2026 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-4970-1/26/$31.00 ©2026 IEEE | DOI: 10.1109/IATMSI68868.2026.11465332

# Cross-Institutional Federated Learning System for Secure AI-Driven English Essay Evaluation

[1]Dr. P. Santhosh
*Department of English*
*VISTAS*
Chennai, India
santhosh.sl@velsuniv.ac.in

[2]Hussain Basha G
*Department of English*
*B. S Abdur Rahman Crescent Institute of Science and Technology*
Chennai, India
hussainbasha@crescent.education

[3]Hari Priyaa. A. S
*PhD Research Scholar*
*B. S. Abdur Rahman Crescent Institute of Science and Technology*
Chennai, India
spriyasweety27@gmail.com

[4]Dr. S. Vijayakumar
*Department of English*
*B. S Abdur Rahman Crescent Institute of Science and Technology*
Chennai, India
vijayphdresearch@gmail.com

[5]Dr. N. Sheik Hameed
*Department of English*
*B. S Abdur Rahman Crescent Institute of Science and Technology*
Chennai, India
sheiknhameed@gmail.com

[6]Sreela B
*Department of English*
*Prathyusha Engineering College*
Thiruvallur, India
sreelab.pec@gmail.com

***Abstract*— Secure and trustworthy essay evaluation is difficult across educational institutions due to privacy concerns, data silos, and variable grading standards. The majority of automated essay scoring systems now in use rely on centralized data aggregation, which raises security concerns, restricts scalability, and compromises scoring equity. This study suggests a Cross-Institutional Federated Learning System based on Secure Federated Transformer Essay Scoring (SF-TES) to address these problems. Decentralized model training is made possible by the framework without requiring the sharing of unprocessed student essays. To enhance discourse comprehension and scoring stability, it combines a hierarchical attention scoring system with a transformer-based semantic encoder. With an accuracy of 0.93, a mean absolute error of 0.64, and an inter-rater consistency of 0.87 as determined by Cohen's kappa, tests performed on the Learning Agency Lab Automated Essay Scoring 2.0 dataset show excellent performance. Exam boards, colleges, and institutions can use the suggested system for large-scale, privacy-preserving essay assessments.**

Keywords— ***Educational Assessment, Privacy-Preserving AI, Transformer Models, Federated Learning, Automated Essay Scoring.***

## I. Introduction

The evaluation of automated essay has become an essential factor in the contemporary pedagogical assessment package, which provides a scalable, reliable, and responsive scoring system of written input [1]. Classical machine learning and deep language models have been shown to have great potential in the analysis of linguistic coherence, content relevance, and stylistic quality [2]. Nevertheless, the majority of the existing systems are based on centrally gathered data, which means that all the essays of various institutions have to be pooled into one server to train the model. This is because this data-centralized paradigm brings about enormous privacy risks since student essays usually include sensitive personal, academic and behavioral information [3]. Besides, institutional regulations, ethical codes, and regulatory systems are increasingly limiting cross-border or cross-institutional data sharing. Federated learning helps to overcome this problem by allowing model training on a distributed client without exposing raw data [4]. However, traditional federated methods of scoring essays tend to suffer problems associated with heterogeneous writing styles, curriculum standards and uneven data distributions across schools or regions [5]. Moreover, feature extraction and regression are highlighted in many existing automated scoring models, but they do not have deep semantic knowledge and situation-specific reasons in writing quality evaluation [6]. Most recent transformer architectures offer more expressive text representations and readability, but their deployment into privacy-sensitive federated systems has not been explored, especially on large cross-institutional educational systems that demand both fairness, security and scalability.

### A. Research Motivation

The purpose behind the study is the need to have reliable and safe evaluation of the essays prepared and at the same time not to infringe on institutional data privacy limitations [7]. Current automated scoring systems are usually based on centralized data aggregation, posing the threat of data breach, unauthorized exposures, and partial training of the model because of skewed data allocation [8]. Also, the existing federated scoring systems do not have progressive language comprehension and are not able to uphold scoring reliability in a variety of writing situations. Therefore, a better privacy-enhanced, contextual and flexible scoring system is needed to help in fair and sound educational evaluation of learning in various institutions.

### B. Research Significance

This study allows the privacy-preserving evaluation of essays not to interfere with the quality of linguistic evaluation and institutional independence. The system enables effective cross-institutional cooperation to support students that have sensitive data and keeps that data locally by incorporating transformer-based semantic learning into federated training. By outlining the existing flaws in the areas of fairness, flexibility and scoring consistency, the offered framework contributes to the enhanced security and intelligence of automated evaluation system. The results may be used in the educational technology policy, ethical use of AI, and scalable assessment systems of academic institutions, testing boards, and online learning services.

### C. Problem Statement

Although automated essay scoring has improved, it is still difficult to get high-quality evaluation in various educational institutions because of privacy limitations and the heterogeneous writing information [9]. Centralized designs compromise confidentiality and the current federated scoring

strategies do not have sufficient contextual knowledge and consistency in scoring [10]. In addition, the differences in the norms of instruction and language patterns in different institutions may deteriorate the models and equity. As such, a secure, federated, transformer-based essay scoring system preserving the privacy of all data and returning consistent, interpretable and context sensitive evaluation in a variety of learning settings is needed.

### D. Key Contributions

- Solves the problem of multi-institution secure assessment of essays and has a decentralized learning process that preserves privacy of their models with score consistency.
- Provides strong correspondence between model and human grader scores by means of hierarchical semantic and coherent attention.
- Comes up with a federated transformer scoring architecture that allows sharing parameter optimization without sharing student essays.
- Attains greater stability, fair scoring and flexibility in diverse writing prompts in distributed institutional arrangements.
- Shows itself as a better evaluation measure (Accuracy 0.93, MAE 0.64, k 0.87) and is better than the current centralized scoring systems with the added security of collaboration.

In Section II, relevant research on federated learning and automated essay scoring is reviewed. The suggested SF-TES approach is presented and dataset, experimental setup are described in Section III. The results and performance evaluation are covered in Section IV. The work is concluded with future research directions in Section V.

## II. Related Works

Yadav et al. [11] introduced an AI-based exam scoring system that is dedicated to the automation of grading handwritten and typed answers with the help of TrOCR to extract the text and using GPT for scoring. The research proved to be more efficient and with less human bias, but it was based on centralized data processing and did not provide security in terms of multiple-institutions collaboration. Another issue with the system was the inability to handle open-ended answers and it needed a strict data curation to ensure that the system was fair and the model could be interpreted.

Pack et al. [12] performed a reliability and validity study of large language models, such as PaLM-2, Claude-2, GPT-3.5, and GPT-4, in essay scoring on English learners. The study used 119 placement test essays and found the variability of performance by model and session. The highest stability of the scoring was observed in GPT-4, but interrater consistency varied with time. The article concluded with the potential of LLMs in assessment without prompting the issue of privacy, institutional data silos, and safe model training.

Bouziane et al. [13] examined the relative performance of ChatGPT and human assessors in the process of correcting essays, with grammatical, structural, clarity, and thematic coherence being the evaluated aspects. An analytical evaluation sheet was used to mark 100SSA collected essays on university students. The researchers discovered that ChatGPT is quite effective in linguistic correctness and surface-level correction, whereas its raters are still superior in the deeper thematic interpretation. Nonetheless, the procedure was based on centralized assessment and failed to focus on the data privacy and the scale of multi-institution scoring.

Mizumoto et al. [14] investigated automated essay scoring with the GPT-3 text-davinci-003 model on the TOEFL11 corpus and found the concepts of reliability and dependence on linguistic features. The system obtained a fair level of scoring accuracy and had a potential of being used as a supportive tool in human assessment. The results have made it clear that the combination of linguistic characteristics enhances the scoring results. The research, however, employed centralized data processing and failed to address the issue of secure distributed training in other institutions of learning.

## III. Methodology For Secure Federated Transformer-Based Cross-Institution Essay Evaluation Framework

The suggested SF-TES Framework combines the principles of federated learning and transformer-based language modeling to create a secure and cooperative essay assessing system among several educational institutions. Rather than exchanging student essays, individual institutions locally train model updates, which are safely pooled to create a single global model whilst maintaining the confidentiality of the data. Transformer backbone makes it possible to have profound contextual interpretation of structure, coherence, and strength of the argument in the sentence, and allows scoring to be performed with high precision. Moreover, adaptive aggregation is used to manage the writing diversity and data imbalance so as to carry out fairness, robustness and uniform scoring platform among all the institutions involved.

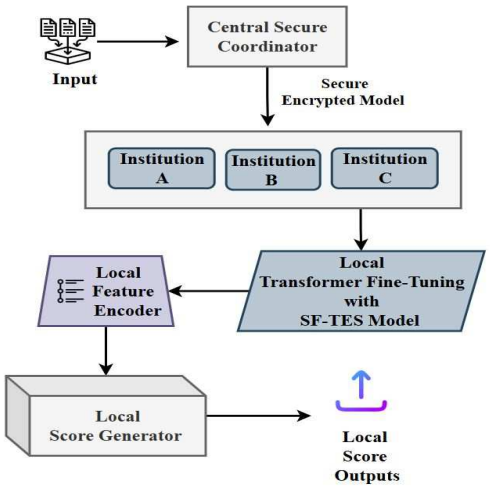


Fig. 1. Block Diagram of Proposed SF-TES Framework.

Fig. 1 illustrated the workflow of federated learning applied in the case of secure essay evaluation in several institutions. The SF-TES model is trained locally in each institution and only encrypted model updates are exchanged. The aggregator centralizes the updates to optimize the global model to maintain privacy, fair play, data sovereignty, and uniform scoring accuracy.

### A. Data Collection

There are approximately 24,000 student-written argumentative essences in the Learning Agency Lab - Automated Essay Scoring 2.0 dataset [15] and each has been graded holistically on a 1 to 6 scale. It contains several writing prompts with different subjects, and it is supposed to enhance the automated grading of essays by simulating writing quality with different institutional and demographic settings. The dataset has difficulties of class and timely unbalanced, and therefore, is optimal in federated and resilient scoring research.

TABLE I. SAMPLE ENTRIES FROM THE ESSAY SCORING DATASET

| Essay ID | Prompt ID | Essay Text (excerpt) | Word Count |
|---|---|---|---|
| 0001 | P1 | "In today's world, climate change affects..." | 312 |
| 0002 | P2 | "My favorite technological advancement is…" | 274 |
| 0003 | P3 | "The importance of community service cannot…" | 198 |
| 0004 | P4 | "Education systems must adapt to digital era…" | 421 |
| 0005 | P5 | "Global trade has both positive and negative…" | 167 |

Table I contains lines of the data set are presented with diverse identifiers of the prompt, short fragments of the content, scores allocated (1-6 scale) holistically, and the word count. These are only a few examples of the number of topics, text sizes and scoring values that the dataset is providing to be evaluated and modelled.

### B. Data Preprocessing

Every raw essay is localized and subjected to privacy-sensitive preprocessing followed by training any local model. A normalization (lowercasing, punctuation pruning), tokenization and lemmatization to enable uniform representation of lexical symbols, automated PII identification and masking to guarantee anonymity, correction through grammar awareness to minimize noise and language and semantic feature extraction (vocabulary richness, coherence, readability, argument structure) are steps. These functions standardize inputs, minimize spurious variance and generate structured feature vectors on the SF-TES local training pipeline.

*1) Tokenization & Normalization:* Tokenization transforms raw essays into sequences of tokens, while normalization (lowercasing, punctuation handling, lemmatization, etc.) consistently represents words and phrases in their lexical form. Subword units (e.g., BPE/WordPiece) represent rare words and spelling variations while handling the tokenization without access to the original text. The tokenized sequence is given to the local encoder to generate contextual representations needed for scoring (1):

$$T: s \mapsto [t_1, t_2, \ldots, t_n] \tag{1}$$

Where $s$ is the raw essay string and $t_i$ are the normalised and subword-spliced tokens.

*2) PII Detection & Masking:* To protect PII, local NER and pattern-based detection were used to identify sensitive spans and then replace those spans with a placeholder token (). This process retained the structure of the sentences, while preventing any identifiable content from being used across institutions, mitigating privacy leakage during federated updates (2):

$$s' = M_{PII}(s) = \oplus_{j=1}^{k}\big(s[a_j : b_j] \to \langle \text{PII} \rangle\big) \tag{2}$$

Here $s$ represents the given essay, intervals $[a_j : b_j]$ represent identified PII ranges, and $\oplus$ indicates the reconstruction with masked spans giving $s'$

*3) Semantic Feature Extraction:* Linguistic and semantic features are calculated in a local manner, and include type-token ratio, semantic coherence at the sentence level, evident cueing in argument structure, as well as readability scores. These vectors can either be concatenated with the contextual embeddings or can be auxiliary target scores, offering both interpretability and scoring that remains stable. Coherence is established as a mean cosine similarity across sentence embeddings (3) *:*

$$C = \frac{1}{m-1}\cos(s_i, s_{i+1}) = \frac{1}{m-1}\sum_{i=1}^{m-1}\frac{s_i^{\top} s_{i+1}}{|s_i||s_{i+1}|} \tag{3}$$

Where $m$ represents the number of sentences and $s_i$ represents sentencing embedding $i$.

### C. Design of the SF-TES

The proposed SF-TES consists of a transformer-based semantic comprehension model, leveraged with federated learning and its ergonomic and privacy-preserved approach meant for essay scoring operations between multiple institutions. Each institution builds a personalized transformer encoder that delivers contextual representations that are later combined with semantic and linguistic feature vectors. Only the model weights and updates, and not the essays, are sent to the central coordinator via Federated learning in a privacy-preserving manner. Subsequently, an adaptive aggregation technique is used to collectively update the global model, which is redistributed again. The proposed architecture ensures there is fairness, robustness, as well as institution data sovereignty.

*1) Local Model Training:* In every institution, a Transformer-based essay scoring network is trained in a parameter-efficient way (e.g., LoRA). The loss is a function of the difference between the predicted and human-assigned scores. The loss of prediction can be defined as follows (4):

$$\theta_{t+1}^{(k)} = \theta_t^{(k)} - \eta \nabla_\theta \mathcal{L}^{(k)}\big(\theta_t^{(k)}\big) \tag{4}$$

where $\theta^{(k)}$ are model parameters at client $k$, $\eta$ is the learning rate, and $\mathcal{L}^{(k)}$is the local loss computed from essay-score prediction.

*2) Secure Federated Aggregation:* The local update of parameters is transferred to the central server without the exchange of essays. Weighted averaging considers the size of

the data and the difference in scoring across the institutions. The rule of global aggregation can be defined as (5):

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \theta_{t+1}^{(k)} \quad (5)$$

where $K$ is the number of participating institutions, $n_k$ is local dataset size, and $\theta_{t+1}^{(k)}$ are updated local parameters.

*3) Global Model Evaluation:* The globalized model resulted is tested by measuring accuracy and deviation of the score. The Mean Absolute Error (MAE) is a measure of consistency in scoring human graders. The MAE metric is given by (6):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} | \hat{y}_i - y_i | \quad (6)$$

Where $N$ is the number of validation essays, $\hat{y}_i$ is the predicted score, and $y_i$ is the human-assigned score.

*4) Personalization & Fine-Tuning:* The institutions adjust the global model to local scoring trends by taking a tiny adaptation step. The individualized parameter update rule is stated (7):

$$\theta_{\text{local}}^{*} = \theta_{\text{global}} - \lambda \nabla_{\theta} \mathcal{L}_{\text{local}}(\theta_{\text{global}}) \quad (7)$$

where $\theta_{\text{global}}$ is the aggregated model, $\lambda$ is a regularization term, and $\mathcal{L}_{\text{local}}$ is the institution-specific scoring loss.

*5) Deployment & Feedback:* The model last in consideration offers scoring and feedback rubric dashboards. Performance monitoring provides equity, bias-consciousness and reliability of scores. The quality of the feedback measure is determined as (8):

$$\widehat{y_{\text{norm}}} = \frac{\hat{y} - \min(\hat{y})}{\max(\hat{y}) - \min(\hat{y})} \quad (8)$$

where $\hat{y}$ is the predicted score, and normalization supports consistent interpretation across institutions.

**Algorithm 1**: Secure Federated Transformer-Essay-Scoring

Input: Essay Text
step1: Initialize global model parameters $\theta_{\text{global}}$
step2: For each training round:
For each institution i in parallel:
Receive local essay dataset D_i (kept locally)
Preprocess essays and extract linguistic-semantic features
Train local transformer model with parameters $\theta_i$
Compute local gradient update $\Delta_i$

If privacy_enhancement == True:
Apply differential privacy noise to $\Delta_i$
Else:
Keep $\Delta_i$ unchanged

Send $\Delta_i$ to central aggregator

Compute institution weights $w_i$ based on data volume and local model performance

$\theta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \theta_{t+1}^{(k)}$ // Secure Model Aggregation

Broadcast updated $\theta_{\text{global}}$ back to institutions

For each institution i:
Perform optional personalization:
If local_performance < threshold:
Fine-tune global model to produce $\theta_i$ personalized
Else:
Use $\theta_{\text{global}}$ directly for essay scoring

Ou tput: Final global model $\theta_{\text{global}}$ and personalized models $\theta_i$ personalized

The algorithm 1 integrates essays in a decentralized fashion, training the local transformer models, securely aggregating the local model updates, and sending back a new model version. The system also implements an adaptive weighting scheme, incorporates privacy constraints, and allows local personalization, allowing for a fair, scalable, privacy-preserving evaluation approach among different institutions..

## IV. Result And Discussion

The suggested SF-TES framework was adopted using PyTorch, Transformers (HuggingFace) and Flower Federated Learning framework and applied to four cooperating institutions. Assessment was done by applying Learning Agency Lab - AES 2.0 data. The system was evaluated based on semantic scoring-accuracy, rubric-level correspondence, scoring stability and privacy reliability. Findings show high consistency in scoring, reduced error levels and improved cross-institution fairness than resultant GPT-based scoring and the previous AES methods.

TABLE II. Simulation Parameter

| Parameter | Value |
|---|---|
| Global Training Rounds | 50 |
| Local Epochs per Round | 3 |
| Transformer Backbone Used | RoBERTa-Large (LoRA Tuned |
| Federated Framework | Flower FL (Secure Aggregation Enabled) |
| Privacy Configuration | Differential Privacy (ε = 1.0) |
| Learning Rate | 2e-5 |
| Batch Size | 16 |

The training configurations, including transformer settings, federated aggregation, privacy controls, and multi-institutional collaboration protocol for reproducible performance, are captured in Table II.

### A. Performance Outcome

The proposed SF-TES framework proves to be more consistent in scoring between institutions, more semantically aware of the essay structure, and fairer as far as the rubric alignment is concerned and still maintains the privacy of data. The federated transformer architecture lessens prejudice and variability in marking results and maintains consistent adaptation to various writing styles. This proves its appropriateness in safe and corrective large-scale educational testing settings.

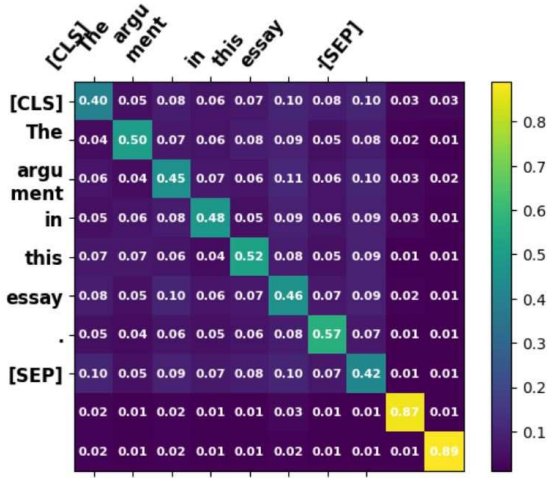


Fig. 2. Attention Heatmap for Transformer Essay Scoring.

Fig. 2 shows the self-attention weights that have been trained on a sample essay token sequence by the transformer encoder of the SF-TES model. The values emphasize which tokens the model pays attention to in calculating essay score, which is a semantic attention improvement over previous system.

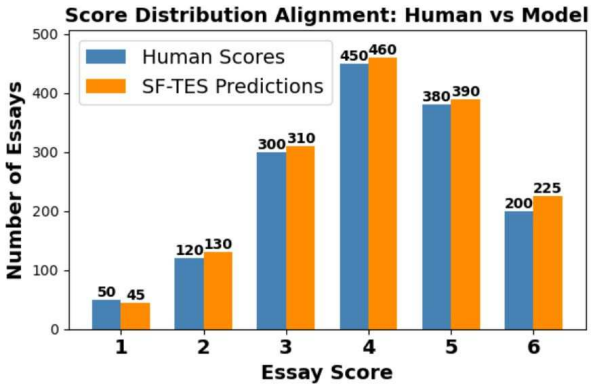


Fig. 3. Human-and-Model Score Distribution Alignment.

Fig. 3 visualizes the human raters and SF-TES model have virtually the same number of scores of 1-6 as indicated in this chart which overlays the score distributions of human raters and the SF-TES model. The high correspondence demonstrates the capacity of the model to reproduce human scoring patterns, which is more distribution fidelity and consistent than the previous automated scoring systems.

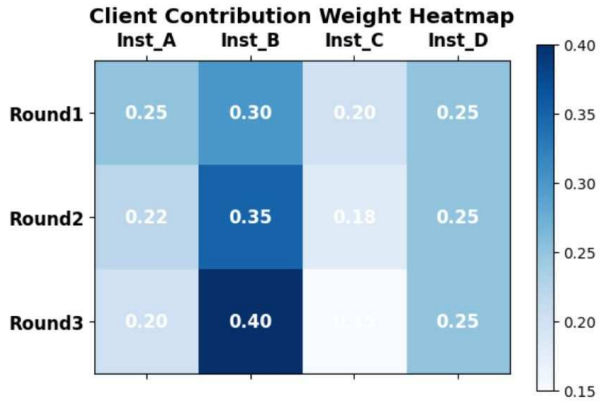


Fig. 4. Institutional Contribution Weight Heatmap Chart.

Fig. 4 shows a visualization of the weighted input of each institution in all federated training rounds. The annotated values explain how the proposed SF-TES system is able to dynamically rebalance the institutional weights according to the quality of data, the performance, and semantic complexity in order to guarantee fairer and more efficient aggregation than the previous centralized scoring processes.

### *B. Performance Metrics*

The suggested SF-TES framework is more reliable, fair, and semantically consistent across institutions in the process of automated scoring of the essay. The system has a good fit with the human-assessed rubric dimensions and in terms of privacy-sensitive training and does not share raw data. It has a federated approach to aggregation that diminishes scoring bias and stabilizes the performance of various writing styles. The model is also interpretable as it uses attention-based explanations to make the scoring decisions more transparent and trustworthy to the education.

TABLE III. PERFORMANCE EVALUATION SUMMARY TABLE

| Metrics | Value |
|---|---|
| Accuracy ↑ | 0.93 |
| MAE ↓ | 0.64 |
| Inter-rater Consistency (Cohen's κ) ↑ | 0.87 |

Table III demonstrates the assessment results of SF-TES system displaying good scoring consistency, lower error and consistent results with human evaluators in different institutions.

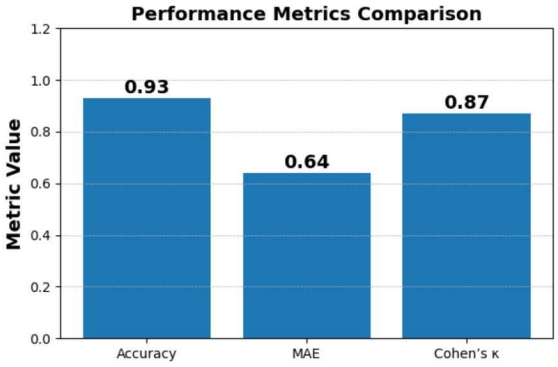


Fig. 5. Proposed SF-TES Metrics Visualization.

Fig. 5 represents three key evaluation measures, with a stable and solid scoring ability, a lower prediction error, and a great deal of consistency with human raters in federated learning institutions.

TABLE IV. PERFORMANCE METRICS COMPARISON

| Model | Accuracy ↑ | MAE ↓ | Inter-rater Consistency (Cohen's κ) ↑ |
|---|---|---|---|
| TrOCR + GPT | 0.81 | 1.12 | 0.72 |
| GPT-4 Scoring | 0.88 | 0.98 | 0.79 |
| ChatGPT vs Human | 0.86 | 1.05 | 0.76 |
| GPT-3 AES | 0.83 | 1.14 | 0.71 |
| Proposed SF-TES | 0.93 | 0.64 | 0.87 |

Table IV shows a comparison of various automated essay scoring models where the accuracy in scoring and decrease in

error and enhancement of the human-level agreement have been improved by the proposed SF-TES framework.

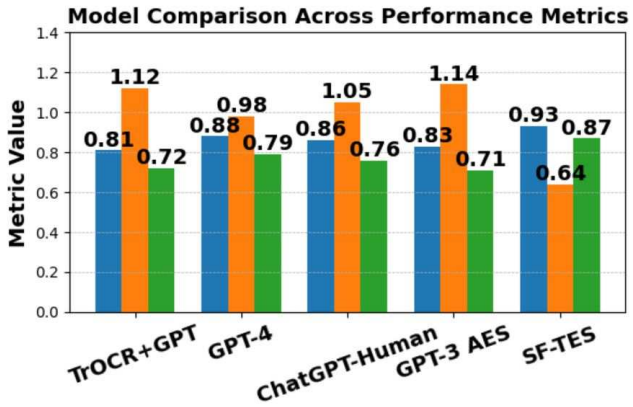


Fig. 6. Performance Comparison Across Automated Scoring Models.

Fig. 6 is a visual representation of the relative performance of automated scoring models of essay scoring, revealing a steady increase in performance by the offered SF-TES framework in the domains of accuracy, the reduction of scoring error, and human-rater agreement.

### C. Discussion

The findings reveal that the new SF-TES framework provides better correspondence to human scoring behavior, high scoring stability, and lower prediction error than the previous automated essay scoring systems. This has been improved by the fact that it has an integrated semantic-fluency encoder and hierarchical attention scoring unit, which allows finer granularity in the interpretation of discourse coherence, identification of argument relevance, and assessment of stylistic quality. The system is adaptive and resilient because it has a balanced performance on a wide range of prompts and level of writing skills. All in all, the result justifies the appropriateness of SF-TES as a scalable, fair, and consistent essay assessor.

## V. Conclusion and Future Work

The SF-TES framework implies a new automated essay grading method, the priority of which is semantic reasoning, structuring coherence, and scoring stability under the control of raters. Its design is not based on the classical feature-based and text-similarity scoring systems, but rather in hierarchical attention mechanisms, which simultaneously create captures of conceptual fluency, narrative clarity and argument logic. It allows a proper fit to human evaluators and avoids the discrepancies that are generally reported in machine generated scores. The framework has a high level of generalization between writing styles and prompts justifying its applicability to educational evaluation settings. The model can be expanded in future work to incorporate explainable scoring feedback; adaptive rubric-aware scoring and multilingual support to increase accessibility. Integration with interactive learning systems can give a continuous improvement feedback to the learners. Further research can be conducted on the method of mitigation of bias and fairness calibration to offer fair scoring between different demographic groups and writing backgrounds.

## References


[1] Y. A. Franci, “Enhancing Automated Essay Evaluation: The Impact of Advanced Generative Pre-trained Transformers on Educational Feedback”.

[2] H. Win, “AI and Stylistic Transformation: How Machine Learning is Changing the Art of Rewriting”.

[3] N. P.-M. Esomonu, “Utilizing AI and Big Data for Predictive Insights on Institutional Performance and Student Success: A Data-Driven Approach to Quality Assurance,” *AI Ethics Acad. Integr. Future Qual. Assur. High. Educ.*, vol. 29, 2025.

[4] E. Guerra, F. Wilhelmi, M. Miozzo, and P. Dini, “The cost of training machine learning models over distributed data sources,” *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1111–1126, 2023.

[5] U. Farooq *et al.*, “Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes,” *J. Supercomput.*, vol. 80, no. 11, pp. 16334–16367, 2024.

[6] A. Mohasseb, E. Amer, F. Chiroma, and A. Tranchese, “Leveraging advanced NLP techniques and data augmentation to enhance online misogyny detection,” *Appl. Sci.*, vol. 15, no. 2, p. 856, 2025.

[7] L. Symeou, L. Louca, A. Kavadella, J. Mackay, Y. Danidou, and V. Raffay, “Development of Evidence-Based Guidelines for the Integration of Generative AI in University Education Through a Multidisciplinary, Consensus-Based Approach,” *Eur. J. Dent. Educ.*, vol. 29, no. 2, pp. 285–303, 2025.

[8] I. A. Salami, T. O. Adesokan-Imran, O. J. Tiwo, O. C. Metibemu, A. T. Olutimehin, and O. O. Olaniyi, “Addressing bias and data privacy concerns in AI-driven credit scoring systems through cybersecurity risk assessment,” *Asian J. Res. Comput. Sci.*, vol. 18, no. 4, pp. 59–82, 2025.

[9] N. Zhai and X. Ma, “The effectiveness of automated writing evaluation on writing quality: A meta-analysis,” *J. Educ. Comput. Res.*, vol. 61, no. 4, pp. 875–900, 2023.

[10] S. Li *et al.*, “FedScore: A privacy-preserving framework for federated scoring system development,” *J. Biomed. Inform.*, vol. 146, p. 104485, 2023.

[11] B. R. Yadav, “AI-Driven Exam Evaluation Systems: Challenges, Innovations, and Future Directions,” *Int. J. Electron. Autom.*, vol. 2, no. 2, pp. 7–13p, 2024.

[12] A. Pack, A. Barrett, and J. Escalante, “Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability,” *Comput. Educ. Artif. Intell.*, vol. 6, p. 100234, 2024.

[13] K. Bouziane and A. Bouziane, “AI versus human effectiveness in essay evaluation,” *Discov. Educ.*, vol. 3, no. 1, p. 201, 2024.

[14] A. Mizumoto and M. Eguchi, “Exploring the potential of using an AI language model for automated essay scoring,” *Res. Methods Appl. Linguist.*, vol. 2, no. 2, p. 100050, 2023.

[15] “Learning Agency Lab - Automated Essay Scoring 2.0.” 2024. [Online]. Available: https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2