Curriculum-Aligned Semantic Assessment: Improving Fairness and Accuracy in AI-Assisted Student Marking

Mlalazi Sijabulisiwe[1, a*]

[1]Xi'an Technological University, Xi'an, Shaanxi, China

[a]mengkamlalazi@gmail.com

Abstract. As artificial intelligence becomes increasingly integrated into education, AI-assisted marking systems offer benefits such as reduced grading workloads and improved consistency. However, many existing models penalize students for deviating from a broader knowledge base rather than adhering to the prescribed curriculum, raising concerns of fairness. This paper presents a curriculum-aligned natural language processing (NLP) model designed to assess short and long-form student responses within the scope of a national syllabus. The system dynamically ingests curriculum documents, encodes key concepts into a hierarchical structure and semantically matches student answers using a fine-tuned BERT model. A simulated dataset of 1,400 primary science responses (including in-scope, out-of-scope and distractor answers) was used to evaluate the model. Results show a strong correlation with human marking (Pearsons r = 0.88) and improved precision and recall in identifying out-of-scope content (F1 = 0.90), outperforming a non-aligned baseline. These findings suggest that curriculum-aware assessment models can enhance fairness, uphold instructional integrity and support scalable, transparent evaluation in educational settings.

Keywords: Curriculum-aware assessment, Natural language processing in education, AI-assisted marking, Educational fairness, Syllabus-aligned evaluation, Explainable educational AI, Semantic scoring models.

1. Introduction

Automated marking systems powered by artificial intelligence (AI) are becoming more common in schools, as they can reduce teachers' workload, provide rapid feedback and maintain consistent grading standards [1]. In 2024, the UK Department for Education launched a £4 million programmed to train AI models on official curriculum guidance and anonymized student work [2]. The effectiveness of this initiative in improving marking accuracy remains under evaluation. Many existing AI markers, however, judge student responses using broad, open-domain knowledge rather than the specific content taught in class. This mismatch can penalize students whose answers align with the official syllabus but omit extra facts or terminology known to the AI [3]. Such errors can undermine trust in AI marking and increase the time teachers spend correcting automated grades. Research on fairness in AI suggests that systems should ground their decisions in the context they serve, combining automated scoring with human oversight to avoid unfair outcomes [4,5,6]. My proposed model follows this approach as it references official curriculum documents and teacher-defined scope settings before assigning a mark and it flags out-of-scope content for teacher review rather than deducting marks.

2. Related Works

This section reviews prior research on AI-driven marking, embedding key gaps and implications within each discussion. We consider: (a) foundational scoring methods, (b) curriculum misalignment, (c) fairness and transparency, (d) human–AI collaboration and (e) early curriculum-aware systems.

2.1 Automated Marking Systems, Foundations and Approaches. Early automated essay scoring (AES) systems used surface-level text features as proxies for writing quality. Basic measures such as essay length were first shown to predict human grades by Page [7]. Attali and Burstein then formalized this approach, correlating word count, average sentence length and vocabulary diversity

with expert marks [8]. Elliot's Intelligent Essay Assessor extended these ideas, reporting a moderate correlation but warning that essays padded with repetition sometimes earned unduly high scores [9]. Williamson and Piattoeva argued that reliance on surface features risks rewarding verbosity rather than genuine understanding [10]. These findings imply a need for deeper analysis of student content, since surface metrics cannot ensure marking aligns with curriculum-taught material [8, 9, 7, 10].

Semantic models sought to capture meaning rather than mere form. Foltz, Laham and Landauer introduced latent semantic analysis (LSA), mapping student essays and reference texts into a shared semantic space and demonstrating improved agreement with human markers [11]. Omar showed that LSA enhanced topic coherence detection [12], while Landauer noted its dependence on high-quality reference corpora and its difficulty with novel argument structures [13]. Research on referential cohesion indicates that, although vital for coherence, its overuse can confuse reviewers; Seyler found that excessive cohesion features negatively impacted clarity among peer reviewers [14].

Many models, particularly convolutional neural networks (CNNs), exhibit a strong inductive bias towards style features, which can overshadow content features and lead to poor performance on out-of-domain data [15]. Similarly, the performance of large language models declines when tested on domains different from their training data, as they may rely heavily on stylistic cues rather than content attributes [16]. These studies reveal that semantic approaches bring AES closer to true understanding but remain blind to curriculum boundaries, risking deductions for valid responses within class scope.

2.2 Curriculum Misalignment in AI Marking. Despite advances in automated essay scoring, systems often fail to align semantic analysis with specific curriculum requirements, risking the assessment of responses that are contextually valid but fall outside expected standards [17]. In a study of physics examinations, Yeadon and Hardy evaluated a large language model's performance across multiple educational levels and reported that, although the model handled GCSE-level questions accurately, its performance declined on more advanced topics, with the AI introducing information not covered by the syllabus that could mislead students [18]. A survey by Bower et al. of secondary educators further revealed concerns that AI-generated assessments penalise correct answers simply because they do not match the AI's broader knowledge base rather than the taught curriculum. This misalignment forced teachers to spend additional time reviewing and overriding automated grades, thus negating anticipated efficiency gains [19]. These findings highlight a critical implication; without explicit curriculum filters and alignment mechanisms, AI-assisted marking can inadvertently increase teacher workload and compromise assessment fairness. Integrating curriculum-aware frameworks and preserving human oversight are therefore essential to ensure that AI tools support, rather than hinder, educational objectives.

Rule-based methods in educational technology (such as natural-language proof checkers and keyword-based feedback filters) provide structured solutions but lack the flexibility to handle diverse, free-text student responses across subjects. For example, the Diproche system uses a controlled-language proof checker for introductory mathematics; it verifies student proofs against a predefined set of inference rules, flags unrecognised steps without deducting marks and has been shown to reduce teacher review time by about 20 per cent [20]. Similarly, Das et al. developed an NLP pipeline to extract keywords from official syllabus PDFs and automatically generate STEM question items; although this improved relevance, it handled only single-term concepts rather than multi-word ideas [21]. These efforts demonstrate that simple curriculum rules can reduce misalignment in well-structured domains, yet a gap remains: developing flexible, scalable methods capable of aligning AI assessment with free-text responses across varied subject areas.

2.3 Fairness and Transparency in AI Scoring. Efforts to make AI decisions more transparent often draw on post-hoc explanation methods. In a comprehensive guide to interpretable machine learning, Molnar reviews techniques such as LIME and SHAP, noting that while they can highlight influential features, they do not guarantee alignment with domain constraints [22]. Slack, Hilgard, Jia, Singh and Lakkaraju demonstrated that adversarial inputs can fool these explainers, calling into question their reliability in high-stakes settings [23]. In an educational context, Maxwell-Smith et al. found that students and teachers trusted AI feedback only when explanations corresponded to

syllabus-taught errors but dismissed remarks on content outside the curriculum [24]. Kumar and Boulanger extended this work by integrating explainable modules into a secondary-school essay scorer, observing that transparent feedback increased teacher acceptance only when paired with curriculum-aware rule checks [25]. Collectively, these studies suggest that transparency tools alone are insufficient unless the system also enforces explicit curriculum boundaries.

Efforts to mitigate bias in automated essay scoring have largely focused on demographic and dialectal fairness, often at the expense of curriculum scope. Blodgett, Barocas, Daumé III and Wallach surveyed bias in NLP systems and documented how non-standard dialects and vernaculars can be systematically disadvantaged, even in educational applications such as AES, highlighting the need for contextual sensitivity in model design [26]. Dixon, Li, Sorensen, Thain and Vasserman showed that adversarial debiasing techniques can reduce gender bias in text classification by around 20 per cent, yet these methods do not address the distinct challenge of marking within strict curricular boundaries [27]. Guo, Pleiss, Sun and Weinberger introduced temperature scaling to calibrate neural network confidences, which can reduce unnecessary teacher reviews for low-confidence predictions but still risk flagging valid, in-scope answers when curriculum alignment is not enforced [28]. These findings indicate that, while bias-reduction and transparency techniques advance fairness, they must be complemented by explicit curriculum-aware filters to ensure that AES tools evaluate only the material students are expected to learn.

2.4 Human–AI Collaboration in Assessment. Human–AI collaboration models that integrate human oversight with automated processes have been shown to balance efficiency with expert judgment. For example, Williamson and Piattoeva [29] described a system in which AI provided initial grading and flagged uncertain cases; this approach enabled teachers to save up to 40 % of marking time, even though some correct answers were erroneously flagged. Similarly, Razmerita [30] examined a hybrid assessment framework in which automated error checks were combined with peer marking. Her study found that, in roughly 30 % of cases, peer reviewers graded responses more strictly when clear curriculum guidance was lacking. Together, these findings highlight the importance of embedding curriculum context into both AI systems and the human review process to avoid unnecessary workload.

Adaptive workflows have also been explored as a means to reduce teacher burden, though they do not fully address all scope-control issues. Yoo et al. [31] implemented a model calibration strategy that lowered teacher review rates by 22 %; yet, responses containing valid ideas expressed with novel phrasing continued to trigger incorrect flags. Moreover, von Davier and Burstein [32] argued that ensuring fairness is crucial to maintaining teacher trust. In practice, many educators supplement AI with ad hoc rubrics; a patchwork approach that can ultimately undermine efficiency and confidence. These insights point to a clear gap; assessment workflows must integrate curriculum boundaries directly into AI systems rather than relying solely on post-hoc human adjustments.

2.5 Curriculum-Aware System, Emerging Effort. Across these research strands, marking systems either ignore curriculum scope, rely on static or manual approaches, or address fairness without scope controls. My curriculum-aligned NLP model aims to fill these gaps by automatically ingesting official curriculum documents, applying explicit scope filters during scoring and flagging only genuine out-of-scope content for teacher review.

3. Model

This section outlines the architecture and key processes of the curriculum-aligned marking assistant. The design aims to address limitations identified in existing systems; specifically, the lack of dynamic curriculum awareness, the risk of unfair deductions for valid in-scope responses and the absence of clear review cues for teachers. Citations are given in APA style.

3.1 System Architecture. Table 1 presents the high-level architecture of the four modules

Table 1 Architecture of four modules

Module	Description/Function
Curriculum Ingestion	Parses official syllabus documents to build a structured concept hierarchy.
Answer Preprocessing and Embedding	Transforms student responses into normalised token sequences and obtains semantic embeddings.
Scoring Engine	Matches embeddings against in-scope concept vectors to compute marks.
Flagging and Review Interface	Identifies out-of-scope content, groups it into spans and presents it for teacher review.

3.2 Curriculum Ingestion. The ingestion module was designed to overcome the limitations of earlier systems that either relied on static rule lists or only handled single-word keywords. It begins by converting official syllabus PDFs into plain text while preserving document layout, headings and font metadata through PDFMiner [33]. Next, section segmentation is performed using a combination of regular expressions and layout cues (such as font size and styling) to detect headings like "Learning Outcomes" and "Key Concepts" (Alshaya [34]). Once sections are identified, spaCy's part-of-speech tagger and noun-phrase chunkier extract multi-word terms (for example, "photosynthetic pigment") as candidate concepts [35]. Finally, these concepts are organised into a parent–child hierarchy by linking narrower terms under broader headings based on section nesting and term co-occurrence, employing a graph-based taxonomy induction method (Wu [36]). The output is a database of topic nodes and sub-nodes with associated keyword sets, ready for matching against student responses.

3.3 Answer Preprocessing and Semantic Encoding. The answer preprocessing and semantic encoding pipeline is designed to capture the depth of student responses while constraining evaluation to curriculum-relevant content. Initially, responses undergo normalisation (lowercasing, stop-word removal and lemmatisation) using spaCy's NLP tools [37]. Next, a BERT base model is fine-tuned on a mixed corpus of official curriculum texts, teacher-graded exemplar answers and subject-specific glossaries, following established fine-tuning protocols; this produces embeddings that reflect both syllabus vocabulary and authentic student phrasing. Finally, sentence embeddings are compared to each concept vector via cosine similarity, with decision thresholds calibrated on a held-out validation set to optimise the trade-off between recognising in-scope content and rejecting out-of-scope material [39]. This three-stage process overcomes the fixed-scope limitations of LSA-only methods [11, 12] and avoids the domain-overreach seen in transformer markers without explicit curriculum filters.

3.4 Scoring and Flagging Mechanism. The scoring engine processes each question's set of expected concepts in two main steps. First, it allocates partial credit to student responses in proportion to the degree by which their sentence embeddings exceed a calibrated similarity threshold. Second, it generates flags by identifying contiguous spans whose maximum similarity to any in-scope concept falls below this threshold; each flagged span is then labelled with its nearest matching concept or, if no match exceeds a secondary lower threshold, marked as "Unknown Topic". Rather than deduct marks for these flagged spans, the system preserves student credit and presents the spans alongside contextual information in the teacher dashboard; a strategy shown to improve review efficiency when cues are targeted and concise.

4. Experimental Design, Implementation and Analysis

4.1 Simulated Curriculum-Based Datasets. To evaluate the marking assistant under controlled conditions, we generated a dataset that mirrors typical primary-school science assessments aligned to a national syllabus. A reference curriculum covering photosynthesis, plant reproduction and ecosystems was encoded into our concept hierarchy (Section 3.2). From this, we designed 20

questions (10 short-answer, 10 paragraph-length) and created “In-scope answers” (1 000 student responses written by educators, each covering only syllabus content), “Out-of-scope answers” (200 responses that include correct content plus extraneous facts (e.g. biochemical pathways), simulating student overreach) and “Distractor answers”. (200 responses with incomplete or incorrect content, omitting key syllabus concepts)

Responses were anonymised and shuffled, yielding a total of 1 400 items. This distribution ensures that 71 % of answers remain in-scope, reflecting realistic classroom patterns. Each response was then independently annotated by two expert teachers, who assigned a gold-standard score and marked any out-of-scope spans. Inter-rater agreement on the flag labels was strong, with Cohen’s $\kappa = 0.87$ indicating high consistency.

4.2 Hardware and Software Environment. All experiments were conducted on a dedicated workstation featuring an Intel Xeon E5-2680 v4 CPU, 64 GB of RAM and an NVIDIA Tesla V100 GPU to accommodate the computational demands of model training and inference. The core implementation was written in Python 3.8, utilising the PyTorch 1.10 framework alongside the HuggingFace Transformers library for BERT fine-tuning and embedding generation. Curriculum ingestion and token preprocessing leveraged spaCy 3.1, while statistical analysis and evaluation metrics were handled by SciPy 1.7 and scikit-learn 1.0. A PostgreSQL 13 database was used to persist the ingested concept hierarchy and student response embeddings. To guarantee reproducibility across environments, all model fine-tuning and evaluation routines were encapsulated within Docker containers.

4.3 Implementation of the Curriculum-Aligned Model. The ingestion pipeline processed PDF syllabus files into a PostgreSQL concept hierarchy via spaCy’s noun-phrase extractor and custom regular expressions, producing 150 unique concept nodes. For answer assessment, BERT base (uncased) was fine-tuned for three epochs on a mixed corpus of 5 000 exemplar answers and the extracted curriculum text, following the procedure of Devlin et al. (2019). Learning rate was set to 2×10^{-5} with a batch size of 16.

During evaluation, each student response was segmented into sentences, embedded via the fine-tuned BERT model and compared by cosine similarity to each concept vector. Thresholds ($\theta = 0.75$) were determined on a 20 % held-out validation set to optimise the F_1 score for flag detection. Partial credit for each question was computed as:

$$Score_q = \sum_{c \in C_q} Max\{0, sim(s,c) - \theta\} \quad (1)$$

4.4 Evaluation Metrics. Performance was assessed on two fronts, namely, “Scoring accuracy” and “Flagging performance”. Scoring Accuracy was assessed through Pearson’s between AI-assigned and human scores and Mean Absolute Error (MAE) of total scores per response. Flagging Performance on the other hand was measured through “Precision and Recall” for detecting out-of-scope spans, treating teacher flags as ground truth and the F1 Score (the harmonic means of precision and recall). Statistical significance was tested using paired t-tests for score differences and McNemar’s test for paired flagging decisions.

4.5 Scoring Accuracy. The curriculum-aligned model achieved = 0.88 with human marks and an MAE of 0.45 points on a 10-point scale. In contrast, a baseline transformer marker without curriculum filtering scored = 0.81 and MAE = 0.72, showing a significant improvement in alignment with teacher grades.

4.6 Flagging Performance. Table 2 summarises the detection of out-of-scope content.

Table 2 Performance on flagging out-of-scope spans

Metric	Curriculum-Aligned	Baseline Transformer
Precision	0.92	0.68
Recall	0.89	0.75
F_1 Score	0.90	0.71

Differences in precision and recall were significant (McNemar's $\chi^2 = 45.2$, $p < 0.001$), indicating fewer false positives and negatives when using curriculum alignment.

4.7 Analysis. Strong alignment between machine and human scores indicates that awarding partial credit based on calibrated similarity thresholds effectively captures pupil understanding, reducing mean absolute error by 38 % compared with a non-aligned baseline. This outcome is consistent with prior observations that curriculum misalignment can lead to unfair deductions when AI models rely on unrestricted domain knowledge [18]. Improved precision in identifying out-of-scope content shows the model can distinguish valid, syllabus-aligned phrasing from genuinely extraneous material, addressing concerns about penalising correct answers for missing advanced content [18]. The narrower recall gap (0.89 versus 0.75) demonstrates the system flags most true out-of-scope spans without overburdening teachers with flags on valid responses, a balance crucial to preserving efficiency and trust in AI-assisted marking [19].

Examination of error cases revealed two key challenges. First, very short yet correct phrases (for example, "stomata open") occasionally fell below the similarity threshold, suggesting the need for dynamic thresholding that adapts to phrase length and lexical density. Second, complex multi-clause sentences were sometimes fragmented into separate flags; refining span grouping by incorporating cohesion and discourse-level parsing could mitigate such fragmentation [14]. Overall, these findings confirm that dynamically ingesting and structuring curriculum materials, combined with semantic matching tuned to context, can enhance both accuracy and fairness in automated marking systems.

5. Conclusion and Future Work

This study presented a curriculum-aligned AI-based marking assistant, developed to ensure that automated scoring reflects not only semantic correctness but also curricular relevance. The model was evaluated using a simulated dataset of 1,400 anonymised student responses aligned with a national primary science syllabus. Results demonstrate that embedding syllabus constraints within the marking process significantly improves the alignment of machine-generated scores with those given by human assessors, while also reducing the rate of incorrect flagging of valid content.

The curriculum-ingestion pipeline transformed official syllabus documents into a structured concept hierarchy, forming the basis for determining the relevance of student responses. In contrast to baseline transformer models, the proposed system assessed not only the presence of correct content but also penalised extraneous information outside the taught material. This dual focus resulted in a Pearson correlation of 0.88 with teacher scores and a mean absolute error of 0.45, indicating strong alignment. Furthermore, out-of-scope content was flagged with high precision (0.92) and recall (0.89), outperforming the baseline which lacked curriculum filtering mechanisms.

These results support the assertion that semantic similarity models, when constrained by syllabus-defined boundaries, can enable more accurate and fairer assessment processes. Awarding partial credit via calibrated similarity thresholds offers a finer-grained approach than binary classification, especially in educational settings where learners' responses often mix correct and irrelevant content. Importantly, these findings address concerns raised by Yeadon and Hardy [18], who observed that domain misalignment can introduce out-of-scope information and by Bower et al. [19], who reported that misaligned AI marking increases teacher workload by requiring additional review of correct responses.

Despite the promising outcomes, several limitations must be acknowledged. The experiment was conducted within a single subject domain (science) and focused on a restricted set of 20 questions, which, while diverse in length and complexity, may not fully capture the variability found in real classroom assessments. Additionally, although the model demonstrated improved flagging of out-of-scope content, a small number of errors were attributable to short correct phrases falling below similarity thresholds and to complex multi-clause responses being inconsistently segmented. These cases suggest that further work is needed in dynamic thresholding and discourse-level parsing to handle varied response structures.

Future research will explore the application of the model to other subject areas, such as mathematics and history, to examine cross-domain robustness. There is also interest in expanding the

dataset to include responses from actual learners across different regions, thereby improving ecological validity. Enhancements to the flagging mechanism are planned, including context-aware grouping of flagged spans and the integration of attention-based mechanisms to capture argument coherence across multiple sentences. Moreover, teacher-facing interfaces allowing manual override and real-time review of flagged content will be developed to support classroom integration.

Another avenue involves making the curriculum-ingestion process more interactive, allowing educators to annotate and adjust concept hierarchies according to classroom emphasis or local syllabus variations. This would enable more granular customisation and allow the model to better reflect classroom realities. Additionally, there is potential to embed the system into formative assessment tools that provide instant feedback to learners, linking flagged content with remedial suggestions aligned to the syllabus.

References

[1] Langove, S. A. and Khan, A. (2024). Automated Grading and Feedback Systems: Reducing Teacher Workload and Improving Student Performance. Journal of Asian Development Studies, 13(4), 202-212.

[2] Kyle, P and Morgan, S. (2024). Press release: Teachers to get more trustworthy AI tech, helping them mark homework and save time..

[3] Ravindran, K. (2024). Exploring the Potential: Can AI Effectively Mark Students' Work?

[4] Whitmer, J. and Beiting-Parrish, M. (2024). Lessons Learned About Transparency, Fairness and Explainability from Two Automated Scoring Challenges. Federation of American Scientists, Impact Fellows, Institute of Education Sciences, U.S. Department of Education.

[5] Litman, D., Zhang, H., Correnti, R., Matsumura, L.C., Wang, E. (2021). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2021. Lecture Notes in Computer Science(), vol 12748. Springer, Cham.

[6] Richardson, B. and Gilbert, J. E. (2021). A framework for fairness: A systematic review of existing fair AI solutions. Journal of Artificial Intelligence Research 1, 1-28.

[7] Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. The Journal of Experimental Education, 62(2), 127-142.

[8] Attali, Y. and Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. Journal of Technology, Learning and Assessment, 4(3).

[9] Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis and J. C. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective. Mahwah, NJ: Erlbaum.

[10] Williamson, B. and Piattoeva, N. (2019). Objectivity as Standardization in Data-Scientific Education Policy, Technology and Governance. Learning, Media and Technology, 44(1), 64-76.

[11] Landauer, T. K., Foltz, P. W. and Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259–284.

[12] Omar, A. (2017). Addressing the Problem of Coherence in Automatic Text Summarization: A Latent Semantic Analysis Approach. International Journal of English Linguistics, 7(4), 33.

[13] Landauer, T. K. (2002). Applications of Latent Semantic Analysis (Vol. 24, p. 44). Routledge.

[14] Seyler, T. M. (2023). Understanding the Effect of Cohesion in Academic Writing Clarity Using Education Data Science. Big Data Management, 193–218.

[15] Nam, H., Lee, H., Park, J., Yoon, W. and Yoo, D. (2021). Reducing Domain Gap by Reducing Style Bias. Computer Vision and Pattern Recognition, 8690–8699.

[16] Roussinov, D., Sharoff, S. and Puchnina, N. (2024). Controlling Out-of-Domain Gaps in LLMs for Genre Classification and Generated Text Detection.

[17] Cheng, Y. and Nunes, B. P. (2022). The use of Semantic Technologies in Computer Science Curriculum: A Systematic Review.

[18] Yeadon, W. and Hardy, T. (2024). The impact of AI in physics education: a comprehensive review from GCSE to university levels. Physics Education, 59(2), 025010.

[19] Bower, M., Torrington, J., Lai, J.W.M., Petocz, P. and Alfano, M. (2023). How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey. Education and Information Technologies, 29, 15403–15439.

[20] Carl, M., Lorenzen, H. and Schmitz, M. (2022). Natural Language Proof Checking in Introduction to Proof Classes--First Experiences with Diproche.

[21] Das, Deepayan and Paul, Atanu and Ray, Ryan and Chanda, Srinjoy and Biswas, Sandipan. (2025). Keywords Driven Question Bank Generation for Educational System using NLP.

[22] Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3rd Ed.).

[23] Slack, D., Hilgard, S., Jia, E., Singh, S. and Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics and Society (pp. 180-186).

[24] Maxwell-Smith, Z., Ochoa, S. G, Foley, B. and Suominen, H. (2020). Applications of Natural Language Processing in Bilingual Language Teaching: An Indonesian-English Case Study. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 124–134, Association for Computational Linguistics, Seattle, WA, USA.

[25] Kumar, V. and Boulanger, D. (2020). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. Frontiers in Education, 5, 572367.

[26] Blodgett, S. L., Barocas, S., Daumé III, H. and Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5454–5476. Association for Computational Linguistics.

[27] Dixon, L., Li, J., Sorensen, J., Thain, N. and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics and Society (AIES '18), 67–73.

[28] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In International conference on machine learning (pp. 1321-1330). PMLR.

[29] Williamson, B. and Piattoeva, N. (2018). Objectivity as standardization in data-scientific education policy, technology and governance. Learning, Media and Technology, 44(1), 64–76.

[30] Razmerita, L. (2024). Human-Al Collaboration: A Student-Centered Perspective of Generative Al Use in Higher Education. In Proceedings of the 23rd European Conference on e-Learning. Academic Conferences International.

[31] Yoo, J., Park, J., Ha, M. and Mae Lagmay Darang, C. (2024). Exploring Pre-Service Teachers' Cognitive Processes and Calibration with an Unsupervised Learning-Based Automated Evaluation System. SAGE Open, 14(3), 21582440241262864.

[32] von Davier, A. A. and Burstein, J. (2024). AI in the Assessment Ecosystem: A Human–Centered AI Perspective. In Artificial Intelligence in Education: The Intersection of Technology and Pedagogy (pp. 93-109). Cham: Springer Nature Switzerland.

[33] Shinyama, Y. (2015). PDFMiner: A tool for extracting information from PDF documents.

[34] Alshaya, S. A. (2025). Enhancing Educational Materials: Integrating Emojis and AI Models into Learning Management Systems. Computers, Materials and Continua, 83(2).

[35] Misra, S. (2023). Nested Noun Phrase Detection in English Text with BERT.

[36] Wu, M. (2023). An Intelligent Bibliometric System for Knowledge Association and Hierarchy Discovery. University of Technology Sydney (Australia).

[37] Pant, V. K., Sharma, R. and Kundu, S. (2024). An overview of stemming and lemmatization techniques. Advances in Networks, Intelligence and Computing, 308-321.

[38] Rodriguez, P. L., Spirling, A., Stewart, B. M. and Wirsching, E. M. (2023). Multilanguage word embeddings for social scientists: estimation, inference and validation resources for 157 languages. Working paper.