and AI or general fairness papers, that did not directly address essay scores.

The screening process was carried out in a staged logical manner similar to the PRISMA. Once identified, duplicated records were deleted. Titles and abstracts were then reviewed to see if a study was directly related to AES or if it helped with one of review themes. Three criteria were used to evaluate the full text for eligibility: relevance to the review scope, methodological or dataset description sufficient for understanding the method, and meaningful contribution to understanding model evolution, fairness, transferability, or robustness. Thematic coding and synthesis were conducted on all of the studies that met the following criteria. This process highlighted conceptual relevance and methodological transparency, instead of the number of citations or novelty.

All studies included were analysed using a structured extraction protocol to facilitate comparative analysis. The following fields were extracted: authors, year of publication, dataset/corpus, AES method, fairness metrics (where applicable), transfer-learning approach (where applicable), generalization findings, and key contribution. The fields chosen were based on (1) their direct association with the questions being addressed in the review, and (2) their ability to facilitate comparison of technically varied studies. In application, the extraction process was also made possible to differentiate studies using the same data but different questions, such as benchmark performance or subgroup differences or cross-prompt portability.

Table 2 summarizes the review design, search scope, screening logic, and thematic coding framework used to construct the final review corpus.

Table 2. Review Methodology and Corpus Selection

Component	Content
Review design	Systematic literature review guided by PRISMA.
Search domains	Educational measurement, NLP, educational technology, AI in education.
Databases	Scopus, Web of Science, IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Google Scholar.
Core search concepts	Automated Essay Scoring, AES, fairness, bias, generalization, transfer learning, domain shift, large language models, educational AI.
Inclusion criteria	Peer-reviewed journal articles and conference papers; benchmark/dataset papers; empirical or methodological AES studies; fairness, transfer, generalization, or LLM-based AES relevance; foundational AES papers.

Exclusion criteria	Non-English publications; non-peer-reviewed items; indirect AES relevance; generic educational AI not tied to essay scoring.
Screening stages	Identification, de-duplication, title/abstract screening, full-text eligibility review, thematic coding.
Extraction fields	Authors, year, dataset/corpus, AES method, fairness metrics, transfer-learning approach, generalization findings, key contribution.
Thematic coding categories	AES foundations; benchmark datasets; traditional AES; deep learning AES; LLM-based AES; fairness and bias; transfer learning; domain generalization; future directions.
Purpose	Ensure reproducibility and support thematic synthesis across methodologically diverse studies.

The table should be interpreted as a methodological map for the review; it indicates that the selection of studies was not only on the basis of relevance to the topic, but also the need to compare work from several methodological traditions. It also explains the way the review changed from a general search of a database to a more specific corpus for critical synthesis and not just descriptive listing.

The final compiled corpus was divided into the following thematic content: AES foundations, traditional AES, benchmark datasets, deep learning AES, LLM-based AES, fairness and bias, transfer learning, domain generalization, and future directions. It was selected because it reflects the evolution of the discipline and the conceptual conflicts that haunted it. It is most importantly designed for an ET&S style review that focuses on synthesizing the patterns, contradictions, and gaps rather than paper-by-paper summarizing. The resulting methodology then offers a defensible path to consider the trajectory of AES research from operational to more general issues of fairness, robustness, and trustworthy use of education.

3. Foundations of Automated Essay Scoring

Automated essay scoring (AES) has been derived from the large-scale development of writing assessment.Automated essay scoring (AES) arose out of the development of large-scale writing assessment. Human reading is an expensive, slow, and difficult to standardize process when thousands of essays need to be read in a timely manner, particularly when assessing essays in an operational testing or classroom assessment setting (Attali & Burstein, 2006; Shermis, 2014). Not yet, however, was AES simply a technical solution to scale. It was also a measurement issue, where any scoring system was intended to capture human judgment as closely as possible and to maintain the writing structure and intended learning value of the written score (Williamson et al., 2012). The conflict between efficiency and validity continues to affect the field.

3.1 Roots of Automated Essay Scoring.

The first models of AES were based on the assumption that the quality of the writing text could be assessed based on patterns of its observable features – length, lexical complexity, syntactic complexity, organization, surface correctness. This logic had strong appeal, as it could be calculated in an accurate manner and scaled to predictions of scores. However, the process also limited the writing structure to that which was most easily measurable. This meant that early AES systems worked well when prompts, populations, and rubrics didn't change, but not as well when they started to focus on broader rhetorical quality or cross-context performance.

3.2 Early AES Systems

Early efforts like Project Essay Grade (PEG) and the Intelligent Essay Assessor (IEA) proved that machine scoring was possible, while also showcasing the assumptions that must be simplified in order to achieve automated assessment. PEG focused on measurable text proxies, while IEA proposed to use latent semantic analysis to measure content similarity and topical relevance (Foltz et al., 2000; Miller, 2003). These systems formed two fundamental concepts in AES: one, that automated scoring can replace human scoring within limited context; two, that efficiency may be obtained at the expense of the richness of constructs. Their historic value is not so much in the way they play today, but in establishing the ground rules for the sport.

3.3 e-rater and Operational AES

e-rater was the most influential early operational system, bringing AES nearer to its use in education by connecting features-based scores to psychometric and assessment logic (Attali & Burstein, 2006). Its architecture was based on the grammar, usage, mechanics, style and discourse features, and the rules used in scoring were more interpretable and defensible for institutional use. A great step forward as it introduced the techniques of NLP with rubric assessment methodology instead of AES being a predictive activity. At the same time, e-rater revealed a structural constraint that is relevant to the present: a requirement that validity be based

on selected features and carefully designed prompts means that performance can become poorer when the scores are used in a new context.

3.4 The ASAP Benchmark and Standardization of AES Research.

The introduction of the Automated Student Assessment Prize (ASAP) marked a significant shift, standardizing prompts, scales and reporting conventions and thus making AES a broadly comparable research area (Shermis, 2014). ASAP enabled researchers to evaluate in a reproducible way and provided them with a common benchmark for methodological innovation. But its success also cut down on what could be considered progress. A high degree of score agreement in many studies can lead to incorrect conclusions about the broader educational validity of the scores, despite ongoing issues of prompt dependence, underperformance by certain groups, and transferability to different contexts.

3.5 Key Challenges in Traditional AES

Although traditional AES set the groundwork, it also bequeathed the heart of the shortcomings that were picked up by subsequent studies. First, prompt dependence was an issue, as systems tended to work best with the prompts for which they had been trained, and to struggle on new prompts or genres (Shermis, 2014; Phandi et al., 2015). Second, feature engineering restricted the coverage of the constructs, leading the models to be based on surface features of the text instead of more complex aspects of the argumentation or discourse quality. Third, fairness concerns were also evident in early systems, as gains in accuracy from a model trained on human ratings can also replicate or exacerbate inequities contained in those ratings (Williamson et al., 2012). These challenges reveal that generalization and fairness are not new issues that stem from the adoption of LLM systems, but rather root issues that are central to AES.

3.6 Move to Modern AES

However, the drawbacks of traditional AES led to the adoption of machine learning, deep learning, transformer-based models, and more recently, LLM-based scoring. Neural models eliminated the need for handcrafted features, enhanced discourse and context representations, and transformer models expanded the ability to model context and longer texts (Nadeem et al., 2019; Taghipour & Ng, 2016; Devlin et al., 2019). Recent research with LLM models has introduced additional levels of flexibility, though it has also raised perennial concerns about fairness, transparency and robustness at the population and setting levels (). Because of this, it is important to revisit the basics of PEG, IEA, e-rater, and ASAP: why the current controversy over fairness, transfer, and domain shift is not a new issue, but a continuation of longstanding debates.

4. Benchmark Datasets and Corpora for Automated Essay Scoring

The key to AES is the role of benchmark datasets and corpora in defining the types of scoring problems for study, the types of comparisons that can be made, and what kinds of statements of validity and fairness can be made. To enhance the cumulative nature of AES research, shared corpora which enable different methods of research to be compared against the same prompts, scoring scales and annotation practices, have been developed (Shermis, 2014; Williamson et al., 2012). Benchmarks are not neutral, however, at the same time. They overrepresent some aspects of the rubric, some sets of students, and some types of writing, and underrepresent others, suggesting that the choice of benchmarks also influences the understanding of what counts as robust, transfer, and equity.

The most influential dataset in AES research has been the dataset of the Automated Student Assessment Prize (ASAP). ASAP created a shared framework of prompts, human scores and reporting guidelines that allowed for evaluation standardization and facilitated easier access to new research (Shermis, 2014). This represented a significant breakthrough in methodology since it enabled the researchers to compare systems under common conditions and paved the way for the adoption of quadratic weighted kappa and others as a common basis for measurement. ASAP, however, also encouraged optimization around a limited number of prompts resulting in a more stable progression than one might have found in new tasks, learners, or educational contexts. In this case, ASAP helped to increase the comparability of the field, yet it also reduced the empirical field's scope.

Scale was not the only thing ASAP contributed to, it was also structure. It has established a public benchmark that transforms AES into a reproducible research program instead of a handful of standalone systems (Shermis, 2014). That was relevant to the type of model developed but also to the types of questions that were asked. Improvements in score agreement can be misinterpreted as general improvements when a benchmark is becoming dominant, even if the system is calibrated to one benchmark family or score condition. One of the reasons why it is important not to equate benchmark performance with educational validity.

Evaluation culture of AES was also defined during the benchmark era. Researchers increasingly started to work on the predictive performance, with fewer studies testing if a model could be transferred to new populations or if it was fair across subgroups. That is, benchmark standardization allowed for easier comparison, but also provided incentives to optimize for the benchmark. Given this balancing act, strong performance on ASAP must be backed up by evidence of the classroom, institutional, and cross-prompt deployment of these items.

A few consequences of using only a handful of shared corpora can be noted. First, many datasets are task specific and hence represent the vocabulary, structure and logic of a narrow task rather than writing in general. Second, demographic information is often limited, restricting consideration of subgroup fairness and representation bias. Thirdly, models can seem more generalizable than they actually are, as the same benchmark is used again and again. In AES the limitations are significant since the educational application case includes real students, different genres, and varying instructional contexts.

Benchmark dependence results in interpretive claims also. An effective system on one corpus can be ineffective on another corpus even with other examples, grade levels or rhetorical expectations (Phandi et al., 2015). This presents a conflict between re-creability and realism. Validity has to be demonstrated through reproducibility of the data sets, while educational validity demands the ability of a model to perform outside the benchmark setting. The simpler the benchmark, the easier it is to compare systems and the more difficult it will be to make solid claims regarding fairness or portability.

For the field of fairness research, the design of datasets is particularly impactful. When a corpus is small and has little demographic information, it is hard to determine whether or not the scoring models act differently with respect to the different groups of learners. Furthermore, if training labels perpetuate inequities, benchmarked systems can further exacerbate inequities, rather than address them (Williamson et al., 2012). That's why benchmark choice is not a technical detail, but a substantive methodological decision.

Newer collections, such as PERSUADE, have brought the field further into a new direction for more complex analysis: the provision of more educationally representative argumentative texts and demographic metadata. These datasets can facilitate more robust research in the areas of discourse structure, subgroup variation, and fairness portability, as well as take research beyond a single benchmark optimization. In addition to their size, they are valuable in providing further analytical tools to examine validity and equity in AES. The table 3 below summarizes the main benchmark tradition and its implications for AES research.

Table 3. Benchmark datasets and their analytic role in AES

Dataset	Scope and Task Type	Approx. Size	Demographic Metadata	Main Strengths	Main Limitations	Review Use
ASAP	Competition benchmark for prompt-specific AES.	Large, prompt-based.	Limited.	Standardized evaluation and reproducibility; catalyzed benchmark research.	Narrow prompt coverage; limited fairness analysis; strong prompt specificity.	Baseline benchmark for early AES and model comparison.
ASAP 2.0	Expanded benchmark emphasizing more authentic writing tasks.	Larger than ASAP.	More useful than ASAP, but still limited for full fairness analysis.	Better source-based and argumentative coverage; improved ecological relevance.	Still benchmark-oriented; not a universal fairness benchmark.	Used to discuss improved transfer and more realistic task settings.
PERSUADE 1.0	Argumentative writing corpus with discourse-	Large.	Rich demographic metadata.	Supports fairness-aware analysis, argument	Still U.S.-centered; demographic coverage	Key corpus for fairness and

	level annotations.			mining, and discourse-sensitive modeling.	not universal.	transfer discussion.
PERSUADE 2.0	Expanded essay corpus with more prompts and demographic variation.	Very large.	Rich and useful for subgroup analyses.	Stronger support for fairness, robustness, and prompt variation studies.	Still insufficient for global multilingual comparison.	Central dataset for modern LLM-based fairness analysis.
ELLIPSE	Corpus focused on English learner writing.	Moderate to large.	Strong ELL metadata.	Valuable for multilingual and second-language fairness research.	Narrower population focus; not a universal AES benchmark.	Important for linguistic fairness and validity.
TOEFL11	Learner corpus for language analysis, not a standard AES benchmark.	Large.	Learner-language metadata.	Useful for broader language and transfer research.	Not designed primarily for essay scoring fairness.	Supports discussion of multilingual transfer contexts.

The table 3 is helpful because it demonstrates that benchmark datasets are more than just a source of training data. They determine the parameters of assessment, how kinds of fairness claims can be made and whether the field focuses on the accuracy in the prompt or robustness more generally. The table does not include the "generic" AES measures for school performance and/or student growth. The generic measures are included because they are the most important to consider when making transfer and fairness decisions, but they are also the ones most likely to limit those decisions if they were treated as representative of the AES as a whole.

Overall, benchmark datasets and corpora have been very important for the development of AES, but with a double edge. They facilitated comparative research and methodological standardization but also, and paradoxically, they limited the range of what the field observed of learners, of writing and of contexts. When a review is needed for ET&S, the benchmarks must be considered as more than resources; they must be epistemic infrastructures that help to form the evidence base for fairness, generalization, and educational validity.

5. Evolution of Automated Essay Scoring Models

AES models represent a trend towards more data driven representation learning in language technology instead of handcrafted feature systems. The transition resulted in increased explainability, fairness, and deployment concerns, while also enhancing the flexibility of scoring. The main issue that has been addressed across all generations is not whether machines can attain a level of accuracy similar to that of human scorers, but whether they can do so in a way that is sufficiently meaningful to learning, transferable, and equitable for meaningful assessment decisions (Attali & Burstein, 2006; Shermis, 2014;).

5.1 Traditional Machine Learning Based AES.

The traditional machine learning based AES systems were developed based on explicit linguistic and psychometric aspects. They used hand-made features like word count, lexical diversity, syntactic complexity, grammar and mechanics indicators, readability measures, discourse markers, and prompt-specific content overlap which were mapped to human scores using regression, support vector machines, random forests, ordinal classifiers, or ensemble methods (Attali & Burstein, 2006; Phandi et al., 2015). The primary strength of their approach was its interpretability, – that is, a connection between the scoring logic and the rubric dimensions and justification of scoring in operational settings.

But it was a limited sort of transparency. The concept of feature engineering always implied some beliefs about what constitutes “good writing” and these beliefs were more constrained than the concept of AES itself. In general, traditional systems were found to be most effective when there was a high level of stability in prompts, genre and scoring rubrics, however, they did not transfer well when these were altered (Phandi et al., 2015).

They also are at risk of using shortcuts like length or surface style which can be associated with demographical patterns, and not with the intended writing construct (Perelman, 2014). In this regard, early AES set the tone for the subsequent analyses regarding the operational logic and concerns of fairness.

5.2 Deep Learning-Based AES

Deep learning transformed AES by taking the responsibility of human feature designer and delegating it to the model. CNNs, RNNs, LSTMs, and attention-based architectures were able to enable systems to learn distributed representations directly from essays, thereby decreasing the reliance on manually engineered predictors (Alikaniotis et al., 2016; Taghipour & Ng, 2016; Dong et al., 2017). It was a significant methodological development because it expanded the range of the text patterns that were able to be measured, such as local syntax, sentence structure, and longer-range dependencies.

The educational value of deep learning was that it could model discourse and meaning at least in principle in more flexible ways than feature-based systems. They were helpful in support of some types of writing such as argumentative and source-based writing, where organization and reasoning are as important as surface correctness (Nadeem et al., 2019; Nguyen 2018). Despite this, deep learning hasn't solved AES's fundamental issues. Improvements in comparison to simpler baselines were frequently small, and the cost of the computation increased and transparency was reduced . Deep learning enhanced representational power, but raised the same concerns with respect to fairness, prompt dependence, and cross-context robustness.

5.3 Transformer-Based AES

The next pivotal change was to introduce contextualized pretraining to essay scoring with transformer-based AES.The next great leap forward was the introduction of contextualized pretraining for the essay scoring with transformer-based AES. The use of relatively small labeled AES sets to fine-tune the model, with the benefit of knowledge captured from extremely large corpora (Devlin et al., 2019) was provided by models like BERT. This in principle should have made the system more lexical flexible, less brittle matching and increased the out-of-domain performance. Transformers also closer AES to the mainstream of NLP practice, making the topics like domain adaptation, style sensitivity and fairness not easy to overlook.

In actualities, the benefits were not consistent. Sometimes these transformer models were similar to or slightly better than earlier models, while other times they were significantly better in computation or with only marginal improvements for educationally relevant transfer. Those trade-offs are relevant in operational settings because schools and assessment organizations have to take into account the costs, latency, and reproducibility of their testing, as well as benchmark scores. Long-context variants (Longformer) tackled the limitation of 512 tokens imposed by standard transformers, particularly important for longer essays where truncating evidence that is relevant to the score could lead to a misrepresentation of the evidence (Beltagy et al., 2020;). Nevertheless, better context handling did not guarantee the fairness of subgroups and portability of the prompts and population.

5.4 Large Language Models in AES

Large language models are the latest advancement in the evolution of AES models. As one can prompt, instruction-tune, or fine-tune LLMs, they are particularly appealing in low-label and cross-prompt scenarios (Brown et al., 2020; Wang & Gayed, 2024). This has rekindled the hope that AES can transcend the realm of benchmark-specific training and greater adaptability cross-context. LLM-based AES is the best current case for scalability and flexible deployment in that regard.

The flexibility, however, brings new potential risk factors. The allure of prompted LLM scoring is that it seems flexible, though it is not clear how reliable, calibrated, or construct valid it is. However, fine-tuned models do not always outperform prompted models, but they heavily rely on the quality of the training data and the composition of the training labels (). LLMs also carry forward and reinvent fairness issues from past problems, not eradicating them. The demographic bias can be incorporated into the pre-training and the label bias can be reproduced during fine-tuning in training sets that are not balanced and in human scores that are not balanced (). As such, LLM's enhance AES and increase transparency and governance.

5.5 Comparative Analysis of AES Model Evolution

Figure 2 helps clarify this historical trajectory by showing that AES did not improve in a simple linear way.

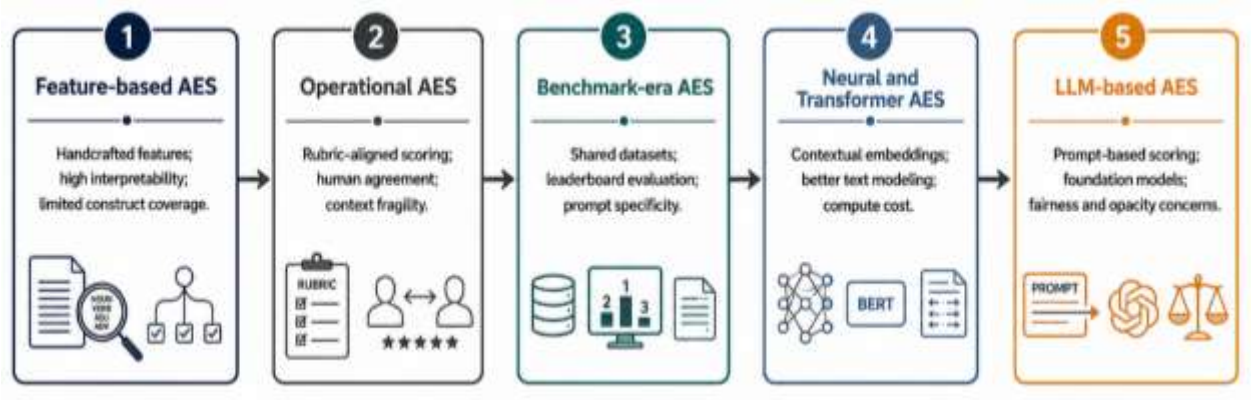

Figure 2 Evolution of Automated Essay Scoring

The figure should be interpreted as an idealized illustration of the tradeoffs between the four types of systems described: feature-based systems are more interpretable and align with the rubric; benchmark-era systems are more reproducible but limited in evaluation focus; neural and transformer systems offer greater representational power but less interpretability; LLM-based systems offer more flexibility but greater opacity and concerns regarding fairness. The value of the figure is that it describes model evolution as a series of shifts in the problem, not as necessarily an improvement.
Table 4 complements that visual narrative by comparing the major model generations side by side.

Table 4. AES Model Evolution Comparison

Model Generation	Core Approach	Typical Strengths	Main Limitations	Fairness Implications	Generalization Implications
Traditional machine learning AES	Hand-crafted features with regression, trees, or ordinal models.	Interpretability, rubric alignment, operational defensibility.	Feature dependence, limited construct coverage, prompt sensitivity.	Can encode shortcuts such as word count or style correlates.	Weak transfer across prompts and genres.
Deep learning AES	CNN/RNN/attention-based representation learning.	Better contextual modeling; less manual feature engineering.	Higher compute cost; weaker transparency.	May inherit label bias and subgroup disparities.	Better than traditional models, but still prompt- and corpus-sensitive.
Transformer-based AES	Fine-tuned BERT/Longformer-style encoders.	Contextual representation; stronger long-text handling.	Computational overhead; limited interpretability.	Fairness varies with corpus and training composition.	Improved transfer potential but still not robust across domains.
LLM-based AES	Prompted or finetuned large language models.	Flexible scoring, few-shot adaptation, potential feedback generation.	Prompt sensitivity, hallucination risk, opacity, calibration issues.	Demographic bias can arise from pretraining and finetuning composition.	Strong apparent transfer, but stability under shift remains uncertain.

This table is helpful because it helps surface the field's trade-offs within four key dimensions: technical approach, strengths, limitations, fairness implications, and generalization implications. Together, the figure and the table reveal the reasons for not being able to evaluate AES by predicting the performance. A model can be more

powerful in terms of computation but less trustworthy in terms of education, if it makes it less transparent, more expensive, or if it is not valid on prompt or population shift.

Thus the key takeaway from model evolution is by no means an improvement, but trade-offs. While traditional machine learning continues to be valuable in situations where interpretability and defensibility are paramount, deep learning offers gains in representation but limited impact on deployment risk, transformers are more capable of modeling context with increased deployment costs, and LLMs provide the highest degree of flexibility and magnify interpretability, bias, and reproducibility concerns. In the context of AES in education, sophistication should never be judged or equated with fairness, transfer, or validity.

6. Fairness and Bias in Automated Essay Scoring

Fairness is at the heart of AES, as it is not only a matter of marking text, but also of formulating educational opportunity, access and interpretation of writing by students. In the assessment context, fairness means that the meaning of scores and the comparability of scores should not change for students who are equally proficient on the test and that the scores should not be distorted by construct-irrelevant factors (American Educational Research Association et al., 2014; Williamson et al., 2012). In AES, the difference in scores can have an impact on placement, feedback, grading, and high-stakes decisions. This makes the AES a prediction problem, a validity problem, and an equity problem .

6.1 Understanding Fairness in Educational AI

Fairness in AES can be interpreted as the congruence between model scores and the purpose of the construct, and non-discrimination on irrelevant attributes. From a technical standpoint, fairness is a question of the distribution of model errors among groups, from an educational standpoint a question of whether those errors detract from the writing construct, and from an ethical standpoint, whether the system reinforces or exacerbates social inequities. Often the views are not congruent. That is why the focus has shifted from aggregate accuracy checking to subgroup-level validation, as a model may have good accuracy on average, but have systematically incorrect performance for certain subgroups. This change is significant because it's possible for a bias to be concealed by an average performance. Achievement on AES does not indicate that the model is fair for gender, race, socioeconomic status, language background, or disability status. Fairness is thus not a fixed, universal property but must be discussed in the context of the claims being made about the system. This is why the fairness literature is growing increasingly focused on multiple metrics and contextual interpretation.

6.2 Sources of Bias in AES Systems

Bias can be introduced at any point in the AES pipeline, and can occur beyond the model phase. Datasets and sampling bias arise when prompts, groups, or writing styles are overrepresented in the corpora, and others are underrepresented, leading to a more balanced distribution of the training data than of the actual educational population. Annotation bias may be introduced by human raters who vary their assessment of severity, who have a different assessment criteria for different groups, or who score based on a personal expectation of social norms. Representation bias can be a serious problem if a corpus is not representative of the target population, as the model will learn a distorted picture of what is typical or acceptable writing.

These sources of bias are all inter-dependent. Representational gaps can be exacerbated by skewed sampling, and biases in labels can obscure the gaps as legitimate ground truth. Thus, algorithmic bias is often a reflection of previous design decisions rather than a characteristic of the scoring model itself. This is the reason that the assessment of fairness in AES requires considering data and model behavior.

6.3 Demographic Fairness

Demographic fairness evidence for AES varies but is usually found to have systematic differences in error pattern by family of models. Gender, racial, and socioeconomic differences in feature-based, neural, and hybrid systems have been reported in small but nontrivial ways, with the direction and magnitude of the difference depending on the corpora and the metrics . This variability implies that fairness is not a characteristic of an AES model alone, but a relationship between the AES model, the dataset and the evaluation design.

Note that interpreting scores on the demographic scale is a bit tricky. Where a group difference is found, it does not necessarily indicate bias in the model as the difference could also be due to genuine variations in ability, opportunity, or writing preparation. The analytic challenge is to sort out construct-relevant variation from construct-irrelevant discrimination. In the last few years, LLM-based studies have confirmed this trend: Holmes et al. (2026) identified demographic bias in the human ratings of the PERSUADE corpus and reported that models trained on more balanced data were less biased by demographic factors, while models trained on fewer data

were more. While the result is positive, it does not rule out the question of whether the distribution of the labels is just.

6.4 English Language Learners and Linguistic Fairness

The issue of fairness is important for ELLs as AES can inadvertently grade language background instead of writing skills. Forms by multilingual and language-diverse writers can vary in terms of their developmental appropriateness or dialectal differences; automated systems might view these forms as errors. If that occurs, it means that construct-irrelevant variance is included in the score, which negatively impacts the validity of the score for ELL students. The nonstandard English features have been found to affect automated scores (Johnson & VanBrackle, 2012)and these shifts can pass through automated scoring pipelines.

Recent transformer and LLM studies propose that the problem is not just a matter of categorizing students as ELLs in the formal sense. Models may not perform as well on essays written by language minority groups and effect size may be different across prompt and scoring tasks (Yamashita, 2025; Yancey et al., 2023). This implies that AES fairness is not just about demographic parity. It also needs to consider linguistic diversity and dialect, as well as the validity of score interpretation across writing communities.

6.5 Fairness Evaluation Metrics

There are a number of fairness measures available for AES, and they serve different purposes. Differential Item Functioning is useful in psychometric applications where it can determine if students of similar ability but different group membership are given different results . Demographic parity is a test for an even distribution of outcomes among groups, but can be misleading if there are differences in base rates. Equal opportunity is more informative when the task is one of correct classification or of agreement with a reference label, when an inquiry is made as to whether true positive rates are the same across groups.

The main methodological challenge is the "fit". A metric should be commensurate with the educational claim being made. A psychometric metric is better than a raw parity check when referring to score comparability across groups. In the case of a claim on classification fairness, group-based error rates might be more helpful. AES fairness metrics are not the same, they are lenses, they highlight various bias signals.

6.6 Fairness in Transformer and LLM-Based AES

Fairness issues do not disappear with transformer and LLM systems, they shift in form. The models are pre-trained on general corpora which can have social and linguistic biases, and fine-tuning can either be a positive or negative effect depending on the nature of the training data . The additional variability of prompted systems stems from the fact that there are other factors that can cause a difference in the score output, such as the wording of the prompt, examples, and/or how the prompt is framed (Wang & Gayed, 2024). Usually, fine-tuned systems are more stable, but they also exhibit the characteristics of the population they are trained on.

This is to differentiate pretraining bias from task-specific bias. Using balanced training data can help to minimize disparity, but not prove construct validity or eliminate historical bias in human ratings. Hence, fairness in LLM-based AES is a data issue and also a governance issue. A general-purpose model, if not systematically monitored by subgroup auditing, can be a neutral model that produces disparate educational outcomes.

6.7. Current Limitations in Fairness Research

For AES, standardized benchmarks and consistent demographic reporting are still lacking, as are common protocols for cross-study comparison of fairness research. Most studies are based on a single corpus, on one demographic axis or on one metric, which complicates synthesis. Additionally, dataset constraints can impact: Some corpora may have limited subgroup metadata or a limited set of genres or even coverage of tasks, which can limit the robustness of a fairness claim.

Another constraint is that the study of fairness tends to come after the development of the model, as opposed to the design of the assessment. That's a bad thing, because fairness shouldn't be an afterthought, it should be a design requirement in AES. More broadly, educational AI bias work is supportive here, but AES is unique in that the score directly influences educational decisions. Therefore, fairness in AES should be associated with validity arguments, not just to statistical parity.

6.8 Synthesis and Key Insights

There are three conclusions in the literature. As the importance of fairness issues in AES has often been overlooked in high-stakes contexts, this is not an excuse to ignore them—first, fairness is a reality in AES, but it's

not a big one . Second, fairness is a multi-dimensionality, and thus there are no a priori defined metrics that capture all the relevant bias signals . Third, representative training data is one of the most promising approaches for minimizing demographic bias, but is not a panacea, as it may still contain historical inequities, and human ratings that can be biased .

The figure 3 should be interpreted in the context of a systems perspective on AES, where upstream decisions about what to sample and label influence what the model learns, and downstream decisions about modeling and evaluating influence how bias is expressed and how it can be detected. The framing is significant because it ties fairness to generalization and domain shift, rather than considering it a distinct issue. Figure 3 captures this logic by showing fairness as the product of data composition, model architecture, and evaluation design.

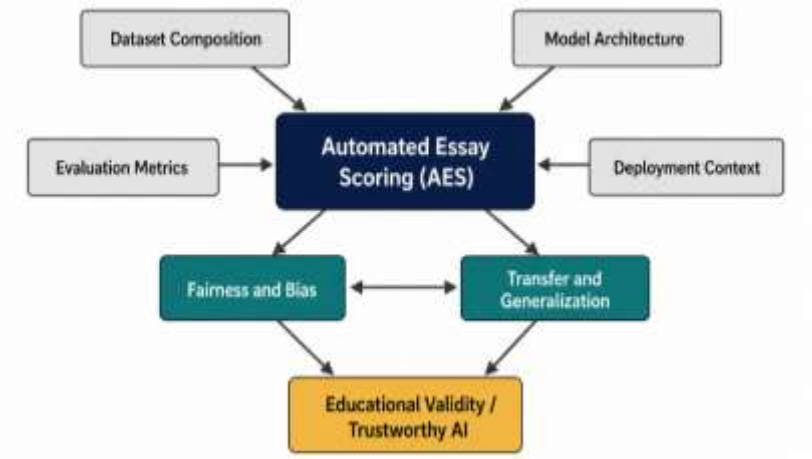

Figure 3: Fairness–Bias–Generalization Framework

The table 5 is helpful because it helps to clarify that findings about fairness in one study cannot be universally applied to the field. There are different patterns and they're different based on the corpora, the subgroup definitions, and the model designs. Table 5 summarizes the main fairness studies and shows how findings vary across model families, corpora, and demographic groups.

Table 5. Fairness Studies Summary

Model Family	Demographics Examined	Main Fairness Finding	Interpretation
Traditional AES	Bias in human ratings.	Biased labels can be learned by AES.	Training labels are a fairness risk, not a neutral target.
Feature-based, neural, hybrid	Gender, race, SES.	Different fairness metrics revealed different biases across models.	Fairness is multidimensional and model-dependent.
Longformer-based LLM-AES	Race/ethnicity, ELL, disability, SES, gender.	Balanced training reduced demographic bias; restricted training increased bias.	Representation bias matters, but balanced data is not a complete solution.
Prompted LLM	Racial descriptors	Identical essays scored differently when race cues were changed.	LLM scoring can be sensitive to construct-irrelevant social signals.
Prompted LLM	English learner background	Agreement with humans varied by language background.	Group effects can appear even without explicit demographic inputs.
Prompted LLM	Race/ethnicity, gender, SES	Different score patterns across subgroups.	Fairness effects are not uniform across demographic axes.
Finetuned LLM feedback models	ELL, SES	Lower accuracy for some groups when predicting effective writing.	Bias can emerge in auxiliary tasks as well as scoring.

The table 6 and the evidence in this section demonstrate that, in AES, methodological pluralism is essential; that is, there is a need for multiple metrics because multiple fairness questions are being asked. In the next section, we turn to the issue of transferability and domain shift, which are closely related because a model that exhibits changes in its behavior across prompts or populations can also show changes in its fairness profile. Table 6 compares the fairness metrics used in AES research and clarifies what each measure can and cannot show

Table 6. Fairness Metrics Comparison

Metric	What It Measures	Strengths	Limitations	Best Use in AES
OSA (Overall Score Accuracy)	Whether AES scores are equally accurate across groups compared with human scores.	Captures group differences in error magnitude.	May miss bias when overall accuracy is similar.	Useful for detecting broad error disparities.
OSD (Overall Score Difference)	Whether AES and human scores differ systematically across groups.	Sensitive to directional score differences.	Can be confounded by true proficiency differences.	Useful for subgroup over- or under-scoring.
CSD (Conditional Score Difference)	Whether group differences remain after controlling for human score.	Best aligned with fairness conditional on ability.	Requires reliable human scores and careful modeling.	Strong option for bias evaluation in AES.
DIF-style analysis	Whether equally proficient students from different groups receive different outcomes.	Psychometrically grounded; supports validity arguments.	More complex to implement in AES settings.	Best for fairness claims linked to assessment validity.
Demographic parity	Whether outcomes are evenly distributed across groups.	Easy to communicate.	Can be misleading if base rates differ.	Limited use in high-stakes AES.
Equal opportunity	Whether true positive rates are equal across groups.	More informative than raw parity.	Depends on target definition and labels.	Useful for classification-like AES tasks.

7. Generalization and Transfer Learning in Automated Essay Scoring

A major drawback of AES is that good benchmark scores do not always lead to valid scores when prompts, populations, genres or institutional settings change. The question of generalization deals with the ability of a model to also generalize the data from one essay collection to another, whereas the question of transfer learning is whether knowledge acquired from a previous task or essay collection can be reused in a new task or essay collection with only a few target labels. These are not just technical nuisance problems, but rather questions of whether AES can shift from a string of positive grades on benchmark tests to a legitimate role in education.

7.1 Understanding Generalization in AES

Generalisation in AES is defined as a model's capacity to make correct scores on essays written on new prompts, topics, genres, populations or assessment settings without losing validity. This is a more rigorous criterion than benchmark accuracy, as many systems are trained and tested on the same data set or prompt family, which can result in replying to the prompt in a specific way. A model can thus show good performance on the cross-validation set and a poor performance on a new educational setting. Generalization is, therefore, a key validity problem not just a machine learning metric. This is particularly important in the school and testing context, in which essay tasks vary both within and across schools and schools. A useful scoring model must be applicable to the conditions in which it would be used, even for a high score, if it fails to apply to those conditions it is not useful. This review thus considers the test of whether AES can be used to maintain the meaning of the score in the context of the training environment to be the test of generalization.

7.2 Cross-Prompt Generalization

One of the major problems for AES is still cross-prompt generalization. Prompt-specific systems typically acquire some combination of writing-related and prompt-specific features, such as the vocabulary in the prompt, discourse structure, and task requirements (Attali et al., 2010. This means that models often fail when given a new prompt that they have never been previously exposed to. The lack of training exposure is not the only

problem, but rather the model might have learned the structure of the prompt rather than the writing quality of the prompt. Previous efforts to minimize this dependence were made using less topic specific feature designs, which had a limited impact. Phandi et al. (2015) demonstrated the effectiveness of correlated linear regression in the case of prompt adaptation in the domain, Despite this, these techniques did not close the scores between prompt dependent and prompt independent. Other more recent transformer models are also found to be sensitive to the prompt structure used for training, indicating that the cross-prompt generalization problem is as much an evaluation problem as a modeling one.

7.3 Transfer Learning in AES

The goal of transfer learning is to avoid having to train large sets of prompts from similar tasks, corpora or pretrained language models, by reusing representations, parameters or knowledge. This encompasses domain adaptation, cross-dataset transfer, representation reuse, and parameter-efficient fine-tuning in AES. The usefulness is clear: for each new prompt or for each new school, having an expert in the classroom to score the outcomes is costly and sometimes impossible, particularly in low resource countries. There are several ways that the transfer can take place. A model can be fine-tuned from a source prompt into a sparsely labeled target prompt, or it can be fine-tuned using a pretrained encoder (such as BERT or Longformer) model on a related scoring task. The next step, cross-dataset transfer, seeks to determine if a model trained on one benchmark can be transferred to another. Previous studies indicate that transfer can lead to data savings and, in some cases, better quality performance, provided that there is substantial similarity between source and target genres, rubric logic, and characteristics of the populations in the two domains (Phandi et al., 2015). The potential use of transfer learning is therefore encouraging, though not certain.

7.4 Cross-Corpus and Cross-Domain Evaluation

Cross-corpus evaluation is the type of evaluation that is most useful for understanding the limits of AES portability. There are differences between the prompts, essay length, genre, age group, scoring scale and the composition of demographics between ASAP, ASAP 2.0, and PERSUADE. These are not superficial differences; they alter the conditions of the construct which are used to generate scores. A system that works for short answers from one source may not work for longer persuasive essays, and a model which fits one criterion-based rubric might not fit the other. Decreases in performance are, therefore, indicative of a change in the construct in these contexts. It is particularly emphasised in recent corpus research. PERSUADE aims to enable more representative analyses in education than the previous benchmarks, and Holmes et al. (2026) find that even in longer context transformer models, representation and population composition have an impact on transfer strength. This implies that transfer performance may not only be influenced by the architecture of the model, but also by the type of prompt, length of the text, the alignment of the rubrics, and the coverage of the demographies. Many claims of scalability of AES are based on within-corpus evidence, as there are relatively few studies that test across corpora.

7.5 Generalization in Transformer and LLM Based AES

The technical foundations for transfer have been enhanced through the creation of transformer and LLM-based systems, however, these systems have yet to address the generalization problem. They have more language knowledge from their pre-training, which may be useful when the data for the target language is limited. One source of distortion is also eliminated by longformer style models, as they do not have to truncate long essays. LLMs take this one step further by allowing them to score a few or no shots, leading to the illusion of good transferability. But evidence is inconclusive. Wang and Gayed (2024) found positive findings for argumentative writing when using LLM-based scoring, yet it hasn't been resolved what should be done for prompt sensitivity, calibration, and aligning with human standards. While finely-tuned models can be more stable than prompting-only models, they also require good and representative data. That is, an LLM could be generalized linguistically but not psychometrically. That's important in an educational context.

7.6 Challenges Limiting Generalization

There are a number of types of shifts that restrict portability of AES. Dataset shift happens when the duration of essays, topic, or demographic composition of the training set is different from the operational set. Prompt shift refers to a change in rhetorical demands and/or familiarity of content that is prompted. Another type of distribution shift is temporal, which occurs, for instance, when the curriculum changes, or when the generative AI is used in writing. The rubric, rater or adjudication system may vary between corpora, resulting in annotation mismatch, which poses another layer when you are scoring the rubrics, scores, or adjudication. These changes are important because the failure to generalize is not uniform. A model that does not work well for some institutions, language backgrounds, or learner groups may be less appropriate, highlighting both a technical and a fairness concern. When the meaning of scores varies from context to context, transfer can be statistically

appropriate but educationally inappropriate. Therefore, it is important that transfer evaluation is considered in the context of bias evaluation, rather than as a standalone issue.

7.7 Educational Applications of Transferable AES Systems

The transferable AES systems are appealing as they offer scale, faster adoption and reduced reliance on local scoring data. It can reduce the barriers of schools and testing programs, and provide a more viable automated feedback option in those environments where developing a custom model is not feasible. But technical portability does not equal educational value. A model can be transferable and yet be misaligned with a local curriculum, with a local population or with the scoring system. The equity issues are significant as well. If the source benchmark is not representative of the target group, then using the transfer method can exacerbate disparities when the target group is underrepresented. AES deployment is thus a cross-context process, and it should be considered as a basic requirement for validity, rather than an optional extra.

7.8 Synthesis and Key Insights

The results of the literature review indicate that AES generalization is still not as strong as what may be suggested by the benchmark results. There are promising approaches like domain adaptation, pretrained encoders, and long-context models, but evidence supporting them in operation is limited both across corpora and populations. The main obstacles include domain mismatch, inconsistencies in annotations, absence of cross corpus benchmarks and problems of linking the transfer performance to educational validity.

Table 7 summarizes the main studies on AES generalization and transfer learning. The table is significant because it illustrates that there are not just one problem and just one solution – the various studies handle the issues of prompt transfer, corpus transfer, and parameter reuse in a variety of ways. It also contributes to differentiate methods that promote better performance on the benchmark from those that actually enhance its portability and validity across education settings. The main idea is that generalization does not exist outside of fairness. A model that doesn't transfer well between prompts/populations will also likely have inconsistent educational impacts. In the following section, we thus immediately focus on the connection between fairness, demographic variation and domain shift.

Table 7. Generalization and Transfer Learning Studies

Transfer Focus	Main Method	Key Finding	Limitation
Cross-prompt/domain transfer	Correlated linear regression / domain adaptation.	Prompt relationships can be exploited to improve transfer.	Still limited by prompt overlap and corpus similarity.
Prompt-independent AES	Two-stage deep neural model.	Prompt-independent results improved but did not eliminate the transfer gap.	Performance remains lower than prompt-specific models.
Transformer transfer and efficiency	BERT fine-tuning.	Transformers can match baseline performance but with major compute cost.	Robustness gains are not sufficient to justify complexity in all settings.
Transfer and fairness under demographic shift	Longformer-based finetuning on balanced and restricted datasets.	Representation bias affects fairness outcomes under transfer.	Cross-domain transfer and fairness portability remain unresolved.
Discourse-aware modeling	Neural/discourse-aware architecture.	Discourse information improves AES in some settings.	Generalization remains corpus-dependent.

8. Fairness Under Domain Shift

Fairness is one of the biggest open questions in education that remains unanswered – when AES models are applied out of their training environment. The present section claims that fairness and domain shift are not two different problems: A model can look fair in one context and automatically be systematically biased in another context when a prompt, institutional or student population changes. In AES, that is important because fairness is a concept that goes beyond prediction of scores, it is related to the concept of educational opportunity.

8.1 Understanding Domain Shift in Educational AI.

Any form of systematic discrepancy between the data used during model training and the data deployed to the model on which subsequently is called domain shift. That mismatch may be at the level of grade, essay type,

scoring guideline, institution or historical time frame of information gathering. With a large gap, model behavior can change in ways that are hard to predict or audit, particularly if there is also a change in the composition of subgroups . That is what is making domain shift a far more than a mere technical annoyance. The true deployment environment is not necessarily the same as the testing environment, and a seemingly stable model might fail in deployment. Some of the fairness concerns are particularly grave if the training and deployment populations are different in terms of language backgrounds, demographic composition, instructional context, etc.

8.2 Types of Domain Shift in AES

here are several types of shift which change in AES. Prompt shift is when a scoring model is used to grade an essay written on a different writing prompt that has different rhetorical requirements, expectations for the topic, or length requirements. Topic shift occurs when the content is familiar to some students and not others, thus presenting a challenge for some groups. When corpora vary in rubric structure, the annotation practice, or in their demographics, then there is a dataset shift. There are other changes that are also important. A demographic shift is a change in the target population and the training population, which is significant in Holmes et al.'s (2026) work. Linguistic shift refers to the variation in either dialect or syntax or discourse structure across groups; institutional shift refers to differences in curriculum, the ways instruction is carried out, and the ways that items are scored locally. This is because the changes tend to happen simultaneously, making it hard to determine which is the primary reason for a loss of fairness. Table 8 shows various examples of studies on fairness under such types of shift. The table 8 is also helpful since it reveals that fairness portability has been tested beyond the cross of demographic groups, and that it is tested across a variety of mismatch types. It also shows that the evidence is not consistent; some studies address population shift, other studies address prompt sensitivity, other studies address language background prompt sensitivity, and others address social cue effects. The results of these studies indicate that the interaction among model family, corpus design and context of the shift is crucial to fairness.

Table 8. Domain Shift and Fairness Portability Studies

Shift Type	Key Result	What It Shows About Fairness Portability
Demographic shift	Balanced training reduced demographic bias; restricted training increased it.	Fairness depends on population composition and is not automatically portable.
Model-family and corpus shift	Different models displayed different fairness patterns under the same data.	Fairness findings do not generalize cleanly across architectures.
Language-background shift	Agreement varied across learner-language groups.	Fairness may vary by linguistic subgroup under the same model.
Prompt/social-cue shift	Racial cues changed scores for identical essays.	Fairness can be altered by construct-irrelevant contextual signals.
Prompt/domain shift	Transformer models still struggled with portability despite contextual modeling.	Generalization and fairness remain linked but unresolved.

8.3 Fairness Portability Across Domains

Fairness portability means that a model's fairness properties are preserved when the model is transferred from one context of assessment to another. This property is often taken as a given in the AES literature, but is not usually tested directly. Even if a model is fair on the development corpus it can be unfair on another new corpus with different demographics, prompts or achievement distributions . That is, fairness is not just whether a model is fair at a specific location but whether that fairness is transferred. It has direct validity implications. However, if meaning is different based on context, the fairness of the model would not be considered to be established if it was found to be acceptable in the development time evaluation. This is particularly relevant for LLM-based AES as it can be implemented rapidly from one institution to another without having to be re-audited for the target population. The conditions of fairness portability, therefore, are essentially technical conditions as well as assessment-validity conditions.

8.4 Evidence from AES Literature

The best evidence of fairness under shift that exists is from controlled studies, where the composition of the training data is varied. Holmes et al. (2026) trained different subsets of the PERSUADE corpus on Longformer based AES models and, showed that restricted training led to greater bias than balanced training. The most significant errors were from models that were demographically limited, such as higher prediction error and over-prediction in some settings of the ELL essays. This means there are no guarantees that subgroup fairness will be

transferable, and that it will vary from population to population. This is consistent with previous research. Warr et al. (2024) showed that modifying a social signal (racial cues) in otherwise identical essays had an impact on ChatGPT-based scoring, indicating that ChatGPT is sensitive to construct-irrelevant social signals. some student groups performed better than others in LLM-based feedback models and found that fairness results are different across different models even when using the same fairness measures. The general trend is that "fairness" doesn't simply transfer across models and contexts.

8.5 Domain Shift in Transformer and LLM-Based AES

Both transformer- and LLM-based systems exacerbate the domain-shift problem in two ways. First, the pretraining data can contain social and linguistic biases, which means that the unfairness can be in place prior to any AES-specific fine-tuning . Second, the subtle socio-demographic cues in vocabulary, discourse, or style may also be present for large models without being targeted. Yancey et al. (2023) reported differences in agreement by GPT-4o across demographic groups and by Yancey et al. (2023) across ELL language background, respectively. The results indicate that there is no such thing as eliminating distributional mismatch even among larger models. Indeed, larger models might be able to capture more subtle patterns for each group in their training datasets. This implies that if it is to be used for educational deployment, generic benchmark evaluation is not sufficient. A fairness audit is necessary for a specific domain and/or population before a model can be used in practice.A fairness audit is required for a specific domain and/or population prior to use of a model in practice.

8.6 Challenges in Measuring Fairness Under Domain Shift

Very little work has been done so far on measuring fairness in the presence of domain shift, as most work considers a specific corpus or a small number of demographic groups. This is difficult to compare results across contexts and hard to know if the difference in fairness is a result of the model, data, or rubric. The reporting is also not comprehensive in many AES studies, particularly if any information is missing or poorly defined about the demographics . Without common cross-corpus protocols, there is no stable base for the fairness portability claims. It is not just a problem of methodology; it's a structural problem. Fairness comparisons can easily be misleading if the corpora have different scoring rules, different coverage or differences in the quality of the annotations. To make any meaningful generalizations about fairness outside of the original training environment, the field must make progress toward more standardized cross-domain evaluation.

8.7 Toward Fair and Robust AES Systems

A few approaches are proposed to be promising in terms of fairness in the domain shift scenario. Absolute training data balance is one of the most powerful levers and Holmes et al. (2026) demonstrate that demographic balance decreases the bias in LLM-based AES. In situations where prompt or institutional change is moderate, domain adaptation can be useful, particularly if there is adequate structure shared between source and target corpora (Phandi et al., 2015). Data augmentation can benefit underrepresented groups or writing styles and fairness-aware objectives can benefit of minimizing differential error across groups, but can sometimes reduce raw accuracy . All of these strategies are needed together. Fairness must be enforced during deployment and not just at deployment time as shift is dynamic. A fair system at its inception can become unfair due to changes in population, curriculum or institutional culture. As a consequence of responsible AES use, continuous auditing is a key component of this.

8.8 Synthesis & key insights

The conclusion is that the concept of fairness is context dependent in AES. Now, it's not an inherent characteristic of the model itself, but rather a characteristic of the model in a specific data and deployment environment. Shifts, demographic shifts, and institutional differences can all influence patterns of errors and meaning of scores within groups. Hence, the issue of fairness portability is still one of the largest gaps in the field.

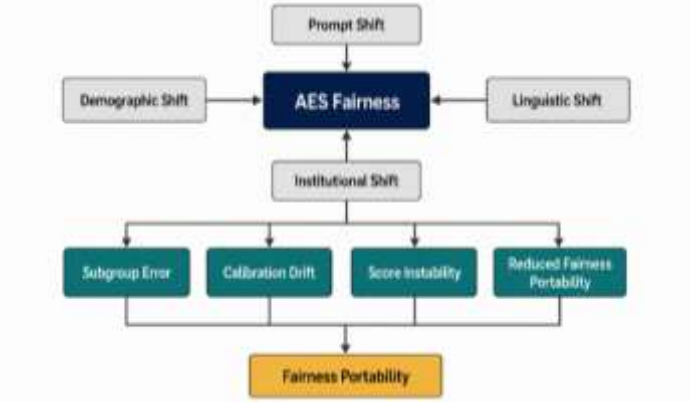

Figure 4: Fairness Under Domain Shift Framework

Figure 4 makes this argument visually by showing domain shift as a cluster of changes in prompt demands, demographic composition, language background, and institutional context. The figure should be understood as a systems view: each shift can alter calibration, subgroup error, and score stability. Its purpose is to show why one-time fairness evaluation is weaker than ongoing fairness governance.

9. LLM-Based AES and Emerging Trends

Large language models (LLMs) mark a revolution in AES as they not just alter the way essays are evaluated, but the very nature of what AES ought to be. The problem of fairness is not robust to domain shift, and in this sense, LLMs extend the problem: they offer new ways to scale, but also new ways in which they become opaque, prompt sensitive, and have new governance requirements .

9.1 Emergence of Large Language Models in AES

The transition from feature based systems to transformer based systems and then LLM based systems started with the introduction of the foundation models like BERT and GPT which have been replacing the engineering of the prompt with pre-trained general purpose representations (Devlin et al., 2019; Brown et al., 2020). The two most prevalent methods in AES are finetuning encoder-based models like Longformer on scored essays and prompting generative models with rubric-aligned instructions to generate scores without having to train a task-specific model (). These are not the same in both terms of fairness and generalization and are both currently of practical interest.

9.2 Capabilities of LLM-Based AES

With AES, LLM-based tools can enliven the scope of automated scoring. These models are able to evaluate more reliably for coherence, argument structure and rhetorical sophistication than systems based on surface features (Nadeem et al., 2019). They also provide less need for task-specific labels in few-shot or zero-shot scenarios and can be used to provide formative rather than summative feedback (Yancey et al., 2023). This makes them appealing to use in the classroom and as a writing aid. But these gains are fraught with significant restrictions. Generative LLM can generate plausible but unreliable results and the scores can vary significantly based on the prompts given . They are also less transparent than feature-based AES, as they are not as easy to explain what they point to. Besides, the computational and infrastructure requirements of big models could render their operational use challenging in many educational contexts.

9.3 Challenges of LLM-Based AES

fairness issues are not addressed by LLM-based AES, but rather in a different manner. Pretraining can incorporate representation bias prior to any fine-tuning for essay scoring; and fine-tuning can reinforce historical bias if the training set is imbalanced or the labels are themselves biased. Empirical research indicates that differences in how students score on an essay can occur for students based on their ELL status, their race/ethnicity, their gender, and their socio-economic status, as well as differences in how students score when only social cues are varied in otherwise identical essays (Warr et al., 2024; Yamashita, 2025). The findings make demographic fairness a “must have” and not an “extra” in terms of validity. Reliable educational AI needs to go beyond the average accuracy. It needs to be transparent, and have someone watch it at all times, accountability for any misjudgements in the scoring process, and it must be reviewed at all times for fairness . While some disparities can be lessened using balanced training data, the data may still reflect some inequities in the original human ratings, and therefore cannot guarantee fairness.

9.4 Fairness and Bias in LLM-Based AES

The most potentially fruitful paths are represented by explainable AES, fairness-aware training, human-AI collaboration, and adaptive assessment. Explainable systems are designed to give scores as well as rubric based rationales, which helps to minimize the gap between feature-based and neural in terms of interpretability.Explainable systems are expected to generate scores and rubric based rationales, which helps to minimize the gap between feature-based and neural in terms of interpretability. Adversarial debiasing, contrastive objectives, and intersectional balancing are some fairness-aware training methods. While teacher judgment remains the cornerstone of assessment, AI-powered tools like LLMs enhance their ability and support them in making decisions. There is also a growing interest in two other approaches to better align models to the writing of students: educational foundation models and multimodal AES.

9.5 Synthesis

The table is useful because it shows that the field is not converging on a single design pattern: finetuned models tend to be more stable, while prompted models offer more flexibility, but both remain vulnerable to bias and prompt effects. The broader conclusion is that LLMs make AES more capable, but also more governance-intensive. Table 9 summarizes the main LLM-based AES studies and the key opportunities and risks identified in the literature.

Table 9. LLM-Based AES Studies Summary

LLM Approach	Strengths Reported	Main Risks Reported	Fairness Finding
Finetuned Longformer AES	Strong performance with balanced data.	Representation bias with restricted training.	Balanced data reduced bias, but did not eliminate fairness concerns.
Prompted GPT-based AES	Flexible scoring without task-specific training.	Language-background sensitivity.	Agreement varied across learner groups.
Prompted ChatGPT scoring	Strong apparent adaptability.	Prompt and social-cue sensitivity.	Identical essays received different scores when race cues changed.
GPT-4o-based scoring	Strong overall scoring performance.	Demographic disparities across subgroups.	Bias patterns varied by subgroup.
Finetuned LLM feedback model	Useful for feedback tasks.	Lower accuracy for some learner groups.	Fairness issues extend beyond holistic scoring.
Finetuned LLM-AES	Balanced training improved fairness.	Limited evidence on broader portability.	Training data composition strongly affects fairness.

10. Research Gaps and Future Research Agenda

It is now well established in the AES literature that validity is not just about fairness, bias and generalization—it is a fundamental component of validity. Despite this fragmentation, however, there is little consensus on how to define fairness, limited transfer studies, and accelerating innovations that rely on LLM are outpacing the tools available to responsibly direct them . The real problem is thus not just to develop more sophisticated models, but to develop a research program that can provide a basis for determining when AES is fair, portable, educationally useful, and trustworthy.

10.1 Major Research Gaps in AES Literature

There are a number of common weaknesses in the existing evidence base. Fairness metrics are not agreed upon, and therefore the same model could be considered fair for one metric and not fair for another . Intersectional analysis and fairness portability is difficult to study due to the often incomplete demographic reporting (). Many corpora are also too small to provide stable data for underrepresented learners, due to the small number of students in subgroups. Very few studies have performed cross-domain evaluations, there is inconsistent evidence for reproducibility due to missing details of preprocessing and prompting, and studies of educational impact are limited for longitudinal evidence (Gao et al., 2024).

10.2 Gaps in Generalization and Transfer Learning Research

There are some strides in generalization research, yet they are very narrow and very under-specified for education. Despite the simultaneous shifts in prompt, institution, genre, and population in real deployment, the cross-prompt evaluation approach remains the predominant one (Phandi et al., 2015). Less studied is the issue of cross-corpus transfer, where corpora may differ significantly and it is hard to distinguish between the effect of transfer and the effect of a difference in construct, rubric and/or population. Most significantly, there is still no agreed-upon methodology for measuring fairness portability and a model that is fair in one context may not be fair in another ().

10.3 Gaps in LLM-Based AES Research

LLM-based AES opens new opportunities but also introduces new unresolved problems. There is still a lack of explainability, as the scores can be due to stylistic regularity or pre-trained connections and do not necessarily correspond to the construct-relevant quality of the writing . Additionally, reliability has not been well explored, particularly in response to changes in wording, examples, decoding options, or API changes (Yoshida, 2024). The issue of hallucination does not only apply to feedback generation, but also to rationales, as plausible explanations may be wrong. Moreover, proprietary systems do not allow transparency, as training data, cycles and safety filtering are normally not viewable for researchers.

10.4 Future Research Directions

There are five directions for future research. First, fairness-aware AES should be fairness-by-design, using balanced sampling, intersectional reporting and highlighting the difference between measurement bias, representation bias and historical bias . Secondly, generalization should be taken beyond prompt-level testing to cross-corpus benchmarking and fairness portability testing with distribution shift (Phandi et al., 2015). Third, LLM-based AES needs to focus on explainable scoring and human-AI collaboration to ensure that AI augments, but does not replace, professional judgment. Fourth, the development of the dataset should be an independent research area, with multi-lingual, cross-cultural and fairer bodies of data that capture a wider range of learners.

Fifth, transparent governance, accountability, and continuous monitoring of AI systems deployed after training are essential for its trustworthiness, particularly in high-stakes assessment settings (Gao et al., 2024).

The table10 is significant as it indicates that the deficiencies of the field are not simply technical problems, but they are connected structural issues that have to do with evaluation, corpus design, reproducibility and educational evidence. These gaps help to understand why there is still uncertainty in the literature in making cumulative claims about fairness and portability.

Table 10. Research Gaps Identified Across AES Literature

Research Area	Identified Gap	Current Limitations	Future Opportunity
Fairness evaluation	Inconsistent metrics and definitions.	Findings are hard to compare across studies.	Standardized fairness reporting and benchmarking.
Demographic reporting	Limited subgroup metadata.	Intersectional analysis is often impossible.	Richer corpus documentation and metadata standards.
Dataset representation	Narrow coverage of populations and genres.	Underrepresented groups remain analytically invisible.	More inclusive and multilingual corpora.
Benchmark design	No universal fairness benchmark.	Limited cumulative progress.	Shared fairness-oriented benchmark suites.
Cross-domain evaluation	Insufficient cross-corpus studies.	Transfer claims remain weak.	Multi-corpus, multi-population evaluation protocols.
Reproducibility	Incomplete reporting of prompts, splits, and preprocessing.	Results are difficult to replicate.	Transparent reporting checklists for AES studies.
Longitudinal evidence	Few studies of learning and institutional impact over time.	Educational consequences remain unclear.	Long-term studies of feedback, teaching, and learner outcomes.
LLM-based AES	Weak explainability and benchmark adequacy.	Prompt sensitivity and opacity limit trust.	Explainable and auditable LLM-AES evaluation.

The table 11 is very useful to turn the review into actionable research program: fairness-aware modelling, improving the transfer protocols, explainable LLMs, better corpora, governance mechanisms are not independent research programs. If AES is to truly become more accurate, it is imperative that this program be implemented in order for AES to be more valid and trustworthy when used in actual education.

Table 11. Proposed Future Research Agenda

Theme	Research Direction	Expected Impact	Educational Relevance
Fairness-Aware AES	Develop fairness-by-design methods, inclusive evaluation protocols, and ability-conditional audits.	Reduces bias before deployment.	Improves validity and equity in high-stakes scoring.
Bias Mitigation	Use balanced training, adversarial methods, and fairness-aware objectives.	Lowers subgroup disparities.	Supports fairer automated grading and feedback.
Generalization and Transfer Learning	Build cross-corpus benchmarks and fairness portability tests.	Improves robustness under shift.	Enables reliable use across schools, prompts, and populations.
Domain Adaptation	Use robust adaptation to new prompts and institutions.	Reduces performance loss in new contexts.	Supports scalable classroom and assessment deployment.

LLM-Based AES	Build explainable LLMs, educational foundation models, and human-AI collaboration systems.	Increases trust and interpretability.	Maintains teacher judgment while using AI efficiently.
Dataset Development	Create multilingual, cross-cultural, fairness-focused corpora.	Expands empirical coverage.	Better reflects diverse learners and writing practices.
Trustworthy Educational AI	Strengthen transparency, accountability, oversight, and governance.	Enhances responsible deployment.	Essential for safe large-scale educational assessment.

The figure 5 should be read as a process model rather than a list of suggestions: it shows that fairness and generalization must be built into the system from the beginning and then monitored after deployment. That framing matters because it shifts AES from benchmark optimization toward responsible educational practice.
The main conclusion is straightforward: future AES research must connect model development, dataset design, subgroup evaluation, cross-domain testing, and governance. Without that integration, progress in LLM-based AES may outpace the field's ability to validate and control it.

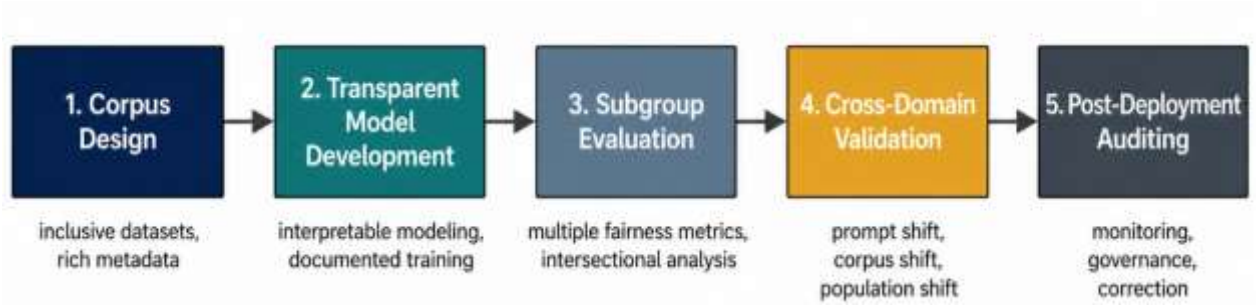

Figure 5 Future Research Roadmap for Fair and Trustworthy Automated Essay Scoring

11. Conclusion

The evolution of automated essay scoring has been a series of compromises or steps. At the same time, there were improved concerns about transparency, fairness, validity, and educational use of early feature-based systems, and also of neural, transformer-based, and LLM based systems that increased the representational power and decreased the reliance on handcrafted features, as demonstrated by Williamson et al., 2012. Throughout this story, benchmark datasets and shared corpora have driven research capabilities in the field: what it measures and what it claims. They simplified the process of comparison and model building, but they also narrowed the field for progress to be evaluated primarily by the limited prompts and corpora.

One of the main findings of the review is that the concepts of fairness and generalization go hand in hand. AES can amplify or reproduce inequities in training data – particularly those related to prior educational bias – as well as cases of weak representation within subgroups . Concurrently, good performance in one prompt/corpus does not imply good performance in the other. However, cross-prompt, cross-corpus and cross-domain evaluation are still in their infancy, and many claims of robustness are still too limited in scope to support their actual use in education. This is particularly applicable to LLM-based AES as there is sometimes a trade-off between flexibility and consistent assessment quality.

The review also reveals that there are multiple standards and norms of fairness. Results of fairness may be different for the same model, depending on the metric used because there are various measures of bias, and it may not be directly comparable across studies . Although there have been strides in bias mitigation using balanced training, feature redesign, and selective debiasing, these techniques are not necessarily effective at combating historical inequities, label bias, or fairness under deployment shift . This is where the LLM-based AES comes in, but with it comes the need for explainability, prompt sensitivity, opacity, and governance more than ever.

The reality of the situation is obvious. The evaluation design of research should be stronger, there is a need for more consistent reporting, and NLP methods should be more closely linked to educational measurement. Fairness, auditability, and contextual validity must be an integral part of the developmental and policy process and not an afterthought. In contexts where even minor systematic error can impact feedback, placement,

progression and opportunity, trustworthiness requires transparency relative to the limits of the model, monitoring for subgroup effects, and responsible human oversight, among other elements.

The review has five main contributions. It unifies fairness, bias, and generalization in one package, instead of considering them as different issues. It maps the evolution of the field from feature based systems to neural/transformer-based and LLM-based models and yet retains the educational relevance as a central element. It attracts attention to the influence of the design of the benchmarks and the choice of the corpus on the knowledge that AES research can have. Relates transferability and fairness portability to the larger literature on bias. It provides a research pathway based on facts and not just optimism.

AES is now in a critical situation. It is no longer only a question of whether student writing can be automatically scored with "acceptable reliability. It now asks 'what for', by whom and when can such systems be trusted? It's not just about the better models, though. Achieving fair and trustworthy AES will need more evidence, more transparency in the implementation and a continued investment in educational equity.

References

1. Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational measurement: issues and practice, 31(1), 2-13.
2. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment, 4(3).
3. Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing, 20, 53-76.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).
5. Loukina, A., Madnani, N., & Zechner, K. (2019, August). The many dimensions of algorithmic fairness in educational applications. In Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications (pp. 1-10).
6. Phandi, P., Chai, K. M. A., & Ng, H. T. (2015, September). Flexible domain adaptation for automated essay scoring using correlated linear regression. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 431-439).
7. Miller, G. A. (2003). The cognitive revolution: a historical perspective. Trends in cognitive sciences, 7(3), 141-144.
8. Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. Interactive learning environments, 8(2), 111-127.
9. Nadeem, F., Nguyen, H., Liu, Y., & Ostendorf, M. (2019, August). Automated essay scoring with discourse-aware neural models. In Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications (pp. 484-493).
10. Taghipour, K., & Ng, H. T. (2016, November). A neural approach to automated essay scoring. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 1882-1891).
11. Dong, F., Zhang, Y., & Yang, J. (2017, August). Attention-based recurrent convolutional neural network for automatic essay scoring. In Proceedings of the 21st conference on computational natural language learning (CoNLL 2017) (pp. 153-162).
12. Nguyen, H., & Litman, D. (2018, April). Argument mining for improving the automated scoring of persuasive essays. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
13. Perelman, L. (2014). When “the state of the art” is counting words. Assessing Writing, 21, 104-111.
14. Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016, August). Automatic text scoring using neural networks. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers) (pp. 715-725).
15. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.
16. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
17. Wang, Q., & Gayed, J. M. (2026). Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. Computer Assisted Language Learning, 39(1-2), 209-237.

18. Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. Assessing Writing, 17(1), 35-54.
19. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In International conference on machine learning (pp. 2790-2799). PMLR.
20. Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. The Journal of Technology, Learning and Assessment, 10(3).
21. Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023, July). Rating short L2 essays on the CEFR scale with GPT-4. In Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023) (pp. 576-584).
22. Gao, F., Jiang, H., Yang, R., Zeng, Q., Lu, J., Blum, M., ... & Li, I. (2024, August). Evaluating large language models on wikipedia-style survey generation. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5405-5418).