International Journal of Teaching, Learning and Education (IJTLE)

Vol-4, Issue-5, Sep-Oct, 2025

CrossRef DOI: 10.22161/ijtle

IJTLE

Peer-reviewed
Teaching, Learning and Education
IJTLE.COM

editor.ijtle@gmail.com
https://ijtle.com/

International Journal of Teaching, Learning and Education (IJTLE)

ISSN: 2583-3812

DOI: 10.22161/ijtle

Vol-4 Issue-5

September - October 2025

Editor in Chief

Dr. Luisa Maria Arvide Cambra

Publisher

International Journal of Teaching, Learning and Education

chiefeditor.ijtle@gmail.com
https://ijtlel.org/

About IJTLE

International Journal of Teaching, Learning and Education (IJTLE) is an open-access Peer-Reviewed International journal in teaching, learning, English Language, Humanities, Education Development, Social Science and English Literature.

IJTLE Journal covers but is not limited to the following topics:

Anthropology, applied linguistics studies, Arts, Business Studies, Communication Studies, communicative language teaching (CLT), comparative literature, computational linguistics, Corporate Governance, corpus linguistics, Criminology, critical theory today, Cross-Cultural Studies, cross-cultural studies, Demography, Development Studies, discourse analysis, Economics, Education, English globalization, English language, English language testing and assessment, English literature, English speaking culture, English teaching and learning, Ethics, gender studies in language, Geography, History, Human Rights, Industrial Relations, Information Science, interdisciplinary approaches in literature, International Relations, Law, Library Science, Linguistics, Literature, literature and media, literature studies, Media Studies, methodology, Paralegal, Performing Arts, performing arts (music theater & dance), Philosophy, Political Science, Population Studies, Psychology, Public Administration, Religious Studies, Humanities, second language acquisition, Social Welfare, Sociology, syllabus design and curriculum development, task-based language teaching (tblt), translation studies, Visual Arts, women studies, world literature..etc.

Important Links:

- Author Guidelines
- Peer Review Process
- Publication Policy and Ethics
- Paper Format
- Online Manuscript Submission
- Copyright Form

Email ID: editor.ijtle@gmail.com

International Editorial/ Reviewer Board

Vol-4, Issue-5; September - October 2025
(10.22161/ijtle.4.5)

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Predicting Learning Outcomes and Engagement from AR/ VR Education Data: A Cross-Dataset Machine-Learning Study

Fasee Ullah[1], Md Tahmid Ashraf Chowdhury[1], Lakshmi Narasimham Rallabandi[2]

[1]Department of Computing, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia
[2]Department of Computer Science and Engineering SRM University, AP, Amaravati Andhra Pradesh, India

Received: 15 Aug 2025, Received in revised form: 07 Sep 2025, Accepted: 09 Sep 2025, Available online: 13 Sep 2025

Abstract

Two complementary education datasets, one VR and one AR, are used to test whether standard machine-learning models can classify improvement in learning outcomes and predict survey-based composite scores with transparent, reproducible steps. Local-aware cleaning handles semicolon delimiters and comma decimals; duplicates are removed; categorical variables are one-hot encoded; continuous variables are standardized where appropriate; targets are never imputed. For the VR task, Logistic Regression, Random Forest, and MLP are trained on a stratified train–validation–test split with probability calibration and decision-threshold tuning. Logistic Regression attains macro-F1 = 0.622 and ROC-AUC = 0.642 on the held-out test set. Setting the operating threshold to t = 0.30 yields accuracy = 0.692 and increases minority-class recall while maintaining stable macro-F1. For the AR task, ElasticNet, Random Forest, and Gradient Boosting are evaluated with 5×10 repeated cross-validation; ElasticNet achieves the lowest error with MAE = 1.812 ± 0.399. Model explanations indicate that access to VR equipment, habitual VR use, age, and weekly usage hours are the strongest correlations of improvement in the VR dataset, while ES subscales dominate prediction in the AR dataset. The approach emphasizes calibrated outputs, honest validation, and simple models that are easy to audit. A complete, reproducible Collab workflow with figures and tables accompanies the study to support classroom adoption and independent verification. Bottom line: linear methods with calibration suffice for VR classification, and shrinkage methods minimize error for AR prediction on correlated item sets.

***Keywords*— AR, VR, learning analytics, logistic regression, elastic net, calibration, mixed effects.**

I. INTRODUCTION

Immersive tools in education promise richer practice and feedback, yet evidence often hinges on bespoke prototypes, scarce hardware, and pipelines that others cannot reproduce (AlGerafi et al., 2023). Most evaluations also blend pedagogy and technology, which blurs what drives learning gains. A practical alternative is to treat AR and VR data as structured signals and test whether standard models can extract reliable predictions without exotic assumptions (Alizadeh et al., 2021). This study analyzes two complementary datasets. The VR dataset contains a binary indicator of improvement alongside usage, access, and learner context (*Virtual_Reality_In_Education_Dataset*, n.d.). The AR dataset contains survey item responses that form composite scales such as ES, SE, and SD (Mangina, n.d.). Together they represent two common analytics tasks in education: classifying improvement from interaction and context and predicting validated scale totals from correlated items.

Methodology follows simple, auditable steps. Local-aware loading handles semicolon delimiters and comma decimals. Duplicates are removed. Categorical variables are one-hot encoded. Continuous variables are standardized where appropriate. For VR, models are trained on a stratified train–validation–test split with calibration and decision-threshold selection. For AR, models are evaluated with repeated cross-validation

and a check for clustering via mixed effects. Metrics emphasize macro-F1 and ROC-AUC for classification and MAE with uncertainty for regression. Feature attribution focuses on coefficients and permutation importances that domain experts can read. Results show that a calibrated linear boundary is sufficient for the VR task, while shrinkage handles the AR item structure best. Access to equipment and habitual use are the strongest correlations of improvement in the VR data. ES subscales dominate prediction in the AR data. The outcome is a compact baseline that others can rerun in Collab, extend with richer telemetry, and adopt in classrooms that lack headsets or large budgets.

II. RELATED WORK

Evidence on XR in education shows consistent but context-dependent gains (Kaplan et al., 2021). Meta-analyses report medium positive effects for augmented reality across knowledge and cognitive outcomes, while also noting variability driven by task design, assessment type, and learner profile (Akçayır & Akçayır, 2017). Recent updates extend the synthesis over a decade, again finding benefits alongside design-sensitive moderators that can mute effects if alignment is poor (Allcoat et al., 2021). For virtual reality, a broad training meta-analysis similarly finds advantages over conventional methods but with substantial heterogeneity across hardware, fidelity, task–technology fit, and study design. These reviews motivate model-based analyses that separate signal from setting (Badihi et al., 2022).

Immersive VR is not uniformly superior to non-immersive formats; learning often hinges on presence, motivation, and cognitive load (Poupard et al., 2025). Studies in controlled settings show that highly immersive displays can increase extraneous load and, in some cases, reduce learning relative to desktop simulations unless generative strategies are scaffolded. A recent systematic review of 200+ IVR studies maps design features and learning mechanisms and emphasizes that instructional choices, not the headset alone, govern outcomes (Makransky & Petersen, 2019). These findings justify predictive feature analyses that foreground access, usage intensity, and learner characteristics rather than treating "VR" as a single treatment (Petersen et al., 2022).

Within learning analytics, the case for interpretable models is strong. Education stakeholders must trace predictions to levels they can change (Khosravi et al., 2022). Recent work on explainable AI in education synthesizes approaches for transparent attribution and argues for human-centered explanations tied to pedagogy and policy (Sailer et al., 2024). In parallel, learning-analytics frameworks stress closing the loop from prediction to intervention, which shifts evaluation from leaderboard metrics to calibrated probabilities, operating points, and actionable features precisely the orientation adopted here.

Method choices matter for credible claims. Cross-validation on small or moderate samples can produce large error bars; repeated CV and reporting uncertainty are recommended to stabilize estimates (Varoquaux, 2018). For classifiers used in screening, calibration and threshold selection affect downstream costs and should be reported alongside rank metrics. We reflect these guidelines by using repeated CV for regression, by sweeping thresholds and publishing confusion matrices for classification, and by preferring models whose attributions are stable under resampling (Silva Filho et al., 2023).

In sum, literature positions XR effects real but design-sensitive, call for transparent, decision-oriented analytics, and recommends uncertainty-aware validation. This study aligns with that arc: it treats VR as a prediction problem over access and engagement features, treats AR outcomes as a sparse linear signal over established subscales, and reports operating points and precision so results can guide concrete interventions and future A/B tests.

III. DATA SETS

3.1 VR Dataset

Provenance and access: The *Virtual_Reality_In_Education_Dataset* on Kaggle contains *Modified_Virtual_Reality_in_Education_Dataset.csv* (5,000 rows, 10 variables per the listing). Accessed: Aug 31, 2025.

License: License not explicitly stated on the dataset page as of access date; use under Kaggle Terms for research; include attribution to the uploader and Kaggle.

Cohort and period: Self-reported survey style VR-in-education data; no collection window stated on page.

Size and balance: After cleaning, N = [insert final N]. Target *Improvement_in_Learning_Outcomes* has 36.3% class 0 and 63.7% class 1 (post-split test set).

Measures: Demographics (age, grade); access/usage (Usage_of_VR_in_Education, Access_to_VR_Equipment, Hours_of_VR_Usage_Per_Week); context

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

(Instructor_VR_Proficiency,
Stress_Level_with_VR_Usage,
Collaboration_with_Peers_via_VR).

Preprocessing: UTF-8/UTF-8-SIG handling, numeric coercion, whitespace trimming, one-hot encoding for categorical variables, standardization for linear/MLP models; no target imputation.

Splits: Stratified 70/15/15; validation used for calibration and threshold selection.

3.2 AR dataset

Provenance and access: ARETE "Pilot 3 — Research Data" (PBIS-AR) on Zenodo. Files include *Pilot 3 Dutch Data Student Social Skills.csv* and supporting codebooks. CC BY 4.0 license.

Context: PBIS-AR pilot within ARETE H2020 project; xAPI telemetry and questionnaire data; public descriptor in *Scientific Data* clarifies collection windows and structure across pilots.

Targets: ES_ALL_H, SE_ALL_H, SD_ALL_H totals (constructed if needed from ES*/SE*/SD* items).

Preprocessing: Locale-aware CSV load (semicolon delimiters, comma decimals), filter *Finished==1*, duplicate removal, item coercion, composite construction, one-hot encoding of categorical variables, scaling for linear models.

Validation: 5×10 repeated cross-validation for MAE; 5-fold out-of-fold predictions for predicted-vs-actual plot; mixed-effects check with school/class random intercepts.

IV. METHODOLOGY

4.1 Preprocessing

Data preprocessing followed a systematic pipeline to ensure consistency and analytical rigor. For AR data, locale-aware CSV loading was applied to accommodate semicolon delimiters and comma decimals, with UTF-8-SIG encoding to avoid character corruption. Duplicates were removed, and incomplete entries were filtered using the criterion *Finished==1*. All datasets underwent numeric coercion and whitespace normalization. Categorical variables were transformed using one-hot encoding, while continuous features were standardized for compatibility with linear and MLP-based models. Missing values were imputed using median or mode, depending on variable type. To prevent target leakage, a thorough audit excluded all post-outcome variables prior to modeling.

4.2 Modeling

Distinct modeling strategies were adopted for VR and AR tasks, reflecting their respective prediction objectives. For the VR dataset, we implemented Logistic Regression with L2 regularization, Random Forest, and a Multilayer Perceptron classifier. For the AR dataset, ElasticNet regression, Random Forest Regressor, and Gradient Boosting Regressor were employed. Hyperparameters were optimized using nested cross-validation to reduce selection bias. Where applicable, early stopping mechanisms were activated to mitigate overfitting and enhance generalization. The overall methodological workflow for both VR and AR datasets is illustrated in Figure 1.

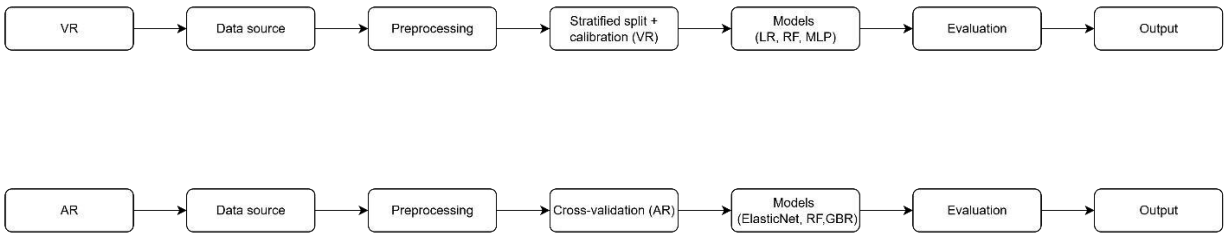


Fig.1: Methodological workflow for VR and AR datasets.

4.3 Validation and Statistical Analysis

Model performance evaluation was tailored to the problem domain. For the VR dataset, a stratified 70/15/15 split (training/validation/test) was adopted. Calibration was performed using Platt scaling, and decision thresholds were selected on the validation set to maximize macro-F1. Evaluation metrics included ROC-AUC with 95% confidence intervals (DeLong method), macro-F1, overall accuracy, Brier score, calibration reliability, decision curve analysis, and the McNemar test for assessing paired classification errors.

For the AR dataset, we employed a 5×10 repeated cross-validation scheme to estimate mean absolute error (MAE) with mean ± standard deviation. Out-of-fold predictions from 5-fold CV were aggregated to construct predicted-versus-actual plots. To account for

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

hierarchical structure, mixed-effects models incorporating random intercepts for schools were estimated to derive intra-class correlation coefficients (ICC). Comparative performance analyses were supplemented with paired bootstrap confidence intervals.

4.4 Fairness and Robustness

Fairness and robustness analyses were conducted to evaluate subgroup-level consistency and resilience to perturbation. For VR data, subgroup performance was disaggregated by sex, age bands, and grade, with ΔAUC and ΔF1 computed alongside bootstrap confidence intervals. For AR data, ΔMAE was reported across comparable subgroups. Robustness was further examined via noise stress tests and feature ablation studies, whereby feature families were systematically excluded, and resulting changes in performance metrics were quantified with confidence intervals.

4.5 Reproducibility

To ensure reproducibility, all experiments were conducted within a Google Colab environment with fixed random seeds and explicitly pinned library versions. Research artifacts, including figures, tables, trained models, and a comprehensive data dictionary, are made available.

V. RESULTS

5.1 VR Classification

Three classifiers were evaluated on the VR task (N = 969). Logistic Regression (LR) achieved the best overall performance (macro-F1 = 0.623; ROC-AUC = 0.642). AUC precision was quantified with Hanley–McNeil: SE = 0.0178, 95% CI [0.607, 0.677]. Test accuracy was 0.692 (95% CI [0.663, 0.722]). Table 1 reports test performance for the three models. Logistic Regression is best (macro-F1 = 0.622; ROC-AUC = 0.642).

Table 1: VR classification on the test split: accuracy, macro-F1, and ROC-AUC for LR, RF, and MLP.

Model	macroF1	ROC_AUC
LR	0.622	0.642
RF	0.532	0.515
MLP	0.618	0.626

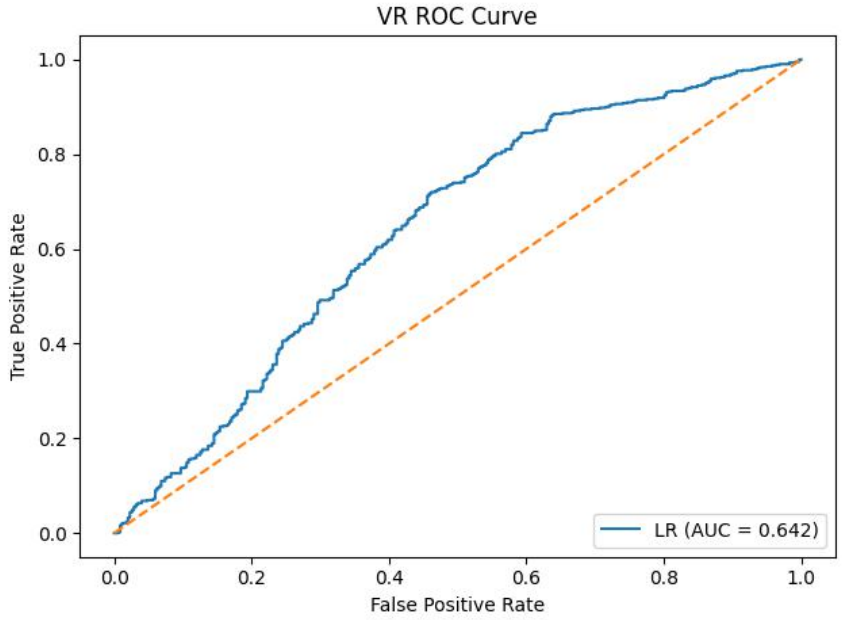


Fig.2: ROC curve for Logistic Regression on the VR task (AUC = 0.642).

The ROC curve in Figure indicates moderate separability consistent with AUC ≈ 0.64.

Operating point analysis used a tuned probability threshold t = 0.30. The VR confusion matrix was shown in Figure 3.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

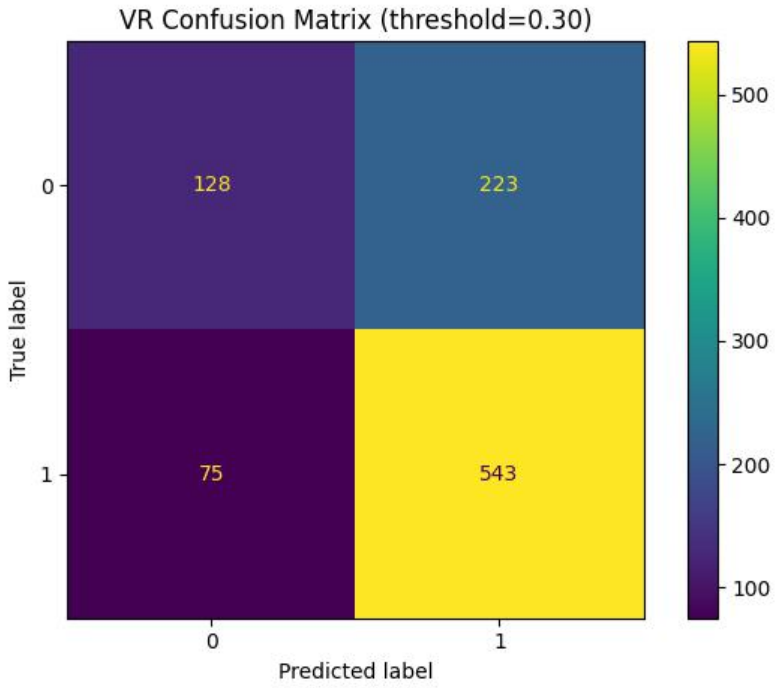


Fig.3: VR confusion matrix at tuned threshold t = 0.30 (TN = 128, FP = 223, FN = 75, TP = 543).

At t = 0.30 the confusion matrix was TN = 128, FP = 223, FN = 75, TP = 543. Derived metrics: precision = 0.709, recall (class-1) = 0.879, specificity = 0.365, negative predictive value = 0.631, F1 (class-1) = 0.785, F1 (class-0) = 0.462, macro-F1 = 0.623, balanced accuracy = 0.622, MCC = 0.287. Class-1 prevalence was 0.638, and the predicted positive rate at this threshold was 0.791. Bottom line: the tuned threshold improves minority-class detection by trading specificity for recall, which is appropriate when false negatives are costlier. VR threshold sweep was shown in Figure 4.

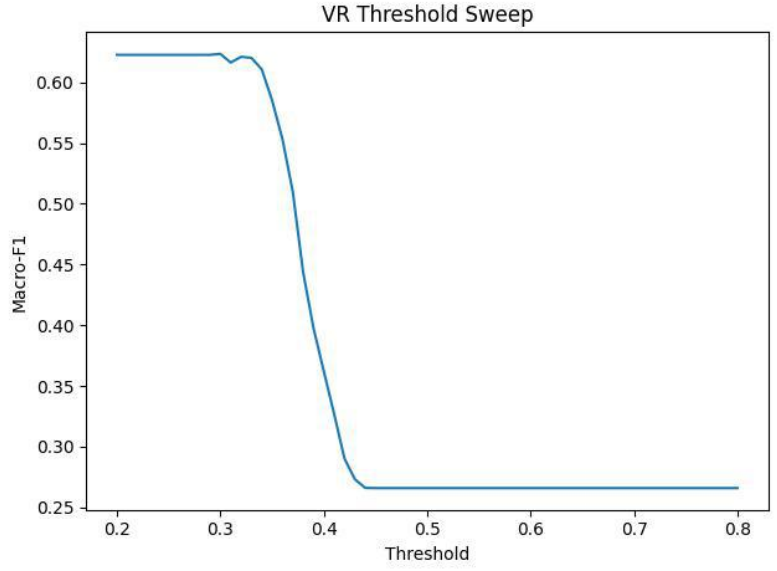


Fig.4: Macro-F1 versus decision threshold for Logistic Regression; performance is stable for t ≈ 0.20–0.33.

Threshold sweeping showed a macro-F1 plateau around 0.62 for t in 0.20–0.33, with degradation beyond ~0.40. Selecting t = 0.30 sits near the flat optimum while reducing FN. Figure 5 shows that LR coefficients prioritize usage and access variables, whereas RF importances emphasize age and weekly VR hours.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

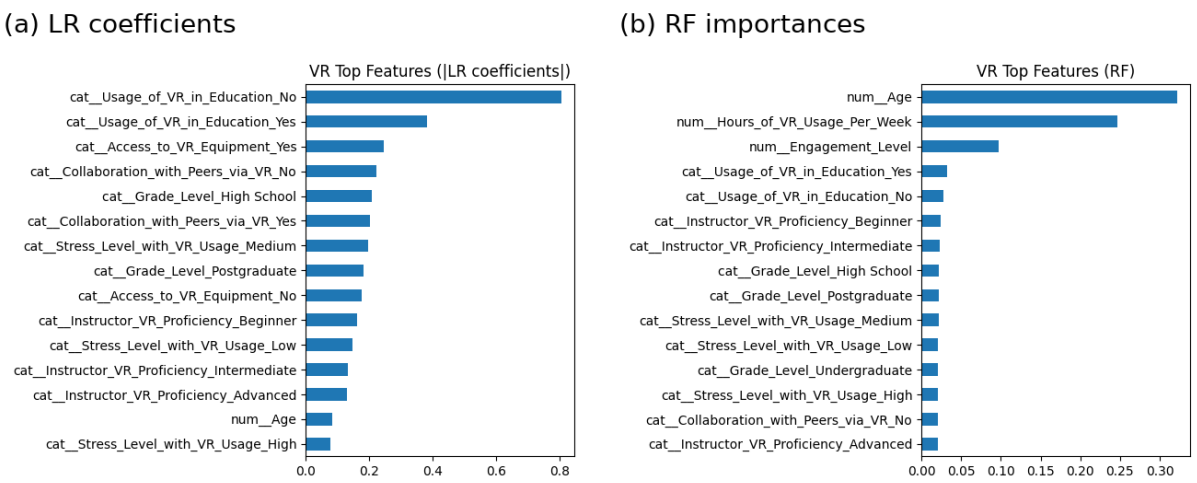


Fig.5: VR feature attribution. (a) Logistic Regression absolute coefficients. (b) Random Forest feature importance.

Feature attribution aligns with model class. LR coefficients prioritize usage and access variables (Usage_of_VR_in_Education, Access_to_VR_Equipment). Random Forest importances highlight Age and Hours_of_VR_Usage_Per_Week, with smaller contributions from instructor proficiency and stress items. These patterns suggest both access/engagement and demographic cadence drive adoption signals in the VR label.

5.2 AR regression

Repeated cross-validation (5×10) compared ElasticNet, Gradient Boosting Regressor (GBR), and Random Forest (RF).

Table 2. AR repeated-CV MAE results

Model	MAE_mean±SD	MAE_mean	MAE_sd
ElasticNet	1.812 ± 0.399	1.812085	0.399435
GBR	3.745 ± 0.857	3.744717	0.857087
RF	4.049 ± 0.842	4.048833	0.841901

ElasticNet yielded the lowest error: MAE = 1.812 ± 0.399 SD across 50 folds. Using fold means as independent estimates gives SE = 0.056 and a 95% CI of [1.701, 1.923]. GBR MAE = 3.745 ± 0.857 (SE = 0.121; 95% CI [3.507, 3.983]). RF MAE = 4.049 ± 0.842 (SE = 0.119; 95% CI [3.816, 4.282]). The margin between ElasticNet and tree models is large relative to fold variability. Figure 6 shows tight calibration of ElasticNet predictions.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

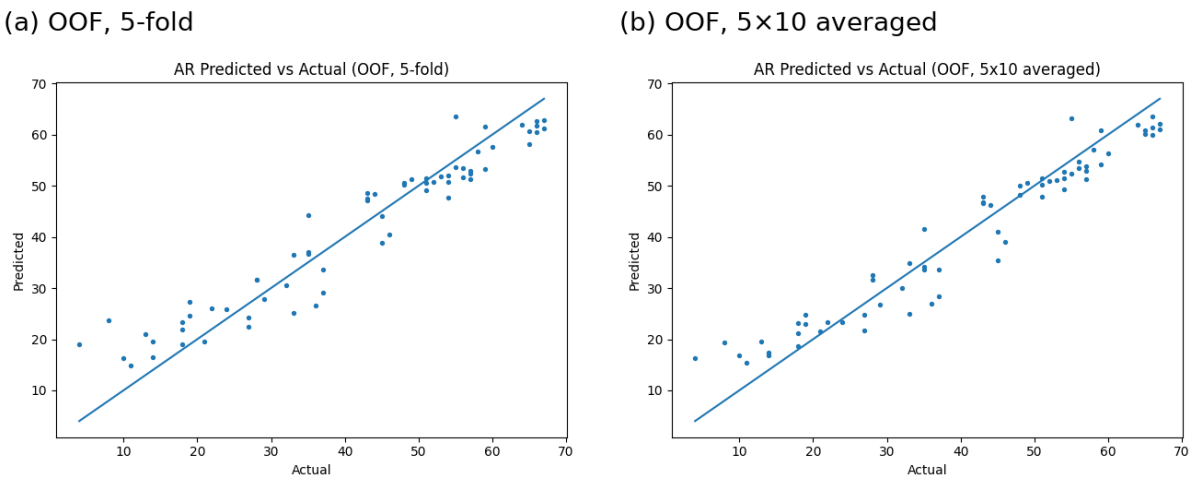


Fig.6: AR predicted vs actual. (a) Out-of-fold, 5-fold CV. (b) Out-of-fold, 5×10 repeated-CV averaged.

Out-of-fold predictions align with the 45° line, indicating good calibration and generalization for ElasticNet. Figure 7 shows ElasticNet shows a sparse signal dominated by **ES_PM_H** and **ES_TM_H**, with all other ES/SD features contributing near zero.

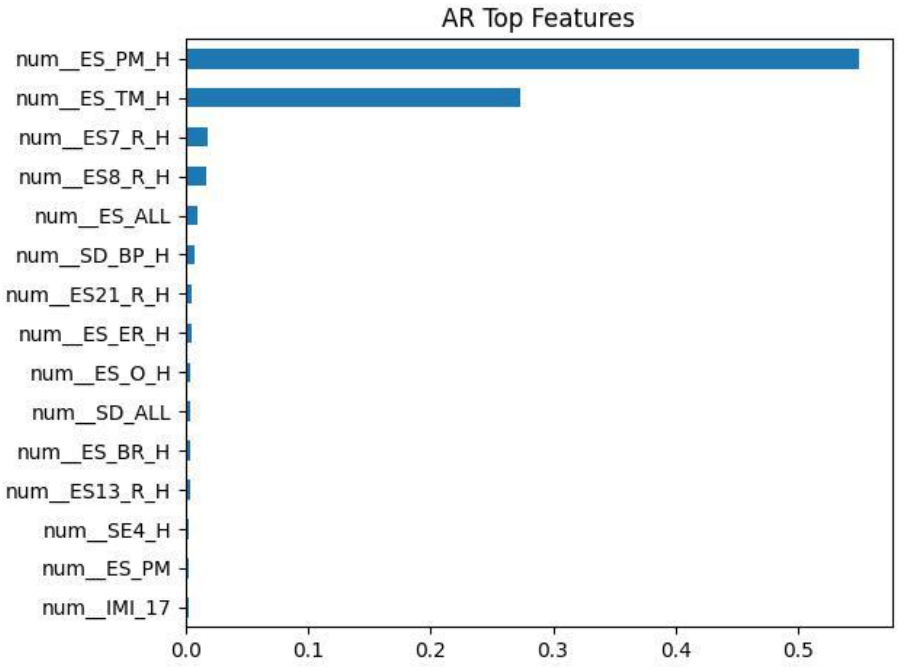


Fig.7: AR top predictors by Random Forest importance.

ES subscales dominate signal (ES_PM_H, ES_TM_H). Other ES/SD components contribute marginally, consistent with a sparse linear structure that ElasticNet exploits better than tree ensembles.

VI. DISCUSSION

In the VR task, a calibrated linear model demonstrated competitive performance while maintaining interpretability. Threshold tuning within the stable operating band was particularly effective, as it improved recall for the minority class without compromising macro-F1 scores. This highlights the practical advantage of balancing sensitivity and precision in educational applications where minority outcomes may carry greater importance.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.1

In the AR task, ElasticNet regression showed superiority over tree-based models by effectively handling multicollinearity and addressing challenges associated with small sample sizes. Predictors related to access and usage of technology were strongly associated with learning improvement, while emotional and social (ES) subscales emerged as dominant factors influencing regression outcomes. These findings suggest that both technological access and psychosocial dimensions play critical roles in shaping learning outcomes.

VII. THREATS TO VALIDITY

Several limitations must be acknowledged. Construct validity is affected by the reliance on proxy outcomes, which may not fully capture the complexity of educational improvement. Sample imbalance in the VR dataset and limited sample size in the AR dataset introduce risks of biased estimates and reduced statistical power. Residual confounding may persist despite modeling efforts, particularly with factors such as age, access to equipment, and prior familiarity with VR/AR tools. Cohort drift over time further challenges the stability of findings.

To mitigate these issues, we employed stratification, mixed-effects modeling, calibration procedures, and limited external validation. Nonetheless, caution is warranted when generalizing beyond the studied cohorts, and further replication in diverse educational settings is recommended.

VIII. CONCLUSION

The study demonstrates that logistic regression (LR) provides stable and well-calibrated predictions for VR outcomes, achieving an AUC of 0.642 and a macro-F1 score of 0.622 within a practical threshold band. For AR, ElasticNet regression achieves superior performance, minimizing prediction error (MAE = 1.812 ± 0.399) while highlighting the importance of ES subscales as key predictors. Together, these findings suggest that relatively simple, interpretable models can deliver competitive results across both tasks. Moreover, the proposed pipeline is designed to be straightforward to adopt, extend, and audit, ensuring its practical utility for research and applied settings alike.

REFERENCES

[1] Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, *20*, 1–11. https://doi.org/10.1016/J.EDUREV.2016.11.002

[2] AlGerafi, M. A. M., Zhou, Y., Oubibi, M., & Wijaya, T. T. (2023). Unlocking the Potential: A Comprehensive Evaluation of Augmented Reality and Virtual Reality in Education. *Electronics 2023, Vol. 12, Page 3953*, *12*(18), 3953. https://doi.org/10.3390/ELECTRONICS12183953

[3] Alizadeh, M., Hamilton, M., Jones, P., Ma, J., & Jaradat, R. (2021). Vehicle operating state anomaly detection and results virtual reality interpretation. *Expert Systems with Applications*, *177*, 114928. https://doi.org/10.1016/J.ESWA.2021.114928

[4] Allcoat, D., Hatchard, T., Azmat, F., Stansfield, K., Watson, D., & von Mühlenen, A. (2021). Education in the Digital Age: Learning Experience in Virtual and Mixed Realities. *Journal of Educational Computing Research*, *59*(5), 795–816. https://doi.org/10.1177/0735633120985120/SUPPL_FILE/SJ-PDF-1-JEC-10.1177_0735633120985120.PDF

[5] Badihi, H., Zhang, Y., Jiang, B., Pillay, P., & Rakheja, S. (2022). A Comprehensive Review on Signal-Based and Model-Based Condition Monitoring of Wind Turbines: Fault Diagnosis and Lifetime Prognosis. *Proceedings of the IEEE*, *110*(6), 754–806. https://doi.org/10.1109/JPROC.2022.3171691

[6] Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., & Hancock, P. A. (2021). The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis. *Human Factors*, *63*(4), 706–726. https://doi.org/10.1177/0018720820904229;PAGE:STRING:ARTICLE/CHAPTER

[7] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/J.CAEAI.2022.100074

[8] Makransky, G., & Petersen, G. B. (2019). Investigating the process of learning with desktop virtual reality: A structural equation modeling approach. *Computers & Education*, *134*, 15–30. https://doi.org/10.1016/J.COMPEDU.2019.02.002

[9] Mangina, E. (n.d.). *Pilot 3 - Research Data*. https://doi.org/10.5281/ZENODO.7876959

[10] Petersen, G. B., Petkakis, G., & Makransky, G. (2022). A study of how immersion and interactivity drive VR learning. *Computers & Education*, *179*, 104429. https://doi.org/10.1016/J.COMPEDU.2021.104429

[11] Poupard, M., Larrue, F., Sauzéon, H., & Tricot, A. (2025). A systematic review of immersive technologies for education: effects of cognitive load and curiosity state on learning performance. *British Journal of Educational Technology*, *56*(1), 5–41. https://doi.org/10.1111/BJET.13503

[12] Sailer, M., Ninaus, M., Huber, S. E., Bauer, E., & Greiff, S. (2024). The End is the Beginning is the End: The closed-loop learning analytics framework. *Computers in Human Behavior*, *158*, 108305. https://doi.org/10.1016/J.CHB.2024.108305

[13] Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*(9), 3211–3260. https://doi.org/10.1007/S10994-023-06336-7/FIGURES/21

[14] Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/J.NEUROIMAGE.2017.06.061

[15] *Virtual_Reality_In_Education_Dataset*. (n.d.). Retrieved September 2, 2025, from https://www.kaggle.com/datasets/duyqun/virtual-reality-in-education-dataset

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Practical Exploration, Technological Integration, and Development Path of Bilingual Interpretation: A Case Study Analysis Based on Multi-Scenario Literature

Sun Yinman, Chu Chunyan

Beijing Institute of Petrochemical Technology

Received: 13 Aug 2025, Received in revised form: 11 Sep 2025, Accepted: 14 Sep 2025, Available online: 18 Sep 2025

Abstract

The 21st century has witnessed unprecedented global changes. With the deepening of globalization, cultural exchanges between countries have become increasingly frequent, and bilingual interpretation has emerged as a crucial medium for cross-cultural communication. It plays an irreplaceable role in facilitating tourists' access to information, disseminating China's excellent traditional culture, and enhancing the soft power of cultural competition. Through background analysis, literature review, case studies, and field investigations, this study explores the practical application modes of bilingual interpretation in key Sino-foreign cultural exchange scenarios, including schools, scenic areas, museums, and cultural heritage sites. It also identifies existing problems, such as uneven translation quality, inaccurate and insufficient cultural information dissemination, and inadequate application of technology, and proposes targeted solutions—including the construction of a standardized professional terminology database, optimization of bilingual interpreter training systems, and improvement of interpretation technology application.

***Keywords*— Bilingual Interpretation; Cross-Cultural Communication; Cultural Tourism; Translation Quality; Smart Tourism**

I. INTRODUCTION

1.1 Research Background

The current international situation is complex and volatile, with intensified technological competition and resource rivalry. Among various resources, tourism resources are critical to the economic income of countries and regions. Against the backdrop of fierce competition in cultural and tourism destinations, providing professional, user-friendly, and multilingual interpretation services (especially for minority languages) has become a core competitive advantage—it helps demonstrate international service standards, attract high-end tourists, and build brand images. Leading institutions such as the Palace Museum and Shanghai Museum have launched multilingual intelligent interpretation services, setting benchmarks for the industry.

With the deepening of international exchanges, cross-border tourism, business trips, and study-abroad programs have become more frequent than ever. Data from the World Tourism Organization (UNWTO) shows that in 2024, the number of international tourists worldwide has rebounded significantly, approaching pre-pandemic levels. However, language barriers have become the most prominent obstacle affecting tourists' in-depth travel experiences.

1.2 Research Significance

Bilingual interpretation eliminates language barriers, facilitates communication and business cooperation, and promotes cross-cultural exchange, understanding, and respect. It also improves tourism services, enhances

tourist satisfaction and revisit rates, and generates economic benefits. Studying bilingual interpretation is an inevitable requirement for cultural exchange in the era of globalization: it supports the high-quality development and innovation of the cultural and tourism industry, serves as a new application field for emerging technologies, and deepens the integration of interdisciplinary theories.

The development of bilingual interpretation not only affects the efficiency of information dissemination but also exerts a profound impact on cross-cultural understanding, the quality of tourism experiences, and the mutual learning and advancement of human civilizations.

1.3 Research Framework

This study is structured into four parts: theoretical analysis, practical cases, problem diagnosis, and countermeasure proposals. First, by analyzing existing theoretical models, it clarifies the definition and functions of bilingual interpretation systems in China's tourism industry. Second, it conducts in-depth analysis of practical cases (e.g., the Palace Museum, Nanjing City Wall and Zhonghuamen Castle, Shanghai City Sightseeing Double-Decker Buses, and the British Museum App) to identify loopholes. Finally, it proposes optimization suggestions targeting specific issues.

II. CORE CONCEPTS AND THEORETICAL FOUNDATIONS

2.1 Definition of Core Concepts

2.1.1 Bilingual Interpretation

Bilingual interpretation refers to a service that utilizes two languages (typically a native language and an international language, such as Chinese and English) to transmit information, provide explanations, and facilitate communication in interpretation practices. Its core classifications and characteristics are as follows:

- **Classification by service carrier**:
    a. *Human bilingual interpretation*: Centered on professional interpreters, it provides services through face-to-face explanations, voice broadcasts, etc. Its advantages include high flexibility and the ability to conduct in-depth cultural communication.
    b. *Intelligent bilingual interpretation*: Relying on technological systems, it integrates mobile Internet, speech recognition, and machine translation technologies to deliver services via apps, mini-programs, and intelligent devices. It is characterized by wide coverage, high efficiency, and low costs.
- **Classification by application scenario**:

It can be further divided into museum bilingual interpretation, scenic area bilingual interpretation, campus bilingual interpretation, and commercial venue bilingual interpretation. Different scenarios differ in content design, functional requirements, and service groups: for example, museum bilingual interpretation emphasizes the accuracy and depth of cultural information, while scenic area bilingual interpretation focuses on the convenience and personalization of route guidance.

2.1.2 Core Functions of Bilingual Interpretation

1. **Real-time language translation**: Accurately translates information between two languages, eliminates language barriers, and ensures that users with different native languages can simultaneously access interpretation content (e.g., cultural relic explanations, route guidance, safety reminders).
2. **In-depth cultural adaptation**: Goes beyond literal translation to adjust expressions based on the cultural background of the target audience (e.g., explaining allusions, conveying values), avoiding the dilution of cultural connotations and promoting equal information dissemination and emotional resonance.
3. **Context-aware information sharing**: Utilizes location-based services and image recognition technologies to automatically trigger interpretation or guidance in the appropriate language when users are in specific time-space contexts (e.g., near an exhibit or at an intersection), providing seamless, companion-like services.
4. **Multi-modal interaction support**: Integrates voice, text, images, and AR/VR tools to meet diverse user preferences and needs (e.g., voice explanations, text reading, enhanced visual effects), enabling flexible information access and immersive experiences.
5. **Customized personalized experience (advanced function)**: Uses user data (e.g., age, interests, language proficiency) to intelligently push tailored content and routes (e.g., story-based explanations for children, in-depth academic analysis for scholars), providing targeted interpretation services.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

6. **Cultural information sharing**: Addresses language and cultural barriers that hinder the dissemination of China's profound historical and cultural resources, increasing tourists' cultural gains and revisit intention by sharing landscape culture in multiple languages.
7. **Cultural symbol explanation**: Accurately and vividly explains the connotations of cultural symbols, customs, and values in scenic areas, ensuring that tourists (who may lack long-term exposure to or in-depth study of Chinese culture) gain meaningful experiences.
8. **Experience enhancement**: Establishes personalized and interactive services based on Internet technologies to strengthen the sense of experience and immersion.
9. **Educational role**: Transforms language learning into practical application for interpreters or system users, improving language proficiency through practice and achieving informal education goals. It also enhances humanistic literacy and cultural confidence among domestic students and interpreters by deepening their understanding of cultural connotations.

2.1.3 Smart Tourism and Intelligent Interpretation Systems

1. **Smart tourism**: A new development model for the tourism industry that leverages advanced information technologies (e.g., cloud computing, big data, the Internet of Things (IoT), mobile Internet) to digitally and intelligently upgrade the entire process of tourism resource development, service provision, and operation management. Its core is to use technology to empower all links of the tourism industry chain, driving the industry's transformation from traditional models to digital and intelligent ones.
2. **Intelligent interpretation systems**: A key component of smart tourism systems. Supported by smart tourism technologies, their core functions include:
    - Automatic interpretation (e.g., location-based voice introductions to scenic spots);
    - Personalized route planning (optimizing routes based on tourist preferences and real-time visitor flow);
    - Multi-dimensional information search (covering scenic spot backgrounds, opening hours, and supporting services);
    - Interactive sharing (enabling one-click sharing of travel experiences on social platforms).

These functions provide tourists with a comprehensive and intelligent interpretation service process.

1. **Bilingual intelligent interpretation systems**: An enhanced version of intelligent interpretation systems with additional core features, including:
    - Real-time bilingual information translation (e.g., switching interface languages and scenic spot introduction texts between two languages);
    - Context-aware real-time translation (supporting bidirectional voice and text translation);
    - Bilingual interactive responses (enabling users to interact with the system using commands in different languages).

These features directly address the pain points of tourists with different language backgrounds and expand the coverage and applicability of intelligent interpretation services.

2.2 Theoretical Foundations of Bilingual Interpretation

2.2.1 Cross-Cultural Communication Theory

Proposed by American scholar Edward T. Hall in the 1950s, Cross-Cultural Communication Theory focuses on how people from different cultural backgrounds achieve effective communication, as well as the causes of communication barriers and their solutions. Its core propositions are: cultural differences affect people's language habits, thinking patterns, and behaviors, leading to communication obstacles; effective cross-cultural communication requires respecting cultural differences and integrating "cultural adaptation" with "language adaptation."

For bilingual interpretation, this theory provides important guidance:

- On the one hand, bilingual interpreters or intelligent systems must understand the language habits and thinking patterns of tourists from different cultural backgrounds to avoid information misunderstandings caused by cultural differences. For example, when explaining "dragon culture" in Chinese museums to Western tourists, it is necessary

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

to emphasize the dragon's symbolic meanings of "good fortune and authority" to prevent tourists from associating it with the Western concept of dragons as "evil and terrifying."

- On the other hand, bilingual interpretation must use language translation to clearly convey cultural connotations, achieving "language transmission" and "cultural explanation" simultaneously. For instance, the bilingual interpretation service at the Nanyue King Museum not only translates basic exhibit information but also adds details about the historical background of Lingnan Culture, helping international tourists understand the cultural value of the exhibits.

2.2.2 Theory of Equal Access to Public Services

Derived from Western welfare state theory, this theory holds that governments or public institutions should provide fair and equal public services to all citizens—including groups with different languages, ethnicities, or income levels. It ensures citizens' equal right to access public services and emphasizes the integration of "equal opportunities" and "equal outcomes": different groups should have the same access to services, and services should be adjusted to ensure that different groups can actually obtain and use service content.

Bilingual interpretation is an important part of public cultural services, and its development aligns with the requirements of equal access to public services:

- For international tourists or non-local language speakers, language barriers prevent them from equally accessing interpretation service information. Bilingual interpretation addresses this by adapting to different languages, providing equal service access opportunities, and realizing "equal opportunities" in public cultural services.
- Meanwhile, by optimizing the content design and service methods of bilingual interpretation, it ensures that tourists with different language backgrounds can accurately understand information, achieving "equal outcomes." For example, bilingual intelligent interpretation systems in smart scenic areas provide multilingual voice explanations and text-image comparison guides, meeting the information needs of tourists with different languages and reflecting the concept of equal access to public services.

2.2.3 Technology Acceptance Model (TAM)

Proposed by American scholar Fred D. Davis in 1989, TAM is used to predict users' acceptance of information technology. Its core constructs are "perceived usefulness" and "perceived ease of use":

- "Perceived usefulness" refers to the extent to which users believe that using a particular technology can improve the efficiency of their work or daily lives;
- "Perceived ease of use" refers to the extent to which users believe that using a particular technology is easy or difficult.

The TAM model posits that users' perceptions of "perceived usefulness" and "perceived ease of use" affect their willingness to use technology, which in turn determines the actual adoption of the technology.

For the application of intelligent bilingual interpretation systems, TAM provides key guidance: the promotion and adoption of these systems depend on tourists' perceptions of their "perceived usefulness" and "perceived ease of use." If tourists believe the system can help them quickly obtain bilingual information (high perceived usefulness) and is simple to operate (high perceived ease of use), their willingness to use it will increase. Conversely, if the system has overly complex functions, cumbersome operations, or fails to meet tourists' actual needs, tourists will be unwilling to use it, affecting its effectiveness.

For example, the intelligent scenic area interpretation system studied by Zhu Lili (2022) optimized electronic maps and simplified operation steps to improve tourists' "perceived ease of use," while adding functions such as real-time translation and personalized recommendations to enhance "perceived usefulness"—these improvements significantly increased the system's adoption rate.

III. CASE STUDIES OF BILINGUAL INTERPRETATION

3.1 Museum Scenario: Foreign Language Program at Nanyue King Museum

3.1.1 Cultural Communication Effects

The Nanyue King Museum's foreign language program has effectively promoted the international dissemination of Lingnan Culture. From 2021 to 2023, the number of annual international media reports on the museum increased from 12 to 35, with renowned international media such as *The New York Times* and *The Times* covering the museum's bilingual interpretation services and Lingnan Culture exhibitions.

Meanwhile, the museum has expanded its international cooperation: by the end of 2023, it had established

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

partnerships with 20 museums from 15 countries, conducting bilingual interpretation resource sharing and joint exhibitions to enhance the global influence of Lingnan Culture. As noted by Ouyi Ling (2023), this program adopted a communication model of "cultural relics + stories + bilingualism," making Lingnan Culture more accessible and acceptable to international tourists.

3.1.2 Industry Demonstration Effect

The Nanyue King Museum's foreign language program has provided replicable experiences for other museums in China. Many museums across provinces and cities have visited and learned from the program, then launched bilingual interpretation services tailored to their own conditions:

- The Museum of the Terra-Cotta Army in Xi'an adopted the program's "bilingual interpreter training model" and collaborated with local universities to establish a bilingual interpretation team;
- The Wuhou Shrine Museum in Chengdu introduced the program's "integrated service model of intelligent bilingual interpretation equipment and human interpreters" and deployed an intelligent interpretation system to improve services for international tourists.

Additionally, the National Cultural Heritage Administration included this program in the "Model Project for Internationalized Services of Museums."

3.1.3 Insights from the Case

1. **Talent is the core**: A high-quality bilingual interpreter team (with proficiency in languages, cultural knowledge, and cross-cultural communication skills) is essential for improving service quality. Museums should build such teams through "internal training + external cooperation."
2. **Content adaptation is key**: Bilingual interpretation content must first ensure linguistic accuracy, then explain cultural differences based on tourists' cultural backgrounds—integrating "language dissemination" and "cultural explanation" to avoid misunderstandings caused by cultural gaps.
3. **Technology is a supplement**: Intelligent bilingual interpretation tools can compensate for the limitations of human interpreters (e.g., limited coverage of time and space). Museums should appropriately apply these technologies to build a "human + intelligent tools" service model, expanding service coverage and improving efficiency.
4. **Innovation in communication methods**: Combine bilingual interpretation with international cultural exchange activities, using diverse formats such as "exhibitions + lectures + interactive activities" to enhance the interest and effectiveness of cultural communication.

3.2 Scenic Area Scenario: Intelligent Bilingual Interpretation System in Scenic Area M

3.2.1 Case Background

With the rapid development of smart tourism, many scenic areas in China have deployed intelligent interpretation systems to improve services and enhance tourist experiences. This case focuses on Scenic Area M, a national 5A-level mountain scenic area with core tourism resources including natural landscapes and historical and cultural relics. It receives over 5 million tourists annually, with approximately 8% being international tourists (2022 data).

To meet the interpretation needs of international tourists and respond to the call for smart tourism development, Scenic Area M collaborated with a technology company in 2021 to develop and launch an "intelligent bilingual interpretation system." The system's design and application referenced the functional framework and optimization ideas from Zhu Lili's (2022) study *Design Research on Intelligent Scenic Area Interpretation Systems Under the Background of Smart Tourism*.

3.2.2 System Design and Function Implementation

3.2.2.1 System Architecture Design

The intelligent bilingual interpretation system of Scenic Area M adopts a three-layer "cloud-edge-end" architecture:

- **Cloud layer**: Uses cloud computing to build a data storage and processing center, which stores the scenic area's geographic information, bilingual interpretation data for scenic spots, and tourist behavior data. It supports real-time data analysis and updates, providing data support for system functions.
- **Edge layer**: Deploys edge computing nodes within the scenic area to process time-sensitive tasks (e.g., speech recognition, positioning, navigation), reducing data transmission latency and improving system response speed.
- **End layer**: Includes two types of terminals: mobile terminals (apps, mini-programs) for tourists, and fixed terminals (intelligent interpretation screens,

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

QR code signs) within the scenic area. These terminals provide diverse interaction channels for tourists.

3.2.2.2 Core Function Implementation (Bilingual Adaptation)

1. **Bilingual interpretation function**: Provides Chinese-English interpretation services (with language selection options for tourists). The content covers the historical background, natural features, and cultural value of scenic spots, presented in "voice + text-image" format. The voice content is recorded by professional bilingual broadcasters, ensuring standard pronunciation and moderate speed; the text-image part adopts side-by-side Chinese-English layout, with key information (e.g., scenic spot names, historical periods) highlighted in bold.
2. **Bilingual electronic map navigation function**: The electronic map uses bilingual labels for scenic spot names, road names, and service facility names (e.g., restrooms, restaurants, parking lots). It supports two positioning methods ("GPS positioning" and "QR code positioning") and a "POI filtering" function, allowing tourists to filter scenic spots by category.
3. **Bilingual travel log sharing function**: Enables tourists to take photos/videos via the system, add Chinese-English text descriptions (with translation assistance), and create a "travel log." The log can be shared on social platforms such as WeChat, Facebook, and Twitter. The system also includes a "popular logs" section to provide references for other tourists.
4. **QR code application function**: Deploys bilingual QR code signs next to each scenic spot and service facility. Scanning the QR code allows tourists to access bilingual interpretation information, locate themselves on the electronic map, and complete ticket verification or interpretation service booking.
5. **Bilingual information query function**: Provides comprehensive Chinese-English information query services (e.g., opening hours, ticket prices) and an "intelligent Q&A" function—tourists can input questions in Chinese or English, and the system returns answers in both languages.

3.2.3 Application Effects and User Feedback

3.2.3.1 Application Effect Data

1. **User coverage**: By the end of 2023, the system had 850,000 registered users, with international users (using the English interface) accounting for 12%—a significant increase from less than 3% before the system's launch.
2. **Service efficiency**: Before the system's deployment, each human interpreter served an average of 30 tourists per day; this number has decreased to 15. The average waiting time for tourists has been reduced from 40 minutes to 15 minutes, and the number of questions asked to staff has decreased by 45%.
3. **Stay time**: The average tourist visit duration has increased from 3.5 hours to 5 hours.

3.2.3.2 User Feedback

A 2023 satisfaction survey collected 2,000 valid questionnaires, with the following results:

- **Overall satisfaction**: 88% of users were satisfied (90% of Chinese users, 82% of English users).
- **Top-recognized functions**: Bilingual electronic map navigation (92% recognition rate) and bilingual interpretation (89% recognition rate) were the most popular functions.
- **User suggestions**:
  a. Add more languages (e.g., Japanese, Korean, French);
  b. Improve speech recognition accuracy;
  c. Enrich cultural details in interpretation content.

3.2.4 Insights from the Case

1. **Adapt to scenario characteristics in technology architecture**: Select a suitable system architecture based on the scenic area's features (e.g., terrain, tourist volume) to ensure stable operation.
2. **User-centric function design**: Focus on core user needs (e.g., navigation, interpretation) and ensure full bilingual support for key functions.
3. **Continuous experience optimization**: Update the system based on user feedback and integrate emerging technologies (e.g., AI speech, AR).
4. **Integrate technology and human services**: Combine intelligent systems with human services to meet complex needs (e.g., in-depth cultural interpretation).

3.3 Comparative Analysis of Museum and Scenic Area Bilingual Interpretation

3.3.1 Difference Analysis

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

Comparison Dimension	Museum Scenario (Nanyue King Museum)	Scenic Area Scenario (Scenic Area M)
Service Objective	Cultural communication; enhance international influence	Tourist experience; improve service efficiency
Content Design Focus	Depth/accuracy of cultural information; cross-cultural comparison	Practicality/convenience of information; clear navigation
Technology Application Focus	Intelligent technology as a supplement to human interpretation	Intelligent technology as core support for automation
User Needs	In-depth cultural information demands (mainly international tourists)	Real-time navigation/query demands (domestic + international tourists)

3.3.2 Common Experiences

1. **Take linguistic accuracy as the foundation**: Ensure translation correctness to avoid misunderstandings.
2. **Take cultural adaptation as the key**: Add explanations of cultural differences to improve communication effectiveness.
3. **Take technology support as the trend**: Use intelligent technology to break the limitations of time and space.
4. **Take user orientation as the principle**: Optimize services based on user preferences and needs.

IV. CHALLENGES AND SOLUTIONS OF BILINGUAL INTERPRETATION

4.1 Challenges

4.1.1 Language and Cultural Challenges

1. **Inaccuracy and lack of professionalism**: Frequent translation errors (e.g., incorrect translation of "汉代玉衣 (Hàn Dynasty Jade Shroud)" as "Han Dynasty jade clothes") and inconsistent professional terminology (e.g., "青铜器 (bronze artifacts)" translated as both "bronze ware" and "bronze artifact").
2. **Insufficient cultural adaptation**: Missing cultural background information (e.g., no explanation of the symbolic meaning of "dragon") and inconsistent expression styles of cultural content (alternating between implicit and direct).

4.1.2 Technological and Talent Challenges

1. **Technological limitations**: Limited language coverage, underutilized functions (e.g., only single-mode voice broadcasting), poor compatibility with other tools (e.g., inability to support foreign-language documents), and high costs.
2. **Talent shortage**: Weak foreign language proficiency of staff and insufficient supply of interpreters proficient in minority languages.
3. **Service inefficiencies**: Complicated tax refund procedures and mismatched timelines between pre-authorization and fund unfreezing.

4.2 Solutions

4.2.1 Technology Empowerment

1. **Application of AI large models**: Use domestic large language models to provide multi-language real-time Q&A and emotional narration.
2. **Tiered interpretation**: Customize content for different audiences (e.g., the British Museum classifies explanations by CEFR levels).
3. **Low-cost solutions**: Adopt offline intelligent translation devices (e.g., those used in Zhangjiajie) and QR code-based interpretation. For example, Ruike Translation reduced costs by 30% through QR code technology.

4.2.2 Cultural Adaptation and Standardization

1. **Terminology standardization**: Formulate *Standards for Foreign Language Translation in Public Services* to unify translations.
2. **Cultural analogy**: Use familiar cultural references to explain unfamiliar concepts (e.g., compare "Zhuge Liang's wisdom" to "Greek Odysseus" when communicating with Western tourists).
3. **Native-language review**: Invite native speakers of the target language to polish content (e.g., labeling taboos on French menus).

4.2.3 Talent Development and Ecosystem Optimization

1. **Talent echelon construction**: Train staff in scenario-specific foreign language application and

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

launch "young interpreter programs" (e.g., the Wuhou Shrine's program covering 8 languages).

2. **Industry standard promotion**: Promote the adoption of ISO 21707 *International Standard for Bilingual Heritage Interpretation* and build scenic area-specific language corpora.
3. **Ecosystem integration**: Connect to international booking platforms (e.g., Agoda, Klook) and deploy intelligent tax refund terminals supporting direct RMB refunds.

V. CONCLUSION

High-quality bilingual interpretation is a critical tool for cross-cultural communication in the era of globalization. Its development relies on technological innovation, cultural adaptation, talent support, and standardization. By addressing current problems (e.g., uneven translation quality, technological limitations, talent shortages) with targeted solutions, bilingual interpretation can better enhance China's cultural soft power, promote fair cultural communication, and drive the high-quality development of the cultural and tourism industry.

REFERENCES

[1] Ling, O. Y. (2023). An international exploration of museums telling Chinese stories well: A case study of the foreign language program at Nanyue King Museum. *Oriental Collection*, (10), 106-108.

[2] Zhu, L. L. (2022). Design research on intelligent scenic area interpretation systems under the background of smart tourism. *Tourism Overview*, (03), 83-85.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.2

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

The Crossroads of Neuroscience and Project-Based Learning: A New Era in Biology Teaching

Ulya Shirinzade*, Miuccia Li

*PHD Dissertator, Institute of Education, Azerbaijan

Received: 16 Aug 2025, Received in revised form: 18 Sep 2025, Accepted: 22 Sep 2025, Available online: 25 Sep 2025

Abstract

This article explores the integration of neuroscience concepts into biology education through project-based learning (PBL), aimed at enhancing student engagement and comprehension of complex biological processes. The methodology actively engages students through hands-on projects and interdisciplinary approaches, fostering a lively classroom atmosphere characterized by curiosity and creativity. The research highlights the positive effects of PBL on student retention and perceptions of learning, supported by findings indicating increased teacher efficacy and student neuroscience literacy. Collaborations with neuroscience professionals enhance real-world relevance, enriching educational experiences. Ultimately, this approach transforms traditional teaching methods and empowers students to become informed and critical thinkers, ready to tackle real-world challenges.

***Keywords*— Neuroscience, Biology Education, Project-Based Learning, Student Engagement, Interdisciplinary Learning, Teacher Efficacy, Neuroscience Literacy, Active Learning, Educational Methods.**

I. INTRODUCTION

Neuroscience has emerged as a critical discipline, revealing the underlying mechanisms of learning and cognition that can significantly influence educational approaches. As education shifts toward more engaging and effective practices, understanding how to leverage neuroscience in the classroom is vital. This article investigates how integrating neuroscience concepts into biology education through project-based learning (PBL) can significantly enhance student understanding, retention, and enthusiasm for science. By focusing on the active involvement of students in their learning processes, we aim to explore the potential benefits of this innovative pedagogical approach.

The existing body of literature highlights the effectiveness of student-centered learning approaches in engaging learners. Barron and Darling-Hammond (2008) argue that project-based learning fosters deeper engagement, critical thinking, and problem-solving skills by allowing students to explore real-world issues. Within the context of biology education, integrating neuroscience can create intersections that enhance understanding—for instance, linking neural mechanisms to concepts in evolution, genetics, and behavior. Research by Duncan et al. (2020) emphasizes that interdisciplinary methods can facilitate deeper connections in students' learning, thereby enriching their educational experience and improving knowledge retention.

Integrating neuroscience concepts into biology lessons is not merely a trend; it is a transformative method that reshapes how students perceive and engage with science. Traditional teaching often relegates students to the role of passive listeners, merely observing as educators convey information. In contrast, project-based learning actively involves students in their education, transforming them into researchers, creators, and problem solvers. They explore complex biological principles while

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.3

simultaneously linking them to neurobiological concepts.

Recent findings indicate that "at the beginning of the intervention, 68% of teachers reported holding a growth mindset. After the intervention, that number rose to 83%, showing a 15 percentage point increase." Furthermore, "at the start of the intervention, 49.5% of teachers reported above-average teaching efficacy. After the intervention, that number increased to 66%, an increase of 17 percentage points." These results suggest that "a brief intervention (six sessions totaling 270 minutes) that teaches basic neuroscience concepts, such as neuroplasticity, can influence elementary teachers' beliefs" (Abeer, 2025).

The effectiveness of this educational strategy is rooted in the very neuroscience principles it seeks to teach. Studies have shown that active learning methodologies significantly improve retention rates (Sweller et al., 1998). When students engage in inquiry-based activities, their brains form and strengthen neural connections, thereby enhancing their understanding and memory of biological processes.

This dynamic teaching methodology fosters an environment where students are not merely recipients of knowledge; they are explorers utilizing their hands and minds to make discoveries. When students engage directly with the subject matter through tangible projects, they develop a deeper understanding of the intricacies of life sciences.

Studies also suggest that when students engage with neuroscience concepts, they cultivate a higher degree of curiosity and a deeper understanding of themselves as learners. This is essential for fostering lifelong learning habits and adaptability in an ever-evolving academic and professional landscape (Bradberry & Greaves, 2009).

As we advance into the 21st century, integrating neuroscience concepts into biology education through project-based learning signifies the beginning of a new era in teaching. Despite advancements in educational neuroscience, misconceptions about the brain and learning—known as neuromyths—persist among educators, hindering the application of scientific findings in educational settings (Kalyuga, 2007).

Changing entrenched beliefs requires more than presenting factual information; it necessitates approaches that consider educators' existing knowledge and experiences, alongside the intuitive appeal of neuromyths (Pradep et al., 2024). Recent interventions targeting in-service teachers, such as refutation texts and immersive experiences, have demonstrated promise in reducing belief in neuromyths by directly addressing misconceptions and providing accurate neuroscience information.

By equipping students with the tools to understand complex biological systems within real-world contexts, educators empower them to become informed citizens and critical thinkers. This cross-disciplinary approach cultivates a generation eager to explore the mysteries of life and the human brain, armed with the knowledge and skills necessary to navigate the challenges of the modern world. The fusion of neuroscience and biology in the classroom is not merely a pedagogical choice; it is a potent strategy that resonates with our innate curiosity, promising a future filled with discovery and understanding.

II. METHODS

Students in the experimental group engaged in a variety of interdisciplinary projects, including:

1. Model Creation: Building scale models of the human brain where students had to include details about neural pathways and the functions of different neurotransmitters, encouraging them to understand biological structures through a neuroscience lens.

2. Music and Memory Experiment: Designing and conducting experiments to examine how different genres of music affected memory retention. Through this project, students not only applied knowledge of neuroplasticity and synaptic function but also explored concepts related to auditory processing and memory retrieval.

3. Interactive Demonstrations: Creating presentations to illustrate complex neurological processes, such as reflex arcs and neuroplasticity, allowing students to inform their peers and encourage discussions around scientific inquiry.

Before and after the intervention, students completed surveys measuring their levels of engagement, comprehension of biological concepts, and perceptions of the learning process. Additionally, teacher interviews were conducted to assess perceived changes in instructional efficacy and confidence in teaching topics related to neuroscience.

III. RESULTS

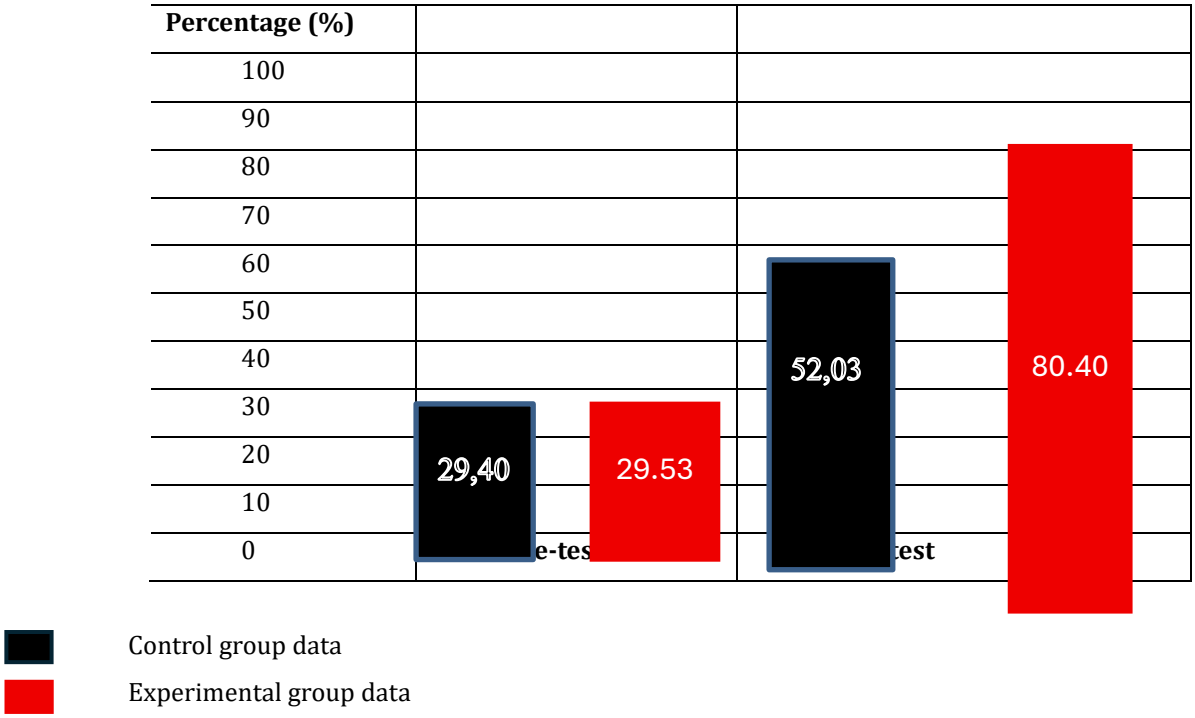
To evaluate the impact of incorporating neuroscience into biology education, a quasi-

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.3

experimental design was employed. A total of 60 high school biology of Azerbaijani students were divided into two groups: an experimental group that engaged in project-based learning with neuroscience integration and a control group that received traditional biology instruction. The intervention lasted six weeks and comprised a series of hands-on projects designed to merge biological concepts with neuroscience principles.

The study took place in a suburban high school located in a diverse district, with 60 students aged 15 to 18 years participating. The cohort was diverse regarding socio-economic status, academic achievement, and prior exposure to neuroscience, ensuring a representative sample for the study.

Percentage of success of each group



Data provided compelling evidence of the effectiveness of integrating neuroscience into high school biology through project-based learning:

- Engagement Scores: The experimental group reported higher levels of engagement, with average scores increasing from 3.5 to 4.6 on a 5-point Likert scale ($p < 0.01$).

- Understanding Biological Concepts: Students in the experimental group outperformed the control group on assessments measuring comprehension of complex biological processes, with an improvement of 25% compared to a mere 5% in the control group.

- Teacher Efficacy: Analysis of teacher interviews revealed that pre-intervention, only 49.5% of teachers felt confident in their ability to teach complex biological topics effectively. Post-intervention, this number rose to 66%, indicating significant growth in teacher confidence and perceived efficacy.

IV. DISCUSSION

The findings from this study underscore the transformative potential of integrating neuroscience concepts into high school biology through project-based learning. Engaging students in hands-on, exploratory projects allows them to become active learners, emphasizing exploration and discovery rather than passive reception of information.

The positive shifts observed in both student engagement and teacher efficacy suggest that PBL methodologies not only improve learning outcomes but also empower educators to adopt more dynamic instructional strategies. By fostering curiosity and creativity, this integrated approach encourages students to see science as interconnected, highlighting the relevance of biological concepts in understanding real-world phenomena.

Reflective of the educational landscape, changes in pedagogical methods may be necessary to prepare

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.3

students for future challenges. The incorporation of neuroscience principles as part of biology instruction empowers students to understand their own learning processes, enhancing their neuroscience literacy and equipping them with tools for lifelong learning.

While this study provides valuable insights into the benefits of integrating neuroscience into biology education, several limitations must be acknowledged. The study's sample is limited to one suburban high school, which may restrict the generalizability of the findings. Future research should aim to expand these inquiries across various educational settings, including urban and rural districts, to assess broader applicability.

Longitudinal studies exploring the long-term impact of such interventions on student pathways, particularly in the sciences, are also warranted. Further investigation into specific pedagogical approaches within PBL and their relationships to student outcomes would enhance the knowledge base for practitioners seeking to employ interdisciplinary strategies.

V. CONCLUSION

This study evidences the powerful effects of integrating neuroscience concepts into biology education through project-based learning. By promoting an engaging and interactive learning environment, this approach enhances student comprehension of biological processes and empowers them to become informed, critical thinkers. As we navigate an increasingly complex world, the merger of neuroscience and biology through innovative educational practices offers a promising path forward for enriching science education and preparing students for future challenges.

REFERENCES

[1] Abeer, F. A. (2025). Neuroscience literacy and academic outcomes: Insights from a cross-sectional study.

[2] Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends in Cognitive Sciences, 12 (5), 193-200.

[3] D'Ardenne, K., Eshel, N., Luka, J., Lenartowicz, A., Nystrom, L. E., & Cohen, J. D. (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. Proceedings of the National Academy of Sciences of the United States of America, 109(49), 1990-1999.

[4] Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. Science, 329 (5998), 1541-1543.

[5] Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychologist Review, 19(4), 509-539.

[6] Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38(1), 23-31.

[7] Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. Brain Research, 1243, 86-103.

[8] Novak, J. D. (1990). Concept maps and Vee diagrams: Two metacognitive tools to facilitate meaningful learning. Instructional Science, 19, 29-52.

[9] Pradep, K., Rajalakshmi, S. A., Priya, T. A., Aswathy, S., Jisha, V. G., & Vaisakhi, V. S. (2024). Neuroeducation: Understanding neural dynamics in learning and teaching. Frontiers in Education.

[10] Sweller, J., van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. Educational Psychologist Review, 10 (3), 251-296.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.3

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Experiential Learning in Engineering Education: A Bibliometric Analysis of Perceptions and Transformations

Bhajan Lal[1], Muhamad Nawaz[1], Haslinda Zabiri[1], Chong Su Li[2], Prashant Kumar[3]

[1]Chemical Engineering Department, Universiti Teknologi PETRONAS, Malaysia
Email: bhajan.lal@utp.edu.my
[2]Centre for Excellence in Teaching and Learning (CETaL), Universiti Teknologi PETRONAS, Malaysia
Email: chong_suli@utp.edu.my
[2]Chemical Engineering and Physical Sciences, Lovely Professional University, Phagwara, Punjab-India
Email: prashant.25338@lpu.co.in

Received: 19 Aug 2025, Received in revised form: 20 Sep 2025, Accepted: 24 Sep 2025, Available online: 27 Sep 2025

Abstract

Experiential learning has emerged as a central approach in engineering education, fostering student engagement, problem-solving, and professional skill development. To map this evolving field, a bibliometric analysis was conducted using data retrieved from Scopus on 1 September 2025, covering 694 publications from 2010 to 2024. Analyses with VOSviewer examined publication trends, keyword co-occurrence, and collaboration patterns. Results show steady growth, with a sharp increase after 2020, driven by digital transformation and the COVID-19 pandemic. Four thematic clusters were identified: (1) project- and problem-based learning, (2) student engagement and perceptions, (3) technology-enhanced learning, and (4) professional development and sustainability. The United States dominates output and citations, while Denmark and Portugal lead through influential scholars and institutions such as Aalborg University and the University of Minho. Collaboration remains fragmented, with limited cross-regional links. The study highlights strengths, gaps, and opportunities, offering guidance for educators, policymakers, and researchers to advance experiential learning in engineering education.

***Keywords*— Experiential Learning, bibliometric analysis, Engineering Education.**

I. INTRODUCTION

In the evolving landscape of engineering education, traditional lecture-based methods are increasingly being complemented—or even supplanted—by experiential learning strategies [1]. Engineering, by its nature, is an applied discipline, and students often benefit from pedagogical approaches that align with hands-on, real-world experiences [2]. The integration of experiential learning—defined broadly as the process of learning through direct experience, reflection, and application—has become a crucial focus for educators aiming to bridge the gap between theoretical knowledge and professional practice [3].

The pedagogical foundation for experiential learning in engineering is rooted in the experiential learning theory (ELT) developed by David Kolb [4]. According to Kolb, effective learning occurs when students move through a cyclical process of concrete experience, reflective observation, abstract conceptualization, and active experimentation. This theory has informed the design of project-based learning (PBL), cooperative education, internships, simulations, and other student-centered teaching methods across engineering curricula worldwide [5].

Over the past two decades, there has been growing interest in investigating the effectiveness, student

perceptions, and transformative potential of experiential learning within engineering contexts [6]. This interest has been fueled in part by demands from industry stakeholders who expect graduates to demonstrate not just academic proficiency but also soft skills such as teamwork, problem-solving, adaptability, and communication [7]. Moreover, the emergence of Education 4.0—driven by Industry 4.0 technologies—has further accelerated the need for adaptive, interdisciplinary, and personalized learning experiences that simulate the complexity of real engineering environments [8].

Engineering education is undergoing a profound transformation due to technological, economic, and societal changes [9]. These changes require engineers who are not only technically proficient but also socially aware, globally competent, and capable of lifelong learning [10]. Experiential learning plays a crucial role in addressing these demands by engaging students in active, authentic learning processes [6].

As emphasized by Tembrevilla and Phillion (2024), experiential learning in engineering programs enables students to understand abstract concepts by applying them in real contexts, thereby improving retention and cognitive depth [6]. Furthermore, such pedagogies encourage reflective practices and help students develop metacognitive awareness—key to innovation and leadership in engineering fields [11].

The rise of student design competitions, living labs, and industry-university partnerships has provided unique platforms for experiential learning [12]. These initiatives have allowed learners to work in multidisciplinary teams, tackle open-ended problems, and engage with external stakeholders [13]. However, the breadth of strategies labeled as "experiential" and the diverse educational outcomes they aim to achieve have led to calls for more systematic, evidence-based analysis of their implementation and impact [6].

Despite the widespread adoption of experiential learning in engineering education, there remains a lack of consolidated knowledge regarding its scope, theoretical evolution, effectiveness, and research trends [9]. A bibliometric analysis serves as an essential tool to map out the intellectual structure of the field. It helps identify key research themes, leading scholars, influential publications, and collaboration networks [14].

Bibliometric studies also reveal the evolution of discourse and highlight shifts in emphasis—from initial studies focused on curriculum innovation to recent work on digital transformation, sustainability, and artificial intelligence in experiential settings [9, 10].

This study seeks to fill this gap by offering a comprehensive bibliometric analysis of experiential learning in engineering education, focusing on how it has been perceived, applied, and transformed over time. In doing so, it contributes to ongoing efforts to enhance evidence-based teaching practices and align engineering programs with global workforce needs.

II. METHODOLOGY

The methodological workflow is illustrated in Figure1, comprising two main phases: literature search and data screening.

The bibliometric dataset was retrieved from the Scopus database on 1 September 2025, as Scopus offers broad coverage of engineering and education-related publications. A comprehensive search strategy was employed using Boolean operators to combine experiential learning terms ("experiential learning" OR "hands-on learning" OR "practical learning" OR "project-based learning" OR "work-integrated learning" OR "active learning") AND (“engineering education" OR "engineering teaching" OR "engineering pedagogy"). To capture studies focusing on perceptions and transformations, additional keywords were included (perception OR attitude OR transformation OR reform OR effectiveness OR outcomes). This initial search yielded a total of 3,497 documents.

A systematic screening procedure was then applied to refine the dataset (Figure 1). First, a time filter restricted the period to 2010–2024, as research on experiential learning in engineering gained significant momentum during this timeframe. Second, the dataset was limited to peer-reviewed journal articles to ensure scholarly quality and rigor. Third, only documents published in the English language were retained to ensure consistency in analysis. Finally, documents unrelated to the context of experiential learning in engineering education were excluded after a manual relevance check. Following this screening process, the final dataset comprised 694 documents, which served as the basis for descriptive statistics, keyword co-occurrence analysis, and collaboration network mapping.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.4

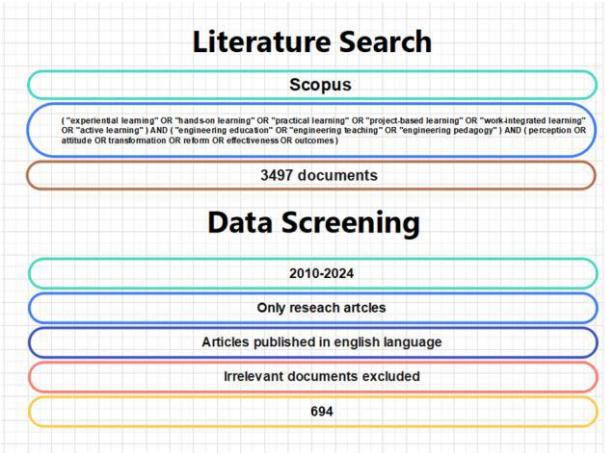


Fig. 1: Literature search

III. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

As shown in Figure 2, research output on experiential learning in engineering education has experienced steady and accelerating growth over the past 15 years. The earliest years of the analysis (2010–2013) reflect a nascent stage, with fewer than 25 publications per year, indicating limited but emerging interest in integrating experiential approaches such as project-based learning and problem-based learning into engineering curricula. A gradual increase is observed from 2014 to 2017, when annual publications consistently surpassed 40 papers, reflecting a growing recognition of active learning methods within engineering education research.

The period from 2018 onward marks a rapid expansion phase, with annual publications rising above 60 and continuing to grow each year. Notably, the number of publications surged significantly after 2020, coinciding with the global COVID-19 pandemic, which accelerated the adoption of blended and online experiential learning practices. By 2024, annual publications reached 114 papers, representing the highest output within the study period.

The cumulative trend line highlights the exponential nature of growth, with the total number of publications increasing from fewer than 100 in 2013 to nearly 700 documents by 2024. This trajectory suggests that experiential learning has transitioned from a marginal pedagogical innovation to a mainstream research focus in engineering education. The sharp upward trend after 2020 also indicates a sustained scholarly commitment to exploring not only traditional models such as project- and problem-based learning but also technology-enhanced approaches including virtual reality, augmented reality, and artificial intelligence.

Overall, the publication trend demonstrates that experiential learning has become a critical and fast-growing research domain within engineering education, reflecting its importance in addressing the evolving demands of industry, sustainability, and digital transformation.

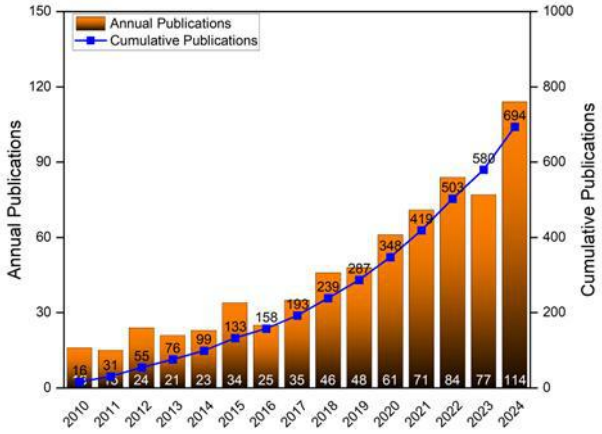


Fig. 2: Publication Tread

The analysis of publication sources reveals that experiential learning in engineering education is a truly multidisciplinary field, with contributions spread across 197 different journals. This breadth reflects the diversity of perspectives — spanning engineering education, pedagogy, technology-enhanced learning, and sustainability. However, a relatively small set of journals contributes disproportionately to the knowledge base, as highlighted in Table 1.

Table 1: Top ten sources

Rank	Journal	TP	TC	CPP	H	YA
1	International Journal of Engineering Education	96	1060	11.04	17	2010-24
2	IEEE Transactions on Education	69	1331	19.29	22	2010-24
3	Journal of Engineering Education Transformations	59	171	2.90	6	2018-24
4	European Journal of Engineering Education	54	2080	38.52	26	2011-24
5	Computer Applications in Engineering Education	23	232	10.09	9	2015-24
6	Advances in Engineering Education	17	290	17.06	8	2010-24
7	Sustainability	17	246	14.47	9	2020-24
8	Education for Chemical Engineers	16	377	23.56	10	2016-24
9	Journal of Engineering Education	16	674	42.13	14	2010-24
10	Education Sciences	14	249	17.79	8	2019-24

The International Journal of Engineering Education (IJEE) leads with 96 publications between 2010 and 2024, confirming its role as a flagship outlet for pedagogical innovation in engineering. The IEEE Transactions on Education ranks second with 69 publications, but it surpasses IJEE in total citations (1,331) and citations per paper (CPP = 19.29), underscoring its high impact within the engineering education research community.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.4

Overall, while research is distributed across nearly 200 journals, these top 10 sources account for a significant share of output and citations, serving as the primary channels for scholarly discourse. This concentration underscores that although the field is diverse, it also maintains a set of core journals where experiential learning research achieves visibility and impact.

3.2 Keyword Co-occurrence Clusters

The keyword co-occurrence network in Figure 3 illustrates the intellectual structure of experiential learning in engineering education. From the total of 4,298 unique keywords, only those appearing 15 times or more were included in the visualization, ensuring that the map reflects the most prominent and influential research themes. The resulting clusters highlight the thematic diversity and interconnections within the field.

The largest node, "engineering education", appears at the center of the network, signifying its role as the dominant anchor term. Surrounding this core, several thematic clusters can be identified:

Cluster 1 – Project- and Problem-based Learning (Red cluster)

Keywords: project-based learning, problem-based learning, curriculum, educational computing, software engineering.

This cluster emphasizes structured pedagogical models that encourage hands-on, inquiry-driven learning, strongly linked to curriculum reform and computing/engineering design contexts.

Cluster 2 – Active Learning and Student Engagement (Blue cluster)

Keywords: active learning, flipped classroom, student learning, student engagement, student perceptions, motivation.

This theme centers on approaches that enhance classroom interaction and learner-centered practices, reflecting growing attention to student experience and perception studies.

Cluster 3 – Experiential and Applied Contexts (Green cluster)

Keywords: experiential learning, higher education, sustainable development, innovation, teamwork, laboratories, design.

This cluster highlights the application of experiential approaches in broader contexts, including sustainability and innovation, aligning engineering education with global challenges.

Cluster 4 – Technology-Enhanced Learning (Yellow cluster)

Keywords: e-learning, online learning, virtual reality, augmented reality, computer-aided instruction, learning environments.

The prominence of these terms reflects the digital transformation of experiential learning, accelerated by the COVID-19 pandemic, and the rise of virtual platforms for immersive and remote education.

The strong interconnections among clusters demonstrate that research in experiential learning is highly interdisciplinary and overlapping, rather than siloed. For example, project-based learning links closely to active learning and student engagement, while sustainability connects both to curriculum innovation and technology-enhanced environments.

The frequency threshold of 15 ensures that only the most consistent themes are highlighted, filtering out sporadic or niche topics. This approach reveals that experiential learning research has matured into four dominant and interconnected research streams: pedagogical frameworks, learner engagement, applied/sustainable contexts, and digital/technology integration.

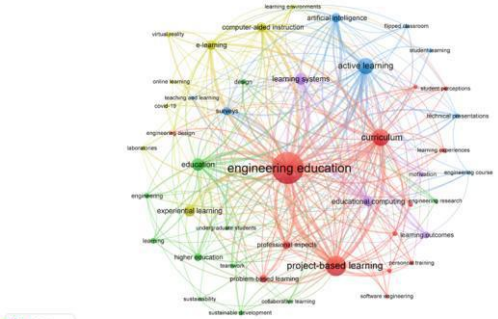


Fig. 3: keywords occurrences

3.3: Prominent Authors, Institutions, and Countries

Out of 2,127 contributing authors (Table 2), only a small group show consistent productivity and high impact. Kolmos, Anette Jensen (Aalborg University, Denmark) leads with 9 publications and the highest influence (652 citations, CPP = 72.44), reflecting her pioneering role in problem- and project-based learning (PBL). Lima, Rui M. (University of Minho, Portugal) and Mesquita, Diana (Catholic University of Portugal) both contribute 8 publications, with CPP

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.4

values of 39.75 and 34.13, respectively, focusing on curriculum and project-based learning. Du, Xiangyun (Aalborg University) and Fernandes, Sandra Raquel Gonçalves (Instituto Politécnico do Porto, Portugal) also rank highly, each with strong citation impacts (CPP > 50). Bhajan et al, (Figure 4) greatly enhancing the learning by doing experience of physical students related to chemical kinetics and phase behaviour through field trips, case studies, adjunct lecture as well as gas hydrate lab visit in Chemical Engineering Department at Universiti Teknologi PETRONAS Malaysia.

Fig. 4: Gas Hydrate lab visits by Physical Chemistry students (Universiti Teknologi PETRONAS)

Overall, the leading scholars are concentrated in Denmark and Portugal, highlighting these regions as hubs of experiential learning research in engineering education. Despite broad participation, intellectual leadership is anchored by this relatively small group of highly cited authors.

Table: 2 Top 5 authors

Rank	author	Institution/Country	TP	TC	CPP	H	YA
1	kolmos, anette jepsen	Aalborg University, Denmark	9	652	72.44	6	2011-23
2	lima, rui m.	University of Minho, Portugal	8	318	39.75	6	2012-24
3	mesquita, diana	Catholic University of Portugal, Portugal	8	273	34.13	6	2011-24
4	du, xiangyun	Aalborg University, Denmark	7	387	55.29	6	2021-23
5	fernandes, sandra raquel gonçalves	Portucalense Infante D. Henrique University, Portugal	6	337	56.17	5	2012-23

Among the 841 institutions contributing to experiential learning in engineering education, only a few demonstrate sustained productivity and influence (Table 3). Aalborg University (Denmark) ranks first with 14 publications and the highest impact (742 citations, CPP = 53.00), reflecting its global reputation as a pioneer of problem- and project-based learning (PBL). Universidade do Minho (Portugal) also produced 14 publications with strong citation performance (504 citations, CPP = 36.00), underscoring Portugal's growing role in this research domain.

Beyond Europe, the Tecnológico de Monterrey (Mexico) contributed 14 publications but with lower citation impact (CPP = 16.93), highlighting emerging leadership from Latin America. Universidad Politécnica de Madrid (Spain) and the University of Michigan, Ann Arbor (USA) complete the top five, both with 9 publications, representing significant contributions from Southern Europe and North America.

Overall, while contributions are widely distributed across hundreds of institutions, intellectual leadership is concentrated in a small set of European universities, complemented by emerging activity in Latin America and the United States. This pattern indicates both global diffusion and regional hubs driving experiential learning research in engineering education.

Table: 3 Top 5 institutions

Rank	*Institution*	TP	TC	CPP	H	YA
1	aalborg university, aalborg, denmark	14	742	53.00	9	2011-23
2	universidade do minho, braga, portugal	14	504	36.00	9	2011-24
3	tecnológico de monterrey, monterrey, mexico	14	237	16.93	8	2011-24
4	universidad politécnica de madrid, madrid, spain	12	247	20.58	8	2015-24
5	university of michigan, ann arbor, ann arbor, united states	9	189	21.00	6	2013-24

The dataset includes contributions from 83 countries, reflecting the global spread of experiential learning research in engineering education. However, output and influence are concentrated in a few leading nations (Table 4).

The United States dominates with 207 publications and 4,148 citations, the highest h-index (32) and CPP (20.04) among the top producers. This confirms the USA's role as the global hub of engineering education research, supported by long-standing traditions of active learning and strong institutional networks.

India ranks second in productivity with 79 publications, but with lower citation impact (CPP = 18, h = 10), suggesting that while research activity is expanding rapidly, international visibility and influence are still developing.

Spain contributes 68 publications with strong scholarly impact (1,153 citations, CPP = 27, h = 19), reflecting its established focus on project-based and student-centered pedagogies. Australia follows with fewer outputs (33 publications) but demonstrates remarkable influence (1,079 citations, CPP = 33, h = 16), indicating highly cited contributions despite

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.4

lower volume. Similarly, the United Kingdom has 32 publications but strong citation performance (648 citations, CPP = 32, h = 14), underscoring its reputation for high-quality, impactful studies in engineering pedagogy.

Overall, while 83 countries contribute to the field, intellectual leadership remains concentrated in a handful of advanced economies, particularly the USA and parts of Europe. Emerging economies such as India are expanding output, but citation impact lags behind, highlighting opportunities for stronger international collaboration and greater global integration.

Table: 4 Top 5 Countries

Rank	country	TP	TC	CPP	H	YA
1	United States	207	4148	56	32	2010-24
2	India	79	406	18	10	2014-24
3	Spain	68	1153	27	19	2011-24
4	Australia	33	1079	7	16	2011-24
5	United Kingdom	32	648	32	14	2010-24

3.4 Collaboration Networks

The co-authorship network (Figure 5) highlights the structure of collaborations among the 2,127 authors contributing to experiential learning research in engineering education. Despite the large pool of contributors, collaboration is concentrated in a few prominent clusters, with many authors publishing independently or in small groups.

The largest and most influential cluster is centered on Kolmos, Anette Jensen (Aalborg University, Denmark), whose extensive work on problem- and project-based learning has established Aalborg as a global hub. Kolmos collaborates closely with colleagues such as Du, Xiangyun (also Aalborg University), forming a strong Scandinavian-led network that is well integrated with other European and Asian researchers.

Another visible hub is formed around Fernandes, Sandra Raquel Gonçalves and colleagues from Portugal and Spain, reflecting the growing prominence of Iberian institutions in project- and curriculum-based approaches. Smaller but emerging clusters are observed in Asia and the Middle East, with authors such as Khalid, Md. Safiuddin and Chandran, M. linking engineering pedagogy with context-specific innovations in developing regions.

Overall, while the network demonstrates the presence of several well-connected leaders, the collaboration landscape is fragmented, with a large number of isolated authors and small clusters. This indicates that experiential learning research is still maturing as a global collaborative field. Strengthening cross-continental partnerships, particularly linking emerging research regions (e.g., India, Latin America) with established hubs in Europe and North America, could enhance knowledge exchange and raise the global impact of this domain.

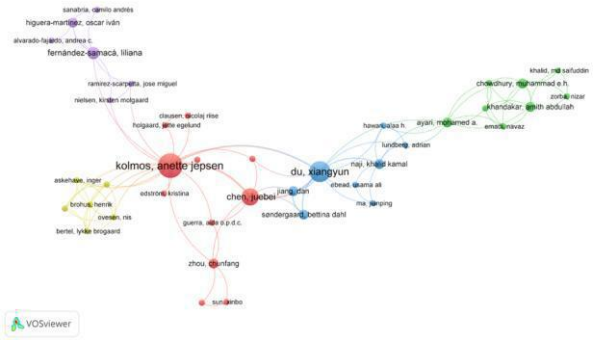


Fig. 5: Collaboration of Authors

The country co-authorship network (Figure 6) maps collaboration patterns among the 83 contributing nations, with only countries producing five or more publications included for clarity. The network is dominated by a few highly productive hubs, particularly the United States, which occupies the central position with extensive collaborative links to Europe, Asia, and Australia. This reflects its role as the leading global contributor in terms of both volume and citation impact.

Spain, the United Kingdom, and Germany form strong European nodes, frequently partnering with the United States as well as with regional neighbors such as Portugal and the Netherlands. This cluster illustrates the strength of intra-European collaboration, which has been instrumental in advancing project- and problem-based learning approaches.

In Asia, India emerges as a productive hub, collaborating actively with both Western countries and regional partners including Malaysia, China, and Singapore. While India contributes substantial output, its collaboration patterns indicate growing but still developing international integration. Australia appears as another active node, linking the Asia-Pacific region with Europe and North America.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.4

Smaller but significant contributors include United Arab Emirates, Saudi Arabia, and Qatar, reflecting rising research interest from the Middle East in engineering pedagogy and experiential learning. These countries often collaborate with Western partners, showing an outward-looking orientation.

Overall, while the network highlights several well-connected clusters, the distribution also shows asymmetry: a few leading nations (United States, Spain, India, UK, Australia) anchor the field, while many of the 83 participating countries remain peripheral with limited international partnerships. Strengthening South–South collaborations (e.g., between Asia, Latin America, and Africa) could enhance diversity and global representation in experiential learning research.

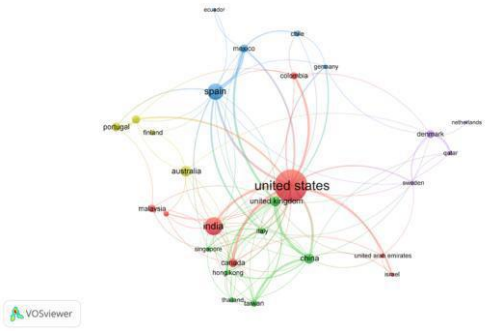


Fig. 6: collaboration of countries

IV. LIMITATIONS AND FUTURE DIRECTIONS

4.1 Limitations

This study has several limitations. First, the analysis relied solely on the Scopus database, which, while comprehensive, has inherent issues such as author name disambiguation (e.g., variations in spelling or formatting leading to duplicate or fragmented author records) and citation counts that differ from other databases like Web of Science or Google Scholar. Second, the bibliometric mapping is sensitive to keyword variations, including synonyms, spelling inconsistencies, and hyphenation (e.g., problem-based learning vs. problem-based learning), which may result in fragmented clusters or overlooked terms. Finally, restricting the dataset to English-language publications and the period 2010–2024 improves focus but inevitably excludes some earlier and non-English contributions.

4.2 Future Directions

Future research should aim to overcome these constraints by combining data from multiple bibliographic databases (e.g., Scopus, Web of Science, Google Scholar, ERIC) to cross-validate publication and citation metrics. Enhanced author profiling and disambiguation tools should be applied to reduce duplication or misattribution of scholarly outputs. Similarly, future studies should employ more robust keyword standardization, possibly through natural language processing (NLP) techniques, to minimize inconsistencies and better capture emerging terms. Expanding to multilingual datasets and extending the timeframe would further enhance coverage, while linking bibliometric insights with educational policy and curriculum reform practices could strengthen the practical relevance of this research.

V. CONCLUSION

This bibliometric study provides a comprehensive overview of experiential learning research in engineering education, mapping its growth, thematic directions, and global distribution. The analysis of 694 publications from 2010–2024 reveals a steady upward trend, with a marked acceleration after 2020 driven by digital transformation and the impact of the COVID-19 pandemic. Four dominant research clusters were identified: project- and problem-based learning, student engagement and perceptions, technology-enhanced approaches, and professional development with sustainability. Together, these themes demonstrate how experiential learning has evolved from traditional classroom reforms to encompassing digital and societal dimensions.

The findings show that research leadership is geographically concentrated. The United States remains the most productive and influential country overall, while European institutions, particularly Aalborg University in Denmark and the University of Minho in Portugal, anchor intellectual leadership through highly cited contributions. At the same time, emerging contributions from India, Spain, and Latin America highlight growing global interest. Collaboration networks, however, remain fragmented, with limited cross-continental ties and many isolated authors.

This study is subject to limitations, including reliance on a single database (Scopus), author name inconsistencies, variations in citation counts across databases, and challenges in keyword standardization. Future research should integrate multiple databases, expand to multilingual datasets, apply stronger author disambiguation tools, and explore thematic evolution

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.4

through longitudinal and systematic analyses. Strengthening international collaboration and linking bibliometric insights with educational policy and curriculum reform will be essential to advancing experiential learning research and ensuring its impact on future engineering education practices.

ACKNOWLEDGEMENTS

Authors would like to thank UTP's Centre for Excellence in Teaching and Learning (CETaL) for encouragement and assistance.

REFERENCES

[1] Salinas-Navarro DE, Garay-Rondero CL, Arana-Solares IA. Digitally Enabled Experiential Learning Spaces for Engineering Education 4.0. Education Sciences2023.
[2] Gadola M, Chindamo D. Experiential learning in engineering education: The role of student design competitions and a case study. International Journal of Mechanical Engineering Education. 2017;47:3-22.
[3] Steele AL. Experiential learning in engineering education: CRC Press; 2023.
[4] Kolb DA. Experiential learning: Experience as the source of learning and development: FT press; 2014.
[5] Botelho WT, Marietto MdGB, Ferreira JCdM, Pimentel EP. Kolb's experiential learning theory and Belhot's learning cycle guiding the use of computer simulation in engineering education: A pedagogical proposal to shift toward an experiential pedagogy. Computer Applications in Engineering Education. 2016;24:79-88.
[6] Tembrevilla G, Phillion A, Zeadin M. Experiential learning in engineering education: A systematic literature review. Journal of Engineering Education. 2024;113:195-218.
[7] Woodcock CSE, Callewaert J, Millunchick J. Synthesizing definitions of professional competencies linked to experiential learning in engineering education: A literature review. Journal of Higher Education Theory and Practice. 2021;21:123-46.
[8] Salinas-Navarro DE, Garay-Rondero CL, Arana-Solares IA. Digitally enabled experiential learning spaces for engineering education 4.0. Education Sciences. 2023;13:63.
[9] Jamison CSE, Fuher J, Wang A, Huang-Saad A. Experiential learning implementation in undergraduate engineering education: a systematic search and review. European Journal of Engineering Education. 2022;47:1356-79.
[10] Sundman J, Feng X, Shrestha A, Johri A, Varis O, Taka M. Experiential learning for sustainability: a systematic review and research agenda for engineering education. European Journal of Engineering Education. 2025:1-31.
[11] Li H, Öchsner A, Hall W. Application of experiential learning to improve student engagement and experience in a mechanical engineering course. European Journal of Engineering Education. 2019;44:283-93.
[12] Gadola M, Chindamo D. Experiential learning in engineering education: The role of student design competitions and a case study. International Journal of Mechanical Engineering Education. 2019;47:3-22.
[13] O'Brien W, Doré N, Campbell-Templeman K, Lowcay D, Derakhti M. Living labs as an opportunity for experiential learning in building engineering education. Advanced Engineering Informatics. 2021;50:101440.
[14] Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. Journal of business research. 2021;133:285-96.

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Auditing the Fairness of AI-Detection Tools: A Comparative Study of ESL, Published, and AI-Generated Texts and Their Misclassification Risks

R. Paul Lege

Graduate School of Law, Nagoya University, Japan

Received: 11 Sep Aug 2025, Received in revised form: 09 Oct 2025, Accepted: 14 Oct 2025, Available online: 18 Oct 2025

Abstract

This study investigated the classification fairness at the threshold level of four commercially available AI detection tools on the Internet: Copyleaks, ZeroGPT, Scribbr, and Quillbot Premium. The research included the submission of three distinct chunks of texts (N=1212) of between 400-5oo words for evaluation. The writing texts came from fully AI-generated examples (N=307), prompted between 2024 and 2025, and published human-written texts (N=302), and ESL graduate student texts (N=303) written before 2021. The texts were analyzed using binary classification thresholds to determine how the three free devices (Copyleaks, ZeroGPT, Scribbr) and the one paid service (QPremium) performed when checking for potentially AI-generated material in each of the writing examples. The study employed a performance metrics to illustrate the issue with threshold application in such devices. The research included the use of the Chi-square test of independence as well as other inferential statistics to assess inter-detector consistency and potential bias patterns. The results indicated that such devices perform well in identifying AI-generated text written artificially; however, significant disparities emerged in the misclassification of human texts. In particular, AI detectors disproportionally flagged ESL writing with false positives. Such findings illustrate the importance of such fairness audits in assessing the linguistic sensitivity in such tools, especially in the educational setting, where misclassification can have academic or reputational consequences.

***Keywords*—Fairness Audit, AI-generated Text, AI-detectors, ESL pattern bias**

I. INTRODUCTION

With the advent of AI technology, an increasing number of educational institutes report problems associated with students submitting assignments created or written by artificial intelligence. In turn, this threat to academic integrity has compelled educational institutions and teachers in general to depend on AI-detectors to counter the problem. In a survey of articles about this problem in the USA and UK, Anara (2005) found that over 70 percent of schools at all levels may be turning to the use of such devices out of a desperate attempt to halt student cheating [1]. In Europe and Asia, the concern and tendency for educators to fight the unethical use of AI engines with AI detectors has been pretty much the same [2], [3]. The *Ashai Simbun* reported in 2024 that while many Japanese educators understood the limitations of such detectors, they saw them as the last defense against the increasing problem of students misusing AI for assignments [4]. However, while some institutions have ambiguous polices that allow learners to use ChatGPT engines, they have little to say about the use of detection tools to curtail potential misuse of AI technology. As a result, teachers lack the proper training or comprehension of how such tools work to actually employ them properly [5].

Many studies continue to show troubling trends. One problem concerns the misunderstanding that many educators have regarding the nature of such machines;

that is, these devices are probabilistic and not actual indicators of a learner's possible guilt. Institutions can correct this problem through proper training and policy development [6]. The second problem revolves around how such detectors function, which is the concern of this study. Growing research indicates that such tools are programmed algorithmically in such a way that they produce far too many false positives or false negatives to be used to judge learner outcomes [7], [8], [9]. Furthermore, several studies now indicate that such technology may be unintentionally biased toward ESL writing [10], [11], [12].

As of 2025, there appear to be at least 50 commercially available detection tools on the market that vary according to the type of detection (text, images, video, and multimedia) and in terms of detection methodology (linguistic heuristics), audience (education, publication), as well as transparency and reliability. While there has been some small regulatory pressures for change and improvements in the accuracy of such devices, overall, only a few of these companies have published validation studies, and even fewer offer transparent evidence that have addressed the concerns of ESL bias [13] A few companies in the industry have attempted to respond to such concerns [14], [15], but only superficially and without independent verification.

Since the pedagogical risks are high, the public at large, and educators specifically, must continuously view such corporate internal evaluations with healthy skepticism. The evolution of AI technology, combined with the multiple ways to assess such profit-making tools, will drive a need for further research. As the industry offers many detectors that include a host of manipulative features that can change over time, this will necessitate independent corroborative research. Consumers, for example, should find it noteworthy when a new detection service claims that other competitive devices on the market produce false positives while their service does not [16]. Educators, in particular, must be concerned with how accurate and fair such tools are in assessing whether students generate assignments with AI technology. Thus, there remains a continuous need for accuracy and fairness studies concerning such detection tools. This paper aims to conduct a fairness audit of four available detectors on the market that may misclassify ESL text as AI-generated when it was not.

1.1 About Fairness

As a matter of fairness, this study is primarily concerned with identifying *who* is impacted when educators employ AI tools to evaluate student assignments. In general, fairness refers to the equitable treatment of learners regardless of their linguistic background, proficiency level, or writing style [17]. In this context, fairness is a multi-dimensional concept associated with proper statistical analysis, structural transparency, contextual impartiality, and educational equity. Such tools should minimize any disparities (such as false positives) across all subgroups. Ideally, such tools require contextual sensitivity in which their features do not penalize for linguistic differences. Fairness also requires full and open transparency in terms of defining the thresholds and providing reproducible metrics in the performance of such machines. In the learning environment, fairness means that such devices should not result in disproportionate harm to the student, such as severe discipline or reputational damage [18]. This paper investigates the threshold levels of four machines to confirm or reject the following hypotheses:

> H0: *The proportion of human-written texts misclassified as AI-generated is the same across all four detectors.*

The alternative hypothesis is:

> H1: *The proportion of ESL texts misclassified as AI-generated is higher than other texts, while scores differ across all four detectors*.

A confirmation of the null hypothesis (H0) would mean that any observed differences in false positives (FP) would be due to random variation and not intentional bias. On the other hand, a rejection of the null hypothesis and confirmation of the alternative hypothesis would provide evidence that such tools can misclassify ESL texts that results in unfairness.

1.2 Understanding AI Detectors

As already noted, commercially available AI detectors come in various types and serve different purposes. While consumers may be naturally confused about which to adopt, the important point is that no such device can predict or verify the absolute truth as to whether an element of writing is AI or human-generated. These are probabilistic machines that measure a number of features such as perplexity, burstiness, repetition, semantic richness, entropy, idiomaticity, and syntactic variety (just to name a few). The definitions of these features derive from a cross-section of theories such as Computational Linguistics, Natural Language Processing, Machine Learning, and Information theory

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

[19]. While there are many features, Fig. 1 below shows four key features that relate to this study.

Feature	Definition	Analyzed at	How measured
Perplexity	Language model confidence in predicting word sequences	Word or sub word level	Averaged across the entire text. Lower Perplexity = more Predictability.
Syntactic Complexity	Level of sentence structure	Clause & sentence level	Uses dependency graphs to assess the use of Subordinate clauses or modifiers
Semantic Richness	Depth and diversity of ideas Despite Syntactic complexity	Phrase and sentence level	Embedding models that assess meaning by phrase coherence and sentence level
Lexical Diversity	Variety of unique words in a text	Word-level & document-level	Uses Type-Token Ratio (TTR) counts unique words versus total words across entire text

Fig. 1: Four Features Commonly Measured and Classified by AI Devices

As Fig. 1 above shows, these features are measured at multiple levels (from word to document), then transformed into a classification model, piped into algorithms and thresholds that provide a probability score as to their origin (either human or AI-generated). However, depending on the brand, the thresholds may be too rigid or uncalibrated for under-skilled or ESL writers [20]. Since ESL writers often use simpler clauses and repetitive vocabulary, this could sway the metrics at different levels. Furthermore, such devices may over- or underemphasize perplexity (which is why many studies focus on this issue) and misclassify authentic writing as AI-generated [21]. Finally, semantic richness is particularly sensitive to idiomatic phrasing and cultural context. A Japanese student might write the following sentence: *Although my friend said John was smart, I was surprised to see how heavy John was.* A detector might parse this sentence as syntactically complex (subordinate clause), but lacking diversity (*John* twice and *was* three times), and perhaps logically unclear, so that it is semantically poor [22]. In addition, it may miss the contextual use of the word "smart," which can mean "thin" to some Japanese learners.

Essentially, these devices act in a similar way to airport screening machines. Fig. 2 provides a basic conceptual model of the two levels of diagnostics that includes measurement systems of algorithms (the scan) and thresholds (settings). The full scan requires a four step process: (1) the raw data (blue); is inserted into the machine (2) the scanning (green) occurs with feature extraction (perplexity, lexical diversity) and then assigns as a score; (3) these scores are matched to the threshold settings (orange) and given a binary label; and (4) an output label of "likely" AI or human generated is delivered to the user [23]. As this detecting measurement system involves two levels of assessment (algorithms and thresholds), this means that problems can arise at either level or both. Problems at the algorithmic level can lead to structural bias depending on how well they are "trained" in classifying linguistic variation, such that errors can result in penalizing ESL writing. Even if a well-trained scan provides a reliable score on a feature, a poorly calibrated threshold (ie, a setting that is too high or low) could result in procedural bias that also misclassifies a text. The scope of this study is at the threshold level.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

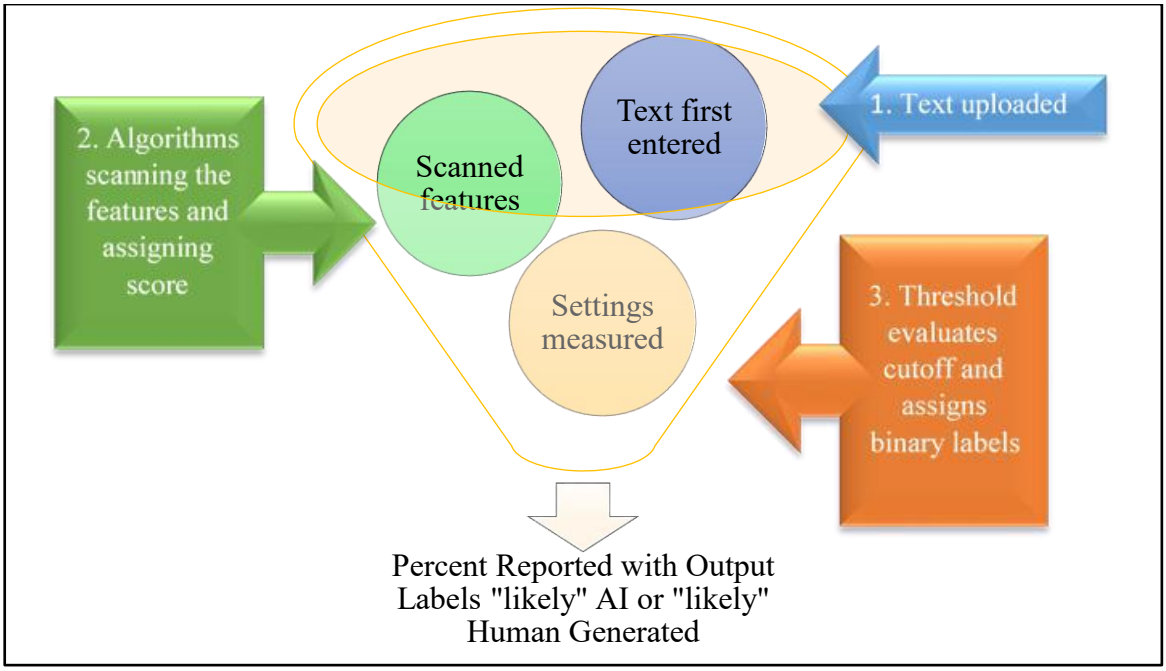


Fig 2: Conceptual Model of AI-detectors as Scanners

II. LITERATURE REVIEW

Gotoman et al. (2024) conducted a systematic literature review of 34 scholarly studies from three online databases in order to assess what the research found regarding commercially available detection devices [24]. They noted eight evaluative approaches, with the main three being concerned with accuracy, reliability, and fairness. The findings revealed that while many detectors achieved above 50% accuracy rates, in total, they remained unreliable. Most studies also indicated that paid or premium machines outperformed free versions. In terms of fairness, the consensus was that such imperfect tools should serve as supportive evidence and not as a final judgment. The authors concluded that such technology needed improvement in terms of transparency, fairness, and the strength of the measurement system. Selectively, the remainder of this review will discuss research that aligns with the aims of this paper associated with investigating how such tools may misclassify human text (especially ESL) as AI-generated.

While the research regarding AI text detection and ESL writing in the educational or multimedia context has expanded, much of it has focused on the algorithmic level. For example, Chaka (2023) reviewed 17 studies by combining corpus analysis and qualitative synthesis to examine how such devices may be misclassifying authentic writing [10]. This evaluation revealed that structural uniformity in the devices triggered false positives in ESL writing, but it was concerned with fairness. The author advocated that educators should triangulate such tools with other devices, along with human judgment.

Meanwhile, Liang et al. (2023) employed stratified sampling and cross-detector benchmarking to evaluate seven commonly used detectors on 40 TOEFL essays [11]. Their study found that these tools misclassified 61% of the ESL writing compared to the high accuracy of essays produced by native authors. This highly cited research revealed that such tools do indeed target linguistic variation common within ESL text. The Liang study maps well with this present study, which is concerned with semantic richness, threshold sensitivity, and stratified fairness auditing.

Echoing some of the same concerns, Price and Sakellarios (2023) sampled 120 essays written by Japanese college students with several commonly available detectors [25]. Their research also found a high number of false positives, especially among lower-skilled learners. They noted that such machines misinterpreted features such as lexical simplicity and syntactic repetition as AI-generated text, and that threshold levels varied widely among the devices. They further concluded that such misclassification can result in pedagogical risks to such learners. This aligns well here because it notes that fairness concerns are dependent on a complete evaluation of both levels of such detection systems.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

Li and Wan (2025) produced a large-scale empirical study using 483,360 student essays to benchmark classifiers with six detectors (3 English and 3 multilingual) based on features such as perplexity and lexical richness [26]. The study employed Random Forest models and stratified sampling across academic fields, finding that at this algorithmic level, false positives occurred at a high rate in both categories, which would impact ESL writing. This study is relevant here because it analyzed several detectors, found data suggesting false positives that can be generalized, and suggested a need for adjustments. While this study implied potential problems in authenticating ESL writing, Pratama (2025) used similar devices to analyze 300 scholarly abstracts from both native and non-native authors [12]. The results revealed that such tools disproportionally flagged ESL abstracts as AI-generated. In addition, the aim of the study was to improve educational integrity.

Pudasaini et al. (2025) tested five detectors with 1,500 writing samples from academic, journalistic, and evasive LLM outputs (ie, camouflaged AI-generated text) [13]. Such a strategy is related to robustness studies rather than a fairness audit. With such testing, they indeed found that thresholds degrade under real conditions. While such a study centers around robustness and accuracy, its findings support fairness-oriented research by showing how such devices perform weakly with linguistic variability in academic and multilingual writing samples.

Together, these studies form a layered map of the linguistic feature analysis, the scope, the approaches, sampling, and methodology that can be used in studying such devices. For comparison, Fig. 3 below provides a summary of the review as each of the studies aligns with this paper. The present study builds on this foundation by integrating semantic richness and cohesion metrics, modeling threshold sensitivity, and visualizing bias through stratified confusion matrices as well as advancing a reproducible framework for ESL-aware fairness auditing.

Study	Scope	Features Measured	Sample Size & Type	# Detectors Analyzed	Approach	Main Methods
Chaka (2023)	Algorithmic	Grammar interference	40 L1 & L2 essays	30	Fairness	Comparison across platforms
Liang et al (2023)	Algorithmic	Stylistic shifts, reviewer behavior	91 TOEFL & 88 L1 essays	7	Adversarial	Corpus-level estimation of LLM influence
Price et al (2023)	Thresholds	Grammar confounders	Japanese university 12 essays	5 Free detectors	Fairness	Manual vs. automated detection comparison
Le & Wan (2024)	Algorithmic	Perplexity false positives	480,000+	6	Adversarial	Inverse perplexity-weighted ensemble
Pratama (2025)	Thresholds	Detection metrics, disciplinary bias	71 academic articles	3	Fairness	Accuracy vs. bias trade-off analysis
Pudasaini (2025)	Algorithmic	Paraphrasing robustness, evasion tactics	6,000 human 6,000 AI texts	3	Adversarial	Benchmarking detectors with adversarial inputs
Lege (2025)	Thresholds	Threshold recalibration	1212 Stratified texts (Japanese university context)	4 (3 free, 1 paid)	Fairness	Performance metrics, Stat. residuals

Fig. 3: Chronological Map of Key Studies on AI-Generated Detection

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

III. METHODS

In terms of inputs, this study investigated and performed an interdependent fairness audit on three free versions of such tools (Copyleaks, ZeroGPT, and Scribbr) and one paid version (Quillbot-Premium) to compare and detect whether written texts (N=1212) were AI-generated or human-produced. These tools were selected due to ease of access, because they are marketed toward education and publications, and the research suggested low to moderate issues with false positives The writing examples consisted of chunks of texts (400-500 words) produced from three classes or groups: actual AI-generated texts (n=307), portions of scholarly published articles (n=302) available on the internet that predate the arrival of Open AI technology, and sections of text from ESL theses (n=303) written at Nagoya University in Japan prior 2021. Thus, with a 100% grounded truth value, this study employed descriptive and inferential analysis to audit the performance of fairness in four AI tools that many educators presently use to evaluate student writing.

Articles by Chaka (2023) and Gotoman et al. (2024) uncovered at least seven methodological approaches to the study of such devices, including adversarial tests, accuracy and error analysis, content obfuscation sensitivity, cross-domain generalization, fairness audits, human-AI discrimination, and watermark detection. The approaches to such studies are typically divided between adversarial (the *why* and *how*) or a fairness (*who* is impacted and *why*). As the column on the scope shows in Fig. 3, several known studies have explored similar risks with such devices, but looked at either algorithmic scanning or the threshold settings. Ideally, both should be done to ascertain a full understanding of the problems associated with such devices, but practical considerations restrict such studies to a single approach.

While two of the studies did look at threshold levels (Price et al. and Pratama), they relied on descriptive statistics to explore *who* was impacted, but did not conduct performance metrics to address *why* this occurs. This present study expands beyond the typical fairness study by investigating *why* such devices may be misclassifying scores. Typically, the adversarial approach incorporates methods to investigate technical weaknesses with the scan (robustness) rather than social vulnerabilities (as to who is impacted) [26]. However, when the case involves unintentional bias, then a fairness audit might be of use to probe how a threshold can fool a device when settings err when scoring demographic groups [27]. Such threshold instability can result in misclassification and occurs when there is over-reliance on algorithmic features that may be exaggerated, flawed, or biased. Though the scope of this study cannot fully explore the scanning level (algorithmic features), inferential statistical analysis will provide clues as to potential problems at that level.

Going back to the airport scanner analogy, if any of the pixels within the lens (algorithmic features) are smudged or misaligned, then this could result in a blurred image or misleading result (threshold score). In this instance, each lens represents one algorithmic feature set. If the scan distorts the scan of any feature, then the algorithm bias the threshold scoring. Such distortion occurs due to poor designing, training, or functioning of the algorithms and is equivalent to a coarse or grainy quality in the scan. This, in turn, can result in algorithmic bias. Meanwhile, the threshold slides up and down between sensitivity and specificity settings that bring a macro or global view of the features [28]. Thus, as a scanner device requires care and calibration, AI-detectors need properly selected features to make fair decisions, particularly with ESL texts.

The methods applied to assist with a fairness audit at these two levels fall under the toolboxes of recalibration (of thresholds) and stylometric profiling (of the algorithmic features) [29]. The main focus in this study is recalibration. As Bellamy et al. (2019) noted, recalibration methods involve post-hoc techniques that assist in showing the strengths and weaknesses of a device while adjusting the scores of probabilistic classifiers (such as detectors) to show possible improvements in correct evaluations [30]. As a form of reverse engineering, investigating threshold calibration helps to identify unfair outcomes, such as false positive rates. This first level of investigation involves using a confusion metric to help show potential misclassification, performance metrics to identify optimal thresholds, a Levene's Test to justify whether threshold instability exists, and a Chi-square analysis to establish statistical justification that there is significant misclassification across the three groups.

With the assistance of this recalibration approach and inferential statistical analysis, residual data will provide clues that problems exist at the algorithmic level, thereby hinting at the reason for the outcome errors. As an example, the study examines four linguistic traits to understand which stylometric signals may be unintentionally causing bias. The four features include perplexity, syntactic-semantic complexity, and

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

lexical diversity in educational equity. While there are indeed many features, these four are selected because they align with the four aspects of fairness (perplexity to statistical analysis, syntactic-richness to structural transparency, semantic complexity to contextual impartiality, and lexical diversity to educational equity) [31]. Furthermore, these features are relevant because research indicates that ESL writing diverges from native writing with all four, and they since they allow for a visual quantification of where the detectors may be misclassifying due to linguistic bias [32]. Thus, the bridge from recalibration or threshold tuning to stylometric analysis allows the study to help educators see the limitations in such devices to ensure that they are applied equitably across diverse writing populations.

The recalibration analysis first incorporates descriptive raw data in the form of a contingency table to assist in showing the averages, mean scores, and standard deviation in the three submitted forms of writing (AI-generated, published text, and student text), which helps identify *who* is most likely affected by such misclassifications. A Levene's test was used to check the significance of some of the variances of the standard deviations. The next step includes the use of performance metrics to illustrate the impact of thresholds in assigning false positives in both the raw data and with recalibration. The study follows with a Post Hoc Chi-square test to compare the significance of the assignment of false positives by each of the devices. To strengthen the results of the Chi-square comparison, the study includes a standardized residual test that hints the problem is not simply at the threshold level but is occurring at the algorithmic level (that is, the *why*), and such misclassifications may be occurring.

IV. RESULTS

The recalibration approach provides insight into how such mechanical devices use thresholds for classification. Using raw data, descriptive and inferential statistics, and adjustments to the threshold, it is possible to reweigh prediction probabilities and visualize *who* may be affected by such scores. With the help of a confusion matrix, performance metrics, Levene's Test, and Chi-square, the study will establish that ESL writers were most likely to be given a false positive score with such tools. Alone, such a method can assist with aligning outputs that improve FP parity. The assistance of a standard residual test will hint that the reason why such scoring occurs may be found at the feature or scanning level as well.

Table 1 represents a form of contingency table or summary matrix showing all the raw data collected from the results of the devices as they were assessed for AI-generation. Here, if any device scored a text (even 1%), then it is listed as being flagged for AI. Each machine evaluated 1212 texts (307 AI-generated, 302 published texts, and 303 ESL texts). As the table shows, all the machines detected automated texts correctly (scores ranged 88-100%), suggesting a large number of true positives (TP). On the other hand, these tools flagged published text 33.6% of the time (406/1208) and, more importantly, ESL text at 69.9% (809/1212). Though not shown, the range for the actual scores for each text was 88-100 for AI-generated text, 1-32 for published, and 1-52 for ESL, indicating a strong true positive rate for the AI texts, and some degree of false positive assignment for the human written texts.

Table 1: Summary Matrix of Raw Data across Four Devices Flagging for AI-generation

Tool	**Binary # Flagged Scores**			
	AI-text (TP)	**Published(FP)**	**ESL (FP)**	**Total**
Copyleaks	307/307	127/302	232/303	1212
Zero	307/307	104/302	219/303	1212
Scribbr	307/307	101/302	211/303	1212
Q-Premium	307/307	74/302	177/303	1212
Totals	1228/1228	406/1208	809/1212	

Continuing with the presentation of the raw data, the apparent disparity between the strength of the devices in terms of recognizing AI-generated text while struggling with some aspects of identifying human texts justifies further investigation. Because the human-generated texts predate the onset of commercially available AI technology, the observed texts have a grounded truth value of 100%. Even though the actual threshold settings for the devices are unknown (could be set between 20-80%), as this tends to be proprietary

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

information, having a strong grounded truth value (knowing that observed values are true) is critical to building performance metrics [33].

At this stage, the main disparity in the raw data between the two human texts hints at potential bias toward ESL writing. The fact that such tools flagged published texts with actual false positives (33.6%) at all is surprising, but the more than double rate of labeling ESL (69.9%) raises even further concerns. This disparity exhibits a systemic fairness issue, suggesting that many of the current detection models on the market may incorrectly conflate linguistic variations in writing differences. Such findings reinforce the need for an evaluation of their actual capabilities. The next step required an examination of the raw continuous scores (1-100) of the devices assigned to each individual text to establish the extent of the difference in scoring between the two human-written texts.

Table 2: Average of the Continuous (1-100) AI-Generation Scores per Writing Text

Group	# Articles	Copyleaks	Zero	Scribbr	QP
AI-gene.	307	96.8	97.1	98.3	98.5
ESL	303	17.2	15.6	15.4	9.4
Published	302	3.9	4.5	5.9	3.1

Table 2 above is a descriptive summary of the average continuous scores that each of the devices gave after submitting them for AI evaluation. As the table shows, all four devices largely detected the actual veracity of AI-generated texts (true positives), though imperfectly. Each of these devices appears quite capable of identifying true positive scoring for AI-text, with Copyleaks performing the weakest, with scores averaging 96.8% and Q-premium the best at 98.5%. Since these devices were not perfect, this raises a question as to the acceptable amount of error (true negative) that would be acceptable. Typically, such an acceptable error rate would depend on the purpose of use and could be less than 2-5% in terms of legal or policy development or for defending academic integrity [34]. The average for true negatives (TN) for the AI-texts in this instance ranged between 1.5 (Q-premium) and 3.2 (Copyleaks), which is calculated by subtracting the average scores from 100. Therefore, the error rate here (total TN average of all four devices), while questionable, is within the standard acceptable margin at 2.34%. Thus, such a device appears strong at correctly identifying AI-produced text, but there could be some issues.

However, Table 2 also shows that the average amount of error or false positive rates (FP) for the human texts (published and ESL) shows averages above the acceptable rates. The range of averages for FP for the published texts is 3.1 (Q-premium) and 5.9 (Scribbr); meanwhile, the range of the average FP scores for the ESL texts is 9.4 (Q-premium) and 17.2 (Copyleaks). While all four devices assigned false positive (FP) scores for all of the human texts, there exists an obvious difference between how the tools evaluated the published texts (total average of 4.15 FP) and the ESL texts (total average of 14.4). As the table indicates, these devices scored the ESL with higher FP scores by 3-4 times relative to the published texts. At this stage, more analysis is needed to confirm that the ESL texts were subjected to unintentional or systematic bias.

Since the error rate of the FP for the published text (4.15) is closer to the TN averages of error for the AI-text (2.34%), a t-test is needed to understand more clearly if the devices are tagging published texts closer to the acceptable TN rate for the AI-generated texts. The results from a t-test compared the TN and FP rates across the four tools revealed a consistent misclassification bias with a difference between the means of 2.025. While the results did not reach an actual statistical significance of $\alpha = .05$ level ($t(3) = 2.66$, $p \approx .08$), the magnitude of the difference in the mean average suggests a practical difference (keeping in mind that TN is an error even at the smallest rate). While these tools did flag some AI-text as human-written, though at small rates, they simultaneously over-flagged human texts that appear to disproportionally affect ratings for ESL writing. Such findings support the need for recalibration, as current thresholds may be misaligned with linguistic diversity. A further look at the mean and standard deviation for each device may illuminate these differences.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

Table 3: Mean Scores and Standard Deviations for the Continuous Scoring

	AI-generated		ESL		Published	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Copyleaks	96.5	3.3	17.2	12.1	3.8	4.2
Zero	96.8	4.5	15.6	13.7	4.6	5.1
Scribbr	98.3	5.8	15.2	14.5	5.5	6.3
Q-Pre	98.4	2.6	9.3	10.2	3.2	4.0

Table 3 above presents the mean scores and standard deviations for each device as they relate to the three different writing forms. As shown, the high means for the AI-generated texts, along with the tight clustering of the SDs, illustrate that these devices are quite adept at identifying when a text is fully AI-written. In contrast, such tools assign FP scores to ESL texts at a much higher rate than native published writing. The table shows much higher means (up to 17.2) and greater variability (SD 14.5) than with the published group, which indicates much lower means and smaller variability. In addition, for the three free versions that measured ESL writing, the SDs had a wider spread but were still more clustered than with the published text. When such tools show low SDs and high false positives, then this is an indicator of algorithms set toward rigid heuristics (targeting unusual grammar patterns, for example) [35]. In general, then, these descriptive findings raise questions about the accuracy and fairness of such detection tools in analyzing ESL writing outcomes.

The lower SDs for the three free versions compared to their mean for the free versions (Copyleaks, Zero, Scribbr) show strong enough clustering of scores that AI may be unintentionally targeting ESL writing for two reasons. First, the means for the FP scores for the published text are small (in fact, closer to the means of true negatives for the AI-generated text). Second, the SDs show less clustering, which may suggess erratic classification rather than bias. Indeed, five of the standard deviations are greater than their means, suggesting either a statistical anomaly or something more subtle. The standard deviation for the Q-premium (10.2) scores for the ESL texts was slightly higher than the mean (9.6), perhaps reflecting the possibility that premium models calibrate to be more sensitive to false negatives than false positives [11], [28].

Furthermore, all the standard deviations for the devices that evaluated the published articles (4.2, 5.1, 6.3, & 4.0) in Table 3 were slightly higher than their corresponding means (3.6, 4.6, 5.5, & 3.2). Essentially, this means that scores were smaller and spread more widely across the published texts group. Compared to the ESL group, the difference suggests several possible things, for example, a wider dispersion of scores due to variations in native writing skills, some form of internal calibration bias, or just a fluke. Thus, these variances between ESL and the native writers require an inferential test for significance.

The present differences in the case of the larger SDs shown in Table 3 could suggest three main things. First, if the writer's scores actually show lower variance, this might suggest bias in the detection devices. Second, a greater variance might hint at inconsistent treatment. Third, if the variance is moderate, then this could reflect various subtleties, measurement error, or statistical noise. To clarify this issue, this study employed a Levene's Test, often used in educational research, to assess the variance of differences across groups [36]. While such a test cannot isolate which of the groups (ESL or native) was subject to biased treatment, it can show that the variance of scores targeted at least one of the groups.

Table 4: Results of Levene's Test across the Four Devices for the ESL and Published Texts

Device	Levene's F score	p-value	Interpretation
Copyleaks	28.37	<.0001	Scores have significant variance
Zero	31.22	<.0001	Strong evidence of unequal dispersion
Scribbr	26.45	<.0001	Scores vary widely
Q-Pre.	19.88	<.0001	Score shows a tighter but erratic spread

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

The results from Levene's Test in Table 4 indicate a high degree of variance for at least one of the writing groups. This is a global test that signals one group may have a wider variance or spread in scores, but cannot identify which group. This test revealed significant variance differences between the two groups across all four detectors (F range: 19.88–31.22, p< .0001), indicating heteroscedasticity (variance is not uniform), which could imply potential fairness concerns. An F-score above 10 is considered quite high and, combined with a p-value of <.0001, suggests that overall score dispersion is unequal between ESL and native writers. Essentially, one group has more clustered FP scores relative to the other. While clustering suggests potential uniform bias, a wide spread in the scores implies an inconsistent application of the scoring when attempting to judge a text as A-generated. As such, when detectors show such inconsistency when assigning false Positives across devices, then this indicates systematic bias [12].

By combining Levene's Test with the higher ESL means and SDS, it becomes apparent that the devices allotted higher FP rates to this group. While such a variance measure in the raw data can signal instability in a tool, it cannot reveal how accurate or fair such a device may classify texts across different groups [37]. As such, the next step is to utilize performance metrics to evaluate the practical implications of the disparity found above. With the assistance of metrics such as precision, recall, and false positive rates, this stage will quantify the extent to which ESL writers are misclassified and assess whether the devices meet equitable standards of reliability and fairness.

Table 5: Confusion Matrix for each Device using ≥30% Threshold

Actual	**Devices and Predictability (AI or Human) AI≥30%**							
	Copyleaks		**Zero**		**Scribbr**		**Q-Pre**	
	P-AI	**P-H**	**P-AI**	**P-H**	**P-AI**	**P-H**	**P-AI**	**P-H**
307 (AI)	307 (TP)	— (TN)	307	—	307	—	307	—
303 (ESL)	37 (FP)	266 (TN)	28 (FP)	275 (TN)	19 (FP)	284 (TN)	2 (FP)	301 (TN)
302 (Published)	1 (FP)	301 (TN)	—	302 (TN)	1	302 (TN)	—	302 (TN)

Table 5 provides the results of a standard confusion matrix set with an idealized threshold of 30%. While typically the standard for such machines is supposed to be 50%, such settings for commercial devices tends to be a trade secret and independent research can only infer such threshold settings[38]. Since the raw data showed many scores less than 50%, the assumption is that these actual market tools were set somewhere between 20-50.% Recalibrating at the 30% threshold, in this context, the classification occurs by designating all scores above 30% as FP. As the table shows, at the 30% threshold, all the detectors accurately identified the AI-generated texts as true positives (TP) with scores well above 30% (88-100). Moreover, only two of the published next showed a score over 30% (Copyleaks 32; Scribber, 32), which indicates that at this level the devices only misclassified two texts with a false positive. On the other hand, the devices misclassified 86 of the ESL writing with false positives (Copyleaks, 37; Zero 27; Scribbr, 19, Q-premium, 2).

As a form of descriptive analysis, the confusion matrix provides only limited insight into the four scoring classifications (TP, TN, FP, and FN). However, as Neeley and Englehart (2025) noted, this form of analysis helps to transition to an even more powerful metric that measures accuracy, precision, recall, specificity, and F1 score [39]. These performance metrics normalize raw counts across group sizes and allow for meaningful comparisons between detectors, especially when assessing fairness toward the ESL writers. For example, precision helps quantify false positives, while recall ensures that AI-generated texts are detected reliably. Performance metrics also help audit the detector calibration, sensitivity, and bias, which is critical in any decision regarding their use, especially in education [40].

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

Fig. 4 below highlights the essential formulas and meaning for these metrics as drawn from the confusion matrix. For example, when calculating the accuracy of the Copyleaks device, the numbers from the binary classification are plugged into the following formula: (TP +TN) ÷ (TP +TN+FP+FN) to provide a value. After submitting the input values from the confusion matrix, the accuracy percent for Copyleaks would be 95.95% calculated from (307 +569) ÷ (307+569+37+0). Thus, the next step.

Metric Analyzed	Formula	Meaning
Accuracy	(TP +TN)÷(TP +TN+FP+FN)	Overall correctness of the device
Precision	(TP) ÷(TP+FP)	Number actually AI-generated
Recall	(TP) ÷ (TP+FN)	Number of actual AI correctly identified
Specificity	(TN) ÷ (TN +FP)	Number of actual Human texts correct
F1 score	2 x (P x R) ÷ (P +R)	Harmonic mean that balances precision and recall

Fig. 4: Clarifying Performance Metrics

The total performance metrics for all four devices at a 30% threshold are presented in Table 6 below. Most notably, the free versions showed lower scores in all the major categories relative to the premium. That is, they were less accurate, less precise, given to assign more FP, and were less balanced (again suggesting thresholds were not set at 50%). The recall for all devices was 100% indicating that such tools could identify an actual AI-generated text at this threshold (no false negatives). The term specificity refers to the degree to which the devices recognized when a text was actually human-generated, and here, Q-premium scored the highest at 99.68% while Copyleaks lagged at 93.98%. The F1 score establishes the degree of balance or harmony between precision and recall. As Table 6 shows, the Q-premium was more balanced (99.34%) than the free versions, suggesting their thresholds may be set between 30-50%.

Table 6: Performance Metrics for all Four Detectors

Tool	Accuracy	Precision	Recall	Specificity	F1 score
Copyleaks	95.95%	89.2%	100%	93.98%	94.29%
Zero	96.77%	91.84%	100%	95.82%	95.73%
Scribbr	97.56%	94.17%	100%	97%	97.44%
Q-premium	99.78%	99.35%	100%	99.68%	99.34%

At a 30% threshold, the performance metrics above demonstrate that while all four tools exhibit strong capability to identify actual AI-generated texts; however, there remains some variability in the treatment of human-generated texts, especially for the ESL writers. Such metrics provide a better picture of the severity of the scoring patterns. Essentially, the devices tend to target ESL writing across the board, but scoring above 30% occurs with less frequency. While these metrics assist in quantifying the overall behavior of the digital tools, they do not test whether the observed differences across writing groups and detectors are statistically significant. As such, the need for statistical significance warrants the use of a Chi-square test of independence. This type of test can assess whether the apparent variation in the rates of FP occurs from systemic bias within the digital technology or if this is due to random chance. This shift from descriptive metrics to inferential statistical analysis strengthens the fairness audit.

This study ran a Chi-square test of independence from the performance metrics threshold of 30% from the observed totals in the recalibration for the four devices. Even at this threshold, the test revealed a significant relationship between the digital tools and the ESL false positives. The degree of freedom (df) was 3, and the critical value was 7.81. The test result exceeded the critical value at 32.03 ($p < .001$), which indicates that the assigned FP scores for the ESL writing were not due to chance but rather the result of calibration bias within the detectors. Thus, these findings reject the null hypothesis and support the hypothesis that the threshold settings in AI-detectors can impact fairness outcomes for ESL writers.

However, the chi-square test above only produced a statistically significant connection between

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

the ESL false positive rates and the four digital tools, but did not clarify which of these devices contributed most to the apparent bias. As Shan and Gerstenberger (2017) opined, a Post Hoc Chi-square comparison would be applicable here as a way to isolate where the significant differences lie between such devices [41]. Such a step is important to help identify if a specific detector may be impacting the overall effect by comparing the tools with each other. Such a comparison provides a more subtle look at the significant difference in bias patterns that would make these results more useful in system calibrations and for educators endeavoring to use the tools in a triangulated way.

Table 7: Post Hoc Chi-square Comparison of the Four Devices for False Positives

Compared Pair	Chi-Square	Deg. of Freedom	P-value	Significance
Copyleaks v. Zero	1.35	1	p<0.245	None
Copy v. Scribbr	6.91	1	P<0.009	Highly
Copy v. Q-pre.	30.91	1	p<0.00001	Extremely
Zero v. Scribbr	2.49	1	P<0.114	None
Zero v. Q-pre.	18.23	1	p<0.0001	Highly
Scribbr v. Q. pre.	9.52	1	p<0.002	Highly

Table 7 provides results from the Post Hoc Chi-square comparison. The point of this test was to isolate which of the technological tools may have skewed the disparity in FP for all of the devices. The table reveals that all of the free versions contrast significantly with Q-Premium, indicating that these three devices were more prone to assigning false positives. Since there were only two categorical variables in each case, the degree of freedom (df) was 1, making the critical value 3.84 At this juncture, the comparison identifies that the bias is significantly concentrated in the free versions, which may inform educators about which of these tools (or a combination) they might adopt or avoid in the learning environment. .The most extreme comparison occurred between Copyleaks and Q-premium (x^2 =30.91, p<.00001), suggesting that, in this instance, the free version was essentially much more biased. However, since the comparative performances between Copyleaks and ZeroGPT, as well as Zero and Scribbr, show no significant difference, another test may be necessary to clarify the residual effects and help to see if more study is necessary at the algorithmic level.

Sharpe (2015) recommended using a standardized residual test to further strengthen the Post Hoc comparison [42]. This additional test assesses how each of the devices' observed values may have deviated from the expected count under the null hypothesis. This type of test represents a micro-level diagnostic that can reveal which of the tools had the largest impact on the overall chi-square signal, thereby improving the fairness audit of such devices. Much like z-scores, standardized residuals measure the degree to which an observed count deviates from its expected count, scaled by the standard deviation (±2). The formula for such a calculation is Standard Residual = (Observed Cell minus Expected Cell) ÷ √ Expected Cell.

Since Copyleaks appeared as the most isolated in the Post Hoc comparison, the data from this device will serve as an example to show how the calculations work and can actually be done by hand. Fig. 5 below illustrates the two essential steps: first, calculating the expected cell values, and second, calculating the residual score. As the figure shows, after calculating the expected values from the 30% threshold contingency table for FP, it is now possible to obtain the standardized residual by plugging the inputs into the formula. The results show that the standardized residual is +3.27, which exceeds the scaled standard deviation of ±2. This, in turn, indicates statistical over-prediction on the part of this tool. By running standardized residuals for all four of the devices, it is possible to confirm the extent to which this device was the main contributor in assigning FP.

Scribbr ESL False Positives		
Step 1. Calculation of expected values		
	Cells	Amount
	Observed FP (O)	37

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

	Total ESL Texts	303
	Total Devices	4
303 x 4 =1212	Total ESL Prediction	1212
(84 ÷ 1212) x 303=21	Expected FP (E)	21
Step 2. Calculation of residual		
(36-21) ÷ √21 = (15) ÷ 4.58 = +3.27	Standardized Residual	+3.27

Fig 5: Illustrating the Calculation of the Copyleaks Device for Standard Residual

Table 8 below displays the results from the calculation of the 2x4 standardized residuals (FP and TN) for the four devices. This dual-row examination offers educators and researchers several insights. First, this approach establishes the importance of how Chi-square tests work and why full contingency tables are essential for interpreting bias in algorithmic systems. Second, the contrast in the results illuminates which of the devices contributes to potential bias. Third, since many teachers are concerned about the pedagogical risks associated with false negatives [7], such an approach illustrates the see-saw effect or inversion between FP and TN when considering such a tool. In this instance, the table clearly shows that while Copyleaks is more sensitive to FP (+3.27) when evaluating ESL writing, it is the least sensitive to TN (-0.39). The opposite is true for Q-premium, which may be setting their algorithms to be more sensitive to TN (+.5), while overcompensating for FP (-4.15). Of the four devices, Table 8 shows the free version of Scribbr to be the most balanced with respect to evaluating ESL writing.

Table 8: Comparison of Standardized Residuals for the Four Devices (FP and TN)

Detector	**Residual (FP)**	**Residual (TN)**
Copyleaks	+3.27	-0.39
Zero	+1.45	-0.17
Scribbr	-0.17	+0.02
Q-Premium	-4.15	+0.5

V. DISCUSSION

In general, the findings suggest that while these tools perform quite well with actual AI-written outputs, their capability to classify human writing diverges significantly. The lack of variation in the scoring of the AI-generated writing establishes that such detectors are fairly effective in identifying synthetic content [43]. On the other hand, the significant variability in FP of the human writing raises concerns about consistency and reliability. The most striking result was the large number of misclassifications of ESL text at both the grounded truth level and the recalibrated 30% threshold. Free versions of Copyleaks and ZeroGPT flagged ESL writing at high rates. In total, this suggests that these tools may erroneously tag linguistic features common in ESL writing as AI-written. The results of the study reject the null hypothesis that the proportion of human-written texts misclassified as AI-generated is the same across all four detectors.

The results clearly support the hypothesis that the proportion of ESL texts misclassified as AI-generated is higher than other texts, while scores differ across all four detectors. To help test the hypothesis, the analysis represented a recalibration of the threshold aspect of these detectors which assist in identifying *who* is more impacted by such scanning tools. The threshold settings assign the outcome based on a scale between sensitivity and specificity (false positives and false negatives) based on the clarity of the algorithmic measures. Commercially available detection tools do not calibrate on grounded truth values, and they do not publish at what level they set thresholds. So, even with a fairness audit (ie, moving the threshold measure between 30-70%), users of such devices cannot assume that such tools reliably confirm that learners have used AI technology for assignments. Misinterpreting how these devices function risks reinforcing assumptions about ESL writing because this learning group may receive allotted FP scores (even at higher thresholds) due to linguistic features similar to AI-generated writing patterns, despite being originally written.

In addition, the fact that there were so many lower scores at the grounded truth level compared to

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.5

the 30% threshold does not guarantee fairness. Any device that flags writing at a 17.2 mean score (Copyleaks) may still misclassify authentic ESL writing at higher threshold settings, particularly if the algorithmic canning amplifies the markers in the linguistic features. Without transparent threshold settings and clear validation of the features, users of such devices should interpret findings cautiously. These outcome statements represent only one possible signal that requires contextual review and are not definitive proof of misconduct [44].

For educational institutes and educators, the results here suggest that they should resist the urge to treat scores from such thresholds as hard evidence rather than as one probabilistic signal. Instead, such scores should prompt those in education to engage more with student writing, including pre-and post-diagnostics, revision history, assignment scaffolding, and continuous dialogue. As such, fairness in AI detection is not simply about accuracy but ensuring that educators are not penalizing ESL writing that reflects their linguistic background [18]. While further study is needed, the use of the standardized residual comparison suggests that problems are occurring at the algorithmic level as well.

2 CONCLUSION

In the educational and evaluative environment, teachers and administrators are increasingly dependent on AI detection tools to protect academic integrity. Such devices are being employed to verify the veracity of student assignments; however, the reliability of such tools varies depending on the thresholds and linguistic characteristics of the inputs. Fairness audits can assist in comprehending the issue at the thresholds, while accuracy audits would analyze the algorithms. A study of both would require more written space as a practical matter.

As a fairness audit, this paper demonstrated that while detectors are effective at identifying fully AI-generated texts, such tools show inconsistent and biased classification of authentic human writing, especially from ESL students. While the sample size could always be larger and more diverse, the amount of text here was sufficient in establishing significance in the findings, especially since the grounded truth value was very strong. As such, the significant number of false positive rates for the ESL writing, even after applying the performance metrics, suggests that such tools may confuse non-native linguistic writing patterns with AI features, which in turn can result in the misclassification of ESL writing.

The findings highlight the importance of fair thresholds while refining detection algorithms to reduce potential bias. The point is that problems with such devices can occur at both levels of the scan. As such, Educators and institutions intending to use such devices must approach these tools with care to ensure that they are applied equitably. Future research should expand the scope of a study to analyze both thresholds and algorithms (specific features) while exploring mitigation strategies that promote responsible and fair use of such detection technology.

REFERENCES

[1] Anara, M.K. (2025, July 15). *AI in education statistics: How artificial intelligence is transforming higher education*. Anara. https://anara.com/blog/ai-in-education-statistics (Accessed 6 August 2025).

[2] Holmes, W. (2023). *Asian and European teachers' perspectives on AI and education*. Asia-Europe Foundation (ASEF). pp 1-56. https://asef.org/publications/asian-and-european-teachers-perspectives-on-ai-and-education (Accessed 1 August 2025).

[3] Son, J., Ružić, N. & Philpott, A. (2025). Artificial intelligence technologies and applications for language learning and teaching. *Journal of China Computer-Assisted Language Learning*, *5*(1), 94-112. https://doi.org/10.1515/jccall-2023-0015

[4] Kano, K. & Takahama, Y. (2024, July 3). *Educators fear rise in AI-created essays as tools for detection lag*. The Asahi Shimbun. https://www.asahi.com/ajw/articles/15302691 (Accessed 27 July 2025).

[5] Dwyer, M, and Laird, E. (2023*).* Up in the air: Educators juggling the potential of generative AI with detection, discipline, and distrust. *Center for Democracy and Technology*. (2023, March). https://cdt.org/wp-content/uploads/2024/03/2024-03-21-CDT-Civic-Tech-Generative-AI-Survey-Research-final.pdf (Accessed 6 July 2025).

[6] Gustilo, L, Ong, E, and Lapinid, MR (2024). Algorithmically-driven writing and academic integrity: Exploring educators' practices, perceptions, and policies in the AI era. *International Journal for Educational Integrity, 20*(3). https://doi.org/10.1007/s40979-024-00153-8

[7] Dalalah, D. & Dalalah, OMA. (2023, July).The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT, *The International Journal of Management Education, 21*(2). https://doi.org/10.1016/j.ijme.2023.100822

[8] Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *Journal of Academic Integrity and Technology Ethics, 9*(2), 134–147. https://www.researchgate.net/publication/386998010

[9] Walters, W. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, *7*(1). https://doi.org/10.1515/opis-2022-0158

[10] Chaka, C. (2024). Accuracy pecking order: How 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning and Teaching,* *7*(1). https://doi.org/10.37074/jalt.2024.7.1.33

[11] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns,* *4*(8). https://doi.org/10.1016/j.patter.2023.100779

[12] Pratama, R. (2025). The accuracy–bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science, 11*, e2953. https://doi.org/10.7717/peerj-cs.2953

[13] Pudasaini, A., Zhang, Y., & Lee, J. (2025). *Benchmarking AI text detection: Assessing detectors against new datasets, evasion tactics, and enhanced LLMs*. Proceedings of the 2025 Conference on Generative AI Detection (GenAIDetect), 1(4): 45–62. https://aclanthology.org/2025.genaidetect-1.4/ (Accessed 15 August 2025).

[14] Lavergne, T. (2023, May 24). *AI detectors: Addressing the challenge of false positives*. Winston AI.https://gowinston.ai/ai-detectors-addressing-the-challenge-of-false-positives/ (Accessed 29 July 2025).

[15] Tian, E. (2023, October 24). *ESL bias in AI detection is an outdated narrative*. GPTZero. https://gptzero.me/news/esl-and-ai-detection (Accessed 6 September 2025).

[16] Emi, B., & Spero, M. (2024). *Technical report on the Checkfor.ai AI-generated text classifier* (Version 2) [Technical report]. arXiv. https://arxiv.org/abs/2402.14873v2 (Accessed 1 August 2025).

[17] González-Sendino, R., Serrano, E., Bajo, J., & Novais, P. (2024). A review of bias and fairness in artificial intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence,* *9*(1). https://doi.org/10.9781/ijimai.2023.11.001

[18] Woelfel, K. (2023, December 18). *Late applications: Disproportionate effects of generative AI detectors on English learners* [Policy brief]. Center for Democracy & Technology. https://cdt.org/insights/brief-late-applications-disproportionate-effects-of-generative-ai-detectors-on-english-learners/ (Accessed 16 July 2025).

[19] Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed., draft version). Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book_aug25.pdf (Accessed 29 July 2025).

[20] Chiusano, F. (2023, February 21). *Two minutes NLP: Perplexity explained with simple probabilities*. Medium. https://medium.com/nlplanet/two-minutes-nlp-perplexity-explained-with-simple-probabilities-6cdc46884584 (Accessed 20 August 2025).

[21] Colla, D., Delsanto, M., Agosto, M., Vitiello, B., & Radicioni, D. P. (2025). *Semantic coherence markers: The contribution of perplexity metrics* [Preprint]. University of Turin. https://iris.unito.it/bitstream/2318/1875282/1/colla2022semantic_preprint.pdf (Accessed 13 September 2025).

[22] Khan, A. (2023). *Mastering perplexity AI: A comprehensive guide to understanding and using perplexity in AI and NLP* [Kindle edition]. Amazon Digital Services LLC.

[23] Yeung, S. (2025, March 3). A comparative study of rule-based, machine learning, and large language model approaches in automated writing evaluation (AWE) (pp. 984-991). Lak`25: Proceedings of the 15th International Learning Analytics and Knowledge Conference. https://dl.acm.org/doi/proceedings/10.1145/3706468 (Accessed 15 August 2025).

[24] Gotoman, J. E. J., Luna, H.L.T., Sangria, J.C.S., Santiago, C.S., Barbuco, D. D. (2024). Accuracy and reliability of AI-generated text detection tools: A literature review. *American Journal of Interdisciplinary Research and Development,* *3*(2). https://journals.e-palli.com/home/index.php/ajirb/article/view/3795

[25] Price, G., & Sakellarios, M. D. (2023). The effectiveness of free software for detecting AI-generated writing. *International Journal of Teaching, Learning and Education, 2*(6), 31–38. https://doi.org/10.22161/ijtle.2.6.4

[26] Li, J., & Wan, X. (2025). *Who writes what: Unveiling the impact of author roles on AI-generated text detection*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 26620–26658). Vienna, Austria: Association for Computational Linguistics. https://aclanthology.org/2025.acl-long.1292/ (Accessed 30 July 2025).

[27] Buchert, J.-M. (2025, March 10). *The 6 best AI detectors based on objective studies & usage*. Intellectual Lead. https://intellectualead.com/best-ai-detectors-guide/. (Accessed 2 July 2025).

[28] Sallami, D., & Aïmeur, E. (2024). Fairframe: A fairness framework for bias detection and mitigation in news. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00568-6

[29] Bergsma, S., Post, M., & Yarowsky, D. (2012). Stylometric analysis of scientific articles. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 327–337. Association for Computational Linguistics.

https://aclanthology.org/N12-1033/ (Accessed 2 August 2025).

[30] Bellemy, R.K.E., Hind, M., Hoffman, S.C., Houde, S., & Kannan, K. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5). https://doi.org/10.1147/JRD.2019.2942287

[31] Ramzan, M., & Alahmadi, T. S. (2024). The impact of explicit syntax instruction on ESL learners' writing complexity. *World Journal of English Language, 14*(2), 25103. https://doi.org/10.5430/wjel.v14n2p25103

[32] André, Q. C., Ghosh, S., & Khatri, C. (2023). Detecting AI-generated abstracts using linguistic features: A comparative analysis. *CEUR Workshop Proceedings, 3551*, 18–25. https://ceur-ws.org/Vol-3551/paper3.pdf (Accessed 20 July 2025).

[33] Krig, S. (2016). *Computer vision metrics: Survey, taxonomy, and analysis*. Textbook edition. Springer

[34] Alexander, J., Alghamdi, A., & Alzahrani, M. (2023). ESL lecturers' perceptions of AI-generated writing: A deficit model in disguise? *Teaching English with Technology, 23*(2): 45–61. https://doi.org/10.56297/BUKA4060/XHLD5365

[35] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1). https://doi.org/10.1007/s40979-023-00146-z

[36] Nordstokke, D. W., & Zumbo, B. D. (2007). A cautionary tale about Levene's tests for equal variances. *Journal of Educational Research and Policy Studies, 7*(1): 1–14. https://files.eric.ed.gov/fulltext/EJ809430.pdf (Accessed 1 July 2025).

[37] Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys, 52*(6): 1–39. https://doi.org/10.1145/3345317

[38] Hind, M. (2019). *AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*. IBM Journal of Research and Development, 63(4/5). https://doi.org/10.1147/JRD.2019.2942287

[39] Neeley, T., & Englehart, T. (2025, January). AI vs human: Analyzing acceptable error rates using the confusion matrix. *Harvard Business School Technical Note* (No. 425-049). https://www.hbs.edu/faculty/Pages/item.aspx?num=66718 (Accessed 9 Sept. 2025).

[40] Padilla, R., Netto, S. L., and da Silva, E.A.B. A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Niteroi, Brazil, 2020, pp. 237-242. https://doi:10.1109/IWSSIP48289.2020.9145130

[41] Shan, G., & Gerstenberger, D. (2017). Fisher's exact approach for post hoc analysis of a chi-squared test. *PLOS ONE, 12*(12). https://doi.org/10.1371/journal.pone.0188709

[42] Sharpe, D. (2015). Your chi-square test is statistically significant: Now what? *Practical Assessment, Research & Evaluation, 20*(8), 1–10. https://doi.org/10.7275/tbfa-x148

[43] Samue, A. (2024, November 30). *7 best AI reference finder tools in 2025: A comprehensive review for researchers*. Tenorshare. https://ai.tenorshare.com/comparisons-and-reviews/ai-reference-finder.html (Accessed 16 August 2025).

[44] Kar S.K., Bansal T., Modi S., Singh A. (2024). How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. *Indian Journal of Psychological Medicine 47(*3): 275-278. https://doi:10.1177/02537176241247934

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

The Current Situation, Problems and Countermeasures of History Curriculum Settings in the Major of Ideological and Political Education

Ma Wenrui

College of Marxism, Foshan University（FOSU）

Received: 09 Sep 2025, Received in revised form: 06 Oct 2025, Accepted: 13 Oct 2025, Available online: 18 Oct 2025

Abstract

The major of Ideological and Political Education aims to cultivate versatile talents with good political theoretical literacy, ideological and moral qualities, and scientific and cultural qualities, shouldering the important mission of promoting the dissemination and practice of socialist core values. History courses play a foundational, supportive, and value-guiding role in this major. However, there are currently three major problems: the structural disconnection between history courses and ideological and political education, the hierarchical disconnection between general history teaching and thematic research, and the functional disconnection between the transmission of historical knowledge and the internalization of values. In response, this study, through historical literature review and logical argumentation, proposes that collaborative efforts in curriculum system restructuring, teaching method innovation, and value goal integration are needed. This will enhance students' in-depth historical understanding of Marxist theory and its achievements in the Chinese context, strengthen their ability to analyze real-world issues using the materialist conception of history, cultivate their "sense of history," solidify their political orientation, reinforce the "Four Confidences," and nurture ideological and political education talents for the new era.

***Keywords*— Ideological and Political Education Major; History Courses; Curriculum Reform; Socialist Core Values**

I. INTRODUCTION

The ideological and political education major (undergraduate) in Chinese universities has been established for 40 years. Its purpose is to cultivate versatile talents with solid political theoretical literacy, sound ideological and moral qualities, and scientific and cultural qualities, who are engaged in publicity, organization, management, and ideological and political work based on this major. At the undergraduate level, the focus is particularly on training teachers for the "Morality and Rule of Law" course in primary and secondary schools. Therefore, the ideological and political education major in universities is not only about producing qualified graduates but also bears the significant responsibility of promoting the dissemination and practice of socialist core values in the new era, as well as enhancing the moral standards and awareness of the rule of law among citizens. Xi Jinping

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

pointed out: "We must strengthen the study of history, deeply grasp historical laws, recognize historical trends, summarize historical experiences, and remember historical lessons. Through profound reflection on history, we can better carry out our current work and move toward the future." [1] As an important front for talent cultivation in the ideological field, the ideological and political education major should place great emphasis on fostering students' historical thinking, and offering corresponding history-related courses is a natural and essential part of this process.

II. THE INTRINSIC CONNECTION BETWEEN HISTORY AND IDEOLOGICAL AND POLITICAL EDUCATION

Marx pointed out in The German Ideology, "We know only a single science, the science of history." [2] Although this passage was crossed out by the author in the manuscript, it still reflects Marx's understanding and emphasis on the science of history. In fact, many theories and concepts in ideological and political education have deep historical roots. For example, the birth of the Marxist theoretical system emerged gradually during a specific historical period, alongside the economic development of capitalist society, the intensification of class contradictions, and the rise of the workers' movement. From the initial ideas of utopian socialism to Marx and Engels critically inheriting the theoretical achievements of German classical philosophy, British classical political economy, and French and British utopian socialism, and then establishing scientific socialism—this entire evolution was nurtured within the long river of history. Another example is the theoretical system of socialism with Chinese characteristics, which was gradually constructed through the historical practices of revolution, construction, and reform in modern and contemporary China by continuously summarizing experiences, learning lessons, and exploring innovations. As Engels said, "The theoretical thinking of every era, including that of our own, is a historical product." [3]

During his inspection of the former site of the Southwest Associated University in Yunnan, Xi Jinping proposed the concept of "cultivating a new generation with a sense of history," and the cultivation of this "sense of history" should first be reflected in the training of ideological and political education professionals. Generally speaking, the narratives of ideological and political education (including curriculum-based ideological and political education) mainly consist of two types: mechanism narratives centered on structure and temporal narratives based on events [4]. Therefore, adhering to the materialist conception of history and conducting education from a historical perspective and mindset, enabling students to utilize disciplinary historical resources to counteract historical nihilism [5], is a key focus in the training of ideological and political education professionals.

Returning to the professional development of ideological and political education, cultivating new talents with a "sense of history" and resisting historical nihilism primarily require students to master fundamental historical facts. On this foundation, by employing the analytical tools of historical materialism, students can clearly understand the background, developmental trajectory, and historical challenges faced by Marxism and the theory of socialism with Chinese characteristics. Only then can students grasp the essence and connotation of ideological and political education more profoundly and accurately, comprehending the roots of its scientific and revolutionary nature. Therefore, incorporating relevant historical courses is not only necessary but also highly meaningful. Such curriculum design helps students deepen their understanding of the historical development process, enhances their ability to analyze contemporary social phenomena and issues with greater depth and breadth, thereby laying a solid foundation for cultivating high-quality talents capable of accurately grasping the pulse of the times and guiding social trends.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

III. CURRENT SITUATION AND ISSUES IN THE CURRICULUM DESIGN OF HISTORY COURSES FOR IDEOLOGICAL AND POLITICAL EDUCATION MAJORS

According to the "National Standards for Teaching Quality of Undergraduate Programs in Regular Higher Education Institutions," the ideological and political education major currently does not require courses on general history (both Chinese and foreign) to be listed as elective courses [6]. As a result, only a very small number of higher education institutions, based on their own academic traditions, offer related general history courses (such as "Ancient Chinese History" and "Ancient World History"), like Foshan University. Most institutions, in accordance with the "National Standards for Teaching Quality of Undergraduate Programs in Regular Higher Education Institutions," offer thematic history courses. History courses are commonly offered in ideological and political education majors. However, from the three dimensions of history course design, history teaching methods, and the cultivation of historical awareness, there currently exist three disconnections in the teaching of history courses within the training of ideological and political education professionals:(a) The structural disconnection between history courses and ideological and political education; (b) The hierarchical disconnection between general history teaching and thematic research; (c) The functional disconnection between the transmission of historical knowledge and the internalization of values.

(1) The Structural Disconnection Between History Courses and Ideological and Political Education

The essence of structural disconnection lies in the fracture between knowledge supply and value objectives. In the history courses offered within the ideological and political education major, there often persists a history-centered teaching philosophy that emphasizes "empirical research" in instruction while lacking the "value guidance" stressed in ideological and political education. In other words, there is no effective integration of the methodologies of these two disciplines. Consequently, history courses exhibit superficiality in value guidance, failing to effectively utilize historical teaching to help students understand the intrinsic logical transformation from historical laws to political theories.

(2) The hierarchical disconnection between general history teaching and thematic research.

The hierarchical disconnect essentially reflects an imbalance between broad coverage and in-depth exploration. General history teaching often follows a timeline-based narrative, while thematic courses tend to fall into the trap of "fragmented textual research." Meanwhile, assessments in general history courses often emphasize memorization, whereas thematic courses lack training in critical analysis of historical sources. Together, these factors lead to a disconnect in students' skill development and a flattening of knowledge acquisition. For example, even among students who have systematically completed relevant course training, it is still common to equate "people-oriented thought" directly with "people-centeredness," while overlooking the fundamental differences between "feudal hierarchy" and "socialist democracy." This reveals that the current combination of general history and thematic courses in ideological and political education has failed to lay the foundation for critical thinking. Besides flaws in curriculum design, the lack of an interdisciplinary teacher collaborative lesson-planning mechanism also contributes to this phenomenon.

(3) The functional disconnection between the transmission of historical knowledge and the internalization of values.

Functional disconnection essentially refers to the alienation between cognitive input and emotional identification. Currently, the teaching practice of history courses in ideological and political education still predominantly follows the "lecture-memorization" model, with insufficient application of experiential teaching methods such as scenario simulations and oral history interviews. Moreover, in course assessments, the dimension of value internalization carries relatively low weight. These factors result in inefficient internalization of the value of historical knowledge among students,

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

and over-reliance on textbook language fails to effectively activate students' "sense of historical presence," leading to a lack of emotional resonance that historical knowledge should inherently evoke.

In summary, to change the current situation of "triple disconnection," it is necessary to make coordinated efforts in three aspects: restructuring the curriculum system (structural), innovating teaching methods (hierarchical), and integrating value objectives (functional).

IV. OPTIMIZATION PATHS FOR THE CURRICULUM DESIGN OF HISTORY COURSES IN IDEOLOGICAL AND POLITICAL EDUCATION MAJORS

The issue of the "three disconnects" in the teaching of history courses for ideological and political education majors is essentially a fracture in the systematic coordination among curriculum design, teaching implementation methods, and evaluation feedback mechanisms. To achieve synergistic efforts among these three aspects, it is necessary to break disciplinary barriers through structural integration, bridge the depth of knowledge through hierarchical progression, and connect value recognition through functional transformation.

(1) Construction of Interdisciplinary Course Clusters

Due to their different disciplinary affiliations, history and ideological and political education courses inevitably exhibit distinct training orientations in curriculum design, which is the direct cause of structural disconnection. The mechanical juxtaposition of different disciplinary systems inevitably leads to modular fragmentation. Currently, undergraduate history education places more emphasis on historical fact verification, while the training of ideological and political professionals requires a greater focus on value guidance. This results in long-term parallel states between different disciplines within the same professional training in terms of teaching objectives, knowledge frameworks, and teaching resources. Therefore, it is necessary to construct an interdisciplinary "history-politics" course cluster to improve teaching methods and address the issue of structural disconnection. The development of this course cluster should follow the path of "laying a foundation through general history—deepening through specialized topics."

Drawing on insights from relevant literature, this study proposes that history and ideological-political courses can be modularly integrated into two major course clusters: foundational general history and thematic deepening. Taking the curriculum design of the ideological-political education major at Foshan University as an example, the foundational general history module integrates historical spatiotemporal frameworks with the origins of ideological-political theories, while the thematic deepening module combines historical-political interdisciplinary topics with the cultivation of critical thinking (Table 1 and Table 2).

Table 1 Foundational Course Cluster for the History of Ideological and Political Education

Course Group	General History of China and the World	General Theory	Integrated Logic
Foundation of General History	Ancient Chinese History, Topics in World Ancient History, Modern Chinese History, Topics in World Modern History	Marxist philosophy, Marxist political economy, and the theory and practice of scientific socialism	Historical Spatiotemporal Framework and the Origins of Ideological and Political Theory

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

Table 2 Specialized Deepening Course Cluster for Ideological and Political Education Major

Course Group	General History of China and the World	General Theory	Integrated Logic
Topic Deepening	History of Modern Chinese Political Thought, History of the Chinese Communist Party, Contemporary Chinese Government and Politics, Modern Western Political Theory	Introduction to Chinese Philosophy, Topics in the History of Western Philosophy, Introduction to Historiography, Selected Readings from Marxist-Leninist Classics	The Deep Connection Between Historical Events and Political Theories

The general history module establishes a historical temporal-spatial framework through comprehensive history courses, revealing the laws of historical development from a materialist perspective via theoretical general education courses. Furthermore, it can merge "Ancient Chinese History" and "Topics in Ancient World History" into "An Outline of Chinese and Foreign Civilizations," reducing redundant content and emphasizing comparative studies of Chinese and Western cultures. The thematic deepening module places greater emphasis on the profound connection between historical facts and theories. For example, linking "The History of the Chinese Communist Party" with "The Theory of Marxism with Chinese Characteristics" to analyze how the "Agrarian Revolution" evolved from historical practice into the theory of "Armed Independent Regimes of Workers and Peasants." Another example is combining "Modern Western Political Theory" with "Topics in World Modern History" to critically compare "the global expansion history of liberal thought with the practical responses of the socialist movement."

The construction of course groups focuses on collaboration among teachers from different disciplines, requiring the establishment of a comprehensive collective lesson preparation system. This enables teachers from various subjects to effectively deconstruct and reorganize teaching content, achieving interdisciplinary synergy. Additionally, in the talent development plan, closely related courses should be scheduled in the same semester or consecutive semesters whenever possible.

(2) Innovation in Teaching Methods: Interdisciplinary Collaboration

In line with the new curriculum module design, teaching methods also need targeted innovation, establishing a new interdisciplinary teaching methodology system.

A. Historical Context Reality Comparison Method

The ultimate concern of history should be the actual social conditions of the present. Using historical events as a mirror, the goal is to restore context and create contemporary parallels, thereby activating students' "sense of historical presence" and transforming their understanding of historical patterns into the analysis of real-world issues. This should be one of the keys focuses on the training of students majoring in ideological and political education.

Scenario simulation and role-playing are the preferred methods for this teaching approach. Taking the course "History of the Chinese Communist Party" as an example, a simulated meeting on "Drafting the 1931 Constitution of the Chinese Soviet Republic" can be designed. Students can be required to explore the topic from historical, political, and value-based perspectives. From the **historical perspective**, students should reconstruct the historical context based on economic data from the agrarian revolution period (such as land distribution in rural southern Jiangxi), class conflicts (the opposition between peasants and landlords), and

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

the international environment (directives from the Comintern). From the **political perspective**, students can be divided into groups to role-play as "worker-peasant representatives," "intellectuals," and "military cadres," debating issues such as "whether to retain bourgeois voting rights in the constitution" to experience the logic behind the formation of the "democratic centralism" principle. From the **value perspective**, students can compare clauses on "people's democratic dictatorship" in the "Constitution of the Chinese Soviet Republic" with those in the contemporary constitution to understand the historical continuity of the Chinese Communist Party's political legitimacy.

This approach allows for the integration of interdisciplinary knowledge from courses like "Political Economy" and "Constitutional Law" in specialized deepening modules. Additionally, cross-temporal dialogues can be employed to connect general history content with thematic modules. For example, in the course "Contemporary Chinese Government and Politics," topics from "Modern and Contemporary World History" can be introduced, simulating a cross-temporal dialogue between "the governance structure of the Paris Commune" and "China's grassroots mass self-governance system" to analyze the historical evolution of proletarian political organization forms.

Tracing the roots of contemporary issues is also a crucial approach in teaching through historical context and present-day comparison. For example, in the course "Contemporary Social Issues in China," taking the "urban-rural dual structure" as a topic, one can integrate content from "Ancient Chinese History" and "Modern Chinese History" to trace back to historical milestones such as the "Single Whip Reform" tax policy of the Ming and Qing dynasties, the "Rural Reconstruction Movement" during the Republic of China era, and the "household registration system" in the early years of the People's Republic of China. Through long-term analysis, this method reveals the deep-seated historical logic behind structural contradictions, guiding students to propose governance solutions based on historical experiences (such as drawing wisdom from the ancient "Equal Field System" to optimize land transfer policies).

B. Interdisciplinary project-based learning

Interdisciplinary project-based teaching is centered around problem-driven pedagogy, using questions as the guiding force to integrate resources from disciplines such as history, politics, and philosophy. Through teamwork, it accomplishes research involving historical deconstruction, theoretical analysis, and value reconstruction, thereby achieving the goal of breaking down disciplinary barriers.

Project-based learning requires the coordination of different courses, with the same project theme being approached from different angles in various courses. For example, a project theme could be "The Historical Construction of the 'Grand Unification' Narrative and Contemporary State Governance—From the Qin Dynasty's Commandery-County System to 'One Country, Two Systems.'"

In terms of interdisciplinary collaboration, courses such as Ancient Chinese History, Contemporary Chinese Government and Politics, and Sinicized Marxist Theory can be integrated. The history course could focus on tracing the evolution of the centralized system from the Qin and Han dynasties to the Ming and Qing dynasties. The political science course might require an analysis of the legal foundations of the "unitary system" as a state structure. The Marxist theory course could ask students to explain how "One Country, Two Systems" represents an innovative development of Marxist state theory.

For implementation, students can be divided into different groups, such as a historical research group, a theoretical analysis group, and a value interpretation group. The historical research group could use texts like Records of the Grand Historian and Book of Han to reconstruct the challenges (e.g., resistance from the nobility of the six former states) and strategies in implementing the commandery-county system. The theoretical analysis group might compare the logic of federalism in The Federalist Papers with the discourse on central-local relations in On the Ten Major Relationships. The value interpretation group could

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

analyze cases like the return of Hong Kong and Macau to demonstrate how "One Country, Two Systems" embodies "the unity of principled firmness and strategic flexibility."

The final output could take the form of a jointly written research report, such as from 'Standardizing the Axle Width' to 'System Integration': A Study on the Historical Continuity of China's State Governance.

C. Deconstruction and Reconstruction – Critical Thinking Training

Through a three-stage training process involving historical evidence verification, ideological deconstruction, and value stance defense, students are equipped with the discursive ability to resist historical nihilism, fostering a cognitive logic of "facts - theory - values."

The first stage focuses on cultivating students' ability to critically analyze historical sources and discern multiple narratives. Taking the "Opium War" section from Modern Chinese History as an example, students compare the moral narrative of "barbarians violating authority" emphasized in Qing court memorials, the "clash of civilizations" discourse centered on "free trade obstruction" portrayed in British parliamentary archives, and analytical narratives examining the war's essence from a semi-colonial perspective. Students are required to closely read the texts and identify the underlying interests and ideological assumptions behind each narrative.

The second phase focuses on cultivating students' ability to analyze political discourse structures and engage in logical falsification. Taking the "Golden Decade of the Republic of China" section from Modern Chinese History as an example, students can be guided to conduct three levels of critique: empirical, theoretical, and value-based. For empirical critique, students can be tasked with examining historical data such as the income statistics of workers and peasants from 1927 to 1937 and maps showing the expansion of foreign concessions, exposing the class disparities beneath the "golden" facade. For theoretical critique, students can apply the analytical framework from Analysis of the Classes in Chinese Society to deconstruct the bourgeois standpoint embedded in the "Golden Decade" narrative. For value critique, students can compare the Communist Party of China's land revolution practice of "land to the tiller" during the same period, demonstrating the fundamental differences between the two paths.

The third stage cultivates students' ability to construct discourse from the value standpoint of socialism with Chinese characteristics. Taking the course "Introduction to Historical Classics" as an example, a comparative analysis can be made between "A Short Course on the History of the Communist Party of the Soviet Union (Bolsheviks)" and "A Brief History of the Communist Party of China," examining how the former deifies the authority of leaders through "historical determinism" and explaining how the latter narrates Party history with "people as the subject." Finally, students are required to write corresponding analytical reports, proposing discursive strategies to "resist historical nihilism."

In summary, the innovation in teaching models activates historical perception through contextual immersion, integrates knowledge systems via interdisciplinary practice, and solidifies value stances with critical study, transforming history courses into genuine "thought laboratories" for ideological education. This approach not only resolves the disconnection between history teaching and ideological education but also instills in the younger generation the profound spiritual drive to "draw lessons from history and create the future" through the cyclical process of deconstruction, reconstruction, and creation.

(3) Integration of Value Objectives

The integration of value objectives requires breaking through the current binary opposition in teaching between the "value neutrality" of historical studies and the "value precedence" of political science, in order to achieve a dialectical unity between historical laws and political values. By anchoring the value stance of historical materialism and guiding it with the core socialist values of Chinese characteristics, a progressive logic of "interpretation of historical laws—construction

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

of political legitimacy—generation of value identification" can be established, achieving a deep integration of historical cognition and political belief.

First, there is the valorization interpretation of historical laws, which requires efforts from two aspects. On one hand, it is necessary to transform causal necessity into value necessity, that is, to convert the objective laws of historical materialism into the value basis of political legitimacy. For example, when analyzing the failures of various classes in the historical process of "saving the nation from subjugation and ensuring its survival" in modern China, one can reveal the dual necessity of the proposition that "the leadership of the Communist Party of China is the choice of history and the people"—it is both the objective result of historical laws and the value expression of the people's subjectivity. Another example is when teaching the historical trend of the formation of the centralized system in the Qin and Han dynasties: while clarifying the causal necessity that the centralized system met the needs of the dispersed nature of small-scale peasant economy, one can further distill the Qin and Han concept of "great unity" into a geographical-cultural-political composite law for the survival of Chinese civilization and transform it into the contemporary value proposition that "national unity is the supreme interest of the nation." On the other hand, it is essential to dialectically combine historical critique with value defense. While deconstructing the historical phenomena of autocracy and exploitation in class societies, one should simultaneously construct the value coordinates of socialist fairness and justice. For instance, in analyzing the process of the agrarian revolution abolishing the feudal land ownership system, one should both expose the exploitative nature of landlord land ownership (historical critique) and highlight the political affirmation of labor value in "land to the tiller" (value defense).

The second aspect is the diachronic construction of political values. It is necessary to modernize and contemporarily reconstruct traditional values and revolutionary values (political legitimacy). When teaching traditional values, such as the people-oriented thought and the ideal of a harmonious world in ancient Chinese intellectual history, while tracing their historical evolution, one can further reveal their "elective affinity" with socialist core values and place them within the Marxist theoretical framework of "human emancipation," ultimately interpreting the civilizational continuity and contemporary innovation of the concept of "common prosperity." In teaching revolutionary values, a diachronic narrative can be employed to connect revolutionary values from different periods, such as linking the "anti-imperialism and anti-feudalism" of the New Democratic Revolution period, the "Four Modernizations" of the socialist construction period, and the "national rejuvenation" of the new era into a chain of value progression, demonstrating the self-improvement logic of the socialist value system with Chinese characteristics in historical practice. For example, when teaching the Yuan Dynasty's "provincial system" and the Qing Dynasty's "bureaucratization of native officers" policy, emphasis can be placed on explaining the gradual optimization logic of China's institutions in maintaining the unity of a multi-ethnic nation, providing historical legitimacy for the "one country, two systems" policy.

Finally, there is the generation of value consensus. This can be approached through historical sentiment and value critique. For instance, when teaching Ancient Chinese History, one can use material heritage such as the Great Wall and the Grand Canal as carriers to weave a narrative of engineering history—"the Qin built the Great Wall, the Sui dug the Grand Canal, and the Ming constructed the Forbidden City"—thereby reinforcing institutional consensus on "pooling resources to accomplish major undertakings." Similarly, when covering modern world history, contrasting The Communist Manifesto's critique of capitalist exploitation (e.g., "workers have no country") with Marx's praise for the "people's sovereignty" of the Paris Commune can construct an emotional chain of identification with proletarian liberation.

On the dimension of value critique, a three-tier

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

defense mechanism—"empirical verification, logical falsification, value critique"—can be established to guard against the erosion of erroneous value consensus (e.g., historical nihilism). For example, in discussing the "end of history" thesis in Contemporary World Economy and Politics, one can expose its logical flaws through empirical analysis of the Soviet Union's collapse while reaffirming the value of "socialist superiority" through the practical achievements of socialism with Chinese characteristics. Alternatively, when analyzing the historical cycle of "long division must unite," comparing the prosperity of the Wen-Jing era with the decline of the late Qing Dynasty highlights the governance innovation value of the Chinese Communist Party in breaking free from historical cyclicality.

The essence of value-goal integration lies in bridging the value pathway between "historical truth" and "political virtue," ensuring that history courses not only teach "what happened in the past" but also reveal "why such choices should be made." By interpreting historical laws as the foundation of political values and anchoring political values as the guiding principle of historical cognition, it ultimately constructs a multidimensional value identification system for socialism with Chinese characteristics across the temporal dimension of "history-reality-future" and the spatial dimension of "individual-nation-civilization." This integration avoids the value nihilism of historical relativism while transcending the value dogmatism of political doctrine. It not only resolves the functional disconnection between imparting historical knowledge and internalizing values but also lays a philosophical foundation for cultivating ideological and political talents with historical depth and political direction in the new era.

V. CONCLUSION

History courses play an irreplaceable foundational, supportive, and value-guiding role in the cultivation of ideological and political education professionals. Through an analysis of the current status of history course offerings in ideological and political education programs, this paper reveals three major disconnections. These disconnections hinder the full effectiveness of history courses in cultivating students' historical thinking, deepening theoretical understanding, solidifying value stances, and resisting historical nihilism.

This paper proposes a synergistic approach involving the restructuring of the curriculum system, the innovation of teaching methods, and the integration of value objectives to transform history courses into genuine "thought laboratories" for cultivating ideological and political education professionals. This approach not only enhances students' in-depth historical understanding of Marxist theory and its Sinicized and contemporary achievements but also strengthens their ability to analyze real-world issues using historical materialism. More importantly, it fosters a profound sense of "historical awareness," enabling students to solidify their political orientation through historical reflection and reinforcing their confidence in the path, theory, system, and culture of socialism with Chinese characteristics. Ultimately, this cultivates a new generation of ideological and political education professionals who are capable of shouldering the mission of national rejuvenation, possess deep historical insight, and demonstrate a strong sense of contemporary responsibility. Moving forward, it is essential to further test and refine these optimization strategies in practice, driving the reform of history courses in ideological and political education to deeper levels and better serving the fundamental task of fostering virtue through education.

REFERENCES

[1] Publicity Department of the Central Committee of the Communist Party of China. Outline of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era. Study Press, People's Publishing House, 2019: 245.

[2] Compilation and Translation Bureau of the Works of Marx, Engels, Lenin, and Stalin of the Central Committee of the Communist Party of China. Collected Works of Marx and Engels (Volume 1). People's Publishing House, 2009: 516-

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.6

519.

[3] Compilation and Translation Bureau of the Works of Marx, Engels, Lenin, and Stalin of the Central Committee of the Communist Party of China. Collected Works of Marx and Engels (Volume 9). People's Publishing House, 2009: 436.

[4] Zhong, Qidong. The Ideological and Political Education Narrative of Historical Materialism. Social Sciences of Beijing. 2023, 0(12): 18-28 https://doi.org/10.13262/j.bjsshkxy.bjshkx.231202

[5] You, Zhichun. “Sense of History”of Ideological and Political Education: Theoretical Connotation, Value Significance and Path to Improvement. Academic Exploration, 2023(10): 151-156.

[6] Teaching Guidance Committee for Higher Education Institutions of the Ministry of Education. National Standards for Teaching Quality of Undergraduate Majors in Regular Higher Education Institutions (Volume 1). Higher Education Press, 2018: 58.

International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-5, Sep-Oct 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Next-Gen Language Pedagogy: Leveraging Generative AI to Support Inclusive English Language Learning

Dr. M. Kannadhasan

Assistant Professor, Department of English, Thiruvalluvar University, Vellore, Tamil Nadu, India

Received: 14 Sep 2025, Received in revised form: 11 Oct 2025, Accepted: 16 Oct 2025, Available online: 19 Oct 2025

Abstract— *In an increasingly globalized and linguistically diverse world, English Language Teaching (ELT) must evolve to address the needs of all learners, including those from marginalized, multilingual, and neurodiverse backgrounds. The integration of Generative Artificial Intelligence (AI) into language pedagogy marks a significant paradigm shift toward inclusivity, personalization, and accessibility. This paper explores how generative AI technologies such as ChatGPT, Grammarly, ELSA Speak, and other adaptive platforms can support inclusive ELT practices by aligning with frameworks like Universal Design for Learning (UDL), translanguaging pedagogy, and culturally responsive teaching. This paper highlights how AI tools are democratizing English education, especially in resource-constrained environments. However, the paper also critically examines the ethical and pedagogical challenges associated with AI, including algorithmic bias, overreliance, privacy concerns, and the digital divide. By addressing both the affordances and limitations of generative AI, this study underscores the importance of thoughtful integration strategies grounded in equity and learner empowerment. The research ultimately advocates for a hybrid model of AI-enhanced pedagogy that preserves the irreplaceable role of human teachers while embracing the transformative potential of AI. The findings aim to contribute to ongoing discussions around sustainable, ethical, and inclusive practices in 21st-century English language education.*

Keywords*— *Generative AI, Inclusive Education, English Language Teaching, UDL, Multilingual Learners, Educational Technology, ChatGPT

The 21st century classroom is increasingly characterized by linguistic heterogeneity, neurodiversity, and multiculturalism. With these shifts, traditional "one-size-fits-all" methods in English Language Teaching (ELT) are no longer sufficient. The emergence of generative AI machine learning systems capable of producing human-like text, feedback, and responses offers promising avenues to reimagine English education in ways that are inclusive, dynamic, and learner-centered. Platforms like OpenAI's ChatGPT, Google Bard, and adaptive apps like Elsa Speak or Duolingo Max have begun to revolutionize language instruction, personalizing learning experiences based on each student's pace, proficiency, and cultural context. This paper explores how generative AI can support inclusive ELT practices while addressing the ethical and infrastructural challenges that accompany its implementation.

The digital revolution has dramatically reshaped educational paradigms across the globe, especially in the field of English Language Teaching (ELT). As English continues to function as a global lingua franca in academic, professional, and social domains, the demand for accessible, adaptive, and inclusive language instruction has intensified. Simultaneously, language classrooms have become more diverse than ever before—populated by learners from varied cultural, linguistic, cognitive, and socioeconomic backgrounds. These shifts present both a challenge and an opportunity for educators and institutions striving to ensure equitable access to English language learning.

In this dynamic context, the emergence of Generative Artificial Intelligence (AI), AI systems capable of generating human-like language, feedback, and content represents a groundbreaking development. Tools like OpenAI's ChatGPT, Google's Gemini, Grammarly, and ELSA Speak are transforming the landscape of language pedagogy by offering real-time assistance, context-sensitive support, and adaptive

learning experiences. These tools can not only personalize content delivery based on learners' proficiency levels and needs but also provide multilingual scaffolding, assistive features for learners with disabilities, and culturally relevant engagement strategies.

While the possibilities are expansive, they are not without complications. Generative AI raises critical questions about data privacy, algorithmic bias, technological dependency, and access inequality. Moreover, AI cannot replace the empathy, cultural sensitivity, and pedagogical judgment that human educators bring to the classroom. Therefore, this paper aims to investigate the role of generative AI in fostering inclusivity in ELT by evaluating its affordances, examining real-world case studies, and offering ethical, pedagogically grounded strategies for responsible integration

To anchor the study in sound pedagogical principles, this research draws upon Universal Design for Learning (UDL), a framework that emphasizes flexibility in the ways information is presented, learners express themselves, and students engage with content. Developed by CAST (2018), UDL advocates for inclusive practices that proactively address the needs of all learners, especially those with disabilities or non-normative cognitive styles. In the context of English Language Teaching (ELT), UDL promotes the integration of visual, auditory, and kinesthetic elements into lessons, allowing students to access content in diverse ways. Generative AI tools such as text-to-speech engines, speech-to-text applications, and AI-powered grammar feedback systems align with UDL by offering multiple entry points into language learning and reducing barriers to participation.

Translanguaging, as conceptualized by García and Wei (2014), challenges the notion of linguistic purity in language classrooms and embraces the dynamic interplay between a learner's full linguistic repertoire. In multilingual classrooms, especially in postcolonial or migrant contexts, students often draw upon their home languages to construct meaning and navigate new linguistic systems. Generative AI tools can support translanguaging by offering real-time translation, context-based code-switching, and feedback that respects linguistic hybridity. AI platforms with multilingual capabilities thus become instrumental in validating students' linguistic identities and providing culturally and linguistically responsive support throughout the learning process.

Culturally Responsive Pedagogy (CRP), as articulated by Geneva Gay (2010), advocates for teaching that affirms and reflects students' cultural knowledge, lived experiences, and worldviews. In ELT, this means moving beyond standardized content to include local narratives, idiomatic expressions, and culturally meaningful texts. Generative AI can contribute to CRP by generating prompts, examples, and stories tailored to students' cultural backgrounds. When carefully curated or customized, AI-generated materials can help students feel seen and respected in the curriculum, enhancing motivation and relevance. Thus, the thoughtful integration of AI within these three frameworks UDL, translanguaging, and CRP positions it as a powerful enabler of inclusive, learner-centered English language education.

Generative AI enables real-time customization of content based on learner profiles. Applications like ChatGPT can adjust reading complexity, grammar exercises, or vocabulary tasks to a learner's current proficiency level. For instance, a beginner-level student struggling with past tense verbs can receive instant, level-appropriate practice sentences with corrective feedback.

Generative AI systems can translate content, instructions, or feedback into a learner's native language, facilitating understanding while building confidence. Tools like DeepL, Google Translate APIs, and GPT-based multilingual chatbots support translanguaging pedagogy, especially in contexts where English is a second or third language.

AI-powered text-to-speech, speech recognition, and spelling correction tools assist learners with dyslexia, ADHD, or auditory processing challenges. Grammarly's tone detector and CoWriter's predictive text features, for example, allow neurodiverse students to participate in writing tasks more confidently and independently. Teachers can leverage AI systems to generate instant feedback on writing, pronunciation, or grammar, allowing them to focus on higher-order instruction and emotional support. Moreover, AI tutors are available 24/7, extending learning beyond classroom hours.

AI systems are trained on predominantly Western datasets, which may reflect implicit biases and underrepresent non-Western cultures or idioms. For example, learners from rural India or Africa may find that AI tools fail to recognize region-specific names, accents, or cultural contexts, potentially reinforcing linguistic imperialism. Students may become overly dependent on AI-generated corrections, undermining

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.7

deep language acquisition or critical thinking. The teacher's role in scaffolding and metacognitive reflection remains irreplaceable. Access to AI tools requires stable internet, smart devices, and digital literacy, resources often unavailable to learners in under-resourced schools or rural regions. Inclusion, therefore, also demands systemic investments in infrastructure and training. Many AI tools collect user data, raising concerns about surveillance and informed consent, especially when used by minors or in public education systems.

For generative AI to meaningfully contribute to inclusive English language learning, its implementation must be intentional, ethically grounded, and pedagogically informed. Technology alone does not guarantee equitable outcomes, it is the thoughtful integration of AI within teaching practices that shapes learner experiences. Effective strategies should center on empowering both educators and students while upholding core values such as equity, critical thinking, and cultural relevance.

A key principle in responsible integration is teacher mediation. Generative AI should act as a supportive tool, not a replacement for human instruction. Teachers play a crucial role in interpreting AI-generated content, correcting inaccuracies, and helping learners make sense of the feedback they receive. For instance, a chatbot may provide a grammatically correct sentence, but it might lack cultural nuance or contextual appropriateness. Teachers must guide students in reflecting on AI outputs, making informed choices, and understanding language beyond surface-level correctness.

Another essential strategy involves promoting student agency. Learners should be encouraged to engage critically with AI tools questioning suggestions, exploring multiple responses, and using AI as a creative collaborator rather than a passive answer generator. Assignments that require learners to compare AI-generated outputs with their own writing or to reflect on AI feedback help build metacognitive awareness. Such activities foster autonomy, creativity, and deeper engagement with language, moving beyond rote learning toward active knowledge construction.

Curriculum design should incorporate generative AI in ways that reflect learners' identities, lived experiences, and sociocultural realities. This includes creating tasks that ask students to write AI-assisted narratives based on local events, translate traditional proverbs, or analyze AI-generated texts for cultural relevance. By embedding AI within a culturally responsive and contextually grounded curriculum, educators ensure that the technology supports rather than erases linguistic and cultural diversity.

To implement these strategies effectively, professional development is essential. Teachers need structured training in AI literacy, including how to evaluate AI tools, recognize biases in datasets, and use generative models ethically. Workshops, peer mentoring, and collaborative lesson planning can help educators develop confidence and creativity in incorporating AI meaningfully into their pedagogy. Equipping teachers with these skills is especially critical in underserved schools, where technological unfamiliarity can widen the equity gap. Community and policy engagement must accompany classroom-level interventions. Efforts to promote inclusive AI use in ELT should be supported by broader educational reforms, including national curriculum guidelines, investment in digital infrastructure, and clear policies for data privacy and AI ethics. Stakeholder collaboration, between educators, policymakers, parents, and tech developers is crucial to ensure that AI integration aligns with equity goals and meets the diverse needs of learners at scale. The global landscape of education reflects an increasing interest in the integration of generative AI to bridge gaps in English language instruction. Several pioneering initiatives and platforms provide practical insight into how AI can be deployed to foster inclusion and linguistic diversity.

India's Ministry of Education launched the DIKSHA (Digital Infrastructure for Knowledge Sharing) platform as part of its National Digital Education Architecture (NDEAR). DIKSHA incorporates AI-driven tools for personalized learning and has been exploring the inclusion of AI-based English language modules, particularly in collaboration with regional language interfaces. These initiatives allow students from underserved rural and tribal communities to access English instruction with multilingual support in Hindi, Tamil, Bengali, Marathi, and other languages, aligning with translanguaging pedagogy.

In regions where access to quality English instruction is limited, generative AI embedded within such platforms enables learners to receive real-time feedback, adaptive assessments, and voice-guided reading assistance. The tools not only cater to linguistic diversity but also help teachers personalize learning for large classrooms with varied proficiency levels. ELSA (English Language Speech Assistant) is a generative AI-powered app developed to improve pronunciation through real-time voice recognition. Widely adopted in

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.7

countries like Vietnam, Brazil, and Indonesia, ELSA helps learners with varying accents and local dialects achieve clearer pronunciation in English. The app uses machine learning algorithms to detect phonetic errors and provides corrective exercises based on native-like speech models.

ELSA's success lies in its inclusive design, it accommodates different English varieties and encourages self-paced learning, making it particularly valuable in low-resource or remote settings where qualified English phonetics instructors are scarce. By incorporating gamification, cultural scenarios, and voice repetition drills, ELSA promotes engagement among younger learners and non-traditional adult learners alike.

In university settings, instructors are experimenting with tools like ChatGPT and GrammarlyGO to aid academic writing among ESL students. At institutions in South Korea, the Philippines, and parts of Eastern Europe, educators report that AI prompts help learners brainstorm ideas, refine grammar, and engage in peer-editing simulations.

A pilot study at a university in the Philippines (Flores, 2024) showed that when paired with reflective activities, ChatGPT encouraged students to improve their vocabulary and self-edit their essays before submission. Moreover, learners with writing anxiety reported increased confidence and autonomy when receiving private AI feedback before public critique. These implementations emphasize AI as a co-writer, not a replacement for human creativity. They also show that guided use of generative AI enhances learner agency and reduces linguistic exclusion in academic discourse communities.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) has launched multiple initiatives to ensure equitable AI integration in education. Its 2021 report, AI and Education: Guidance for Policymakers, outlines recommendations for using AI to advance Sustainable Development Goal 4 inclusive and equitable quality education for all. UNESCO-supported projects in Sub-Saharan Africa and Southeast Asia include piloting AI chatbots in English learning for out-of-school youth, AI-powered reading companions for early literacy, and low-bandwidth AI tools for rural educators. These initiatives foreground ethical concerns, cultural contextualization, and gender sensitivity, especially in underrepresented learner communities.

The future of generative AI in inclusive English Language Teaching (ELT) lies in the collaborative efforts of technologists, educators, linguists, and policymakers. For AI to achieve its full potential in supporting diverse learners, its development must go beyond generic solutions and move toward culturally and contextually relevant applications. One pressing need is the creation of region-specific AI corpora that represent local dialects, multilingual expressions, and cultural narratives. This will help reduce algorithmic bias and allow AI tools to reflect the linguistic realities of learners in diverse geographies, particularly in the Global South. Another promising avenue is the development and dissemination of open-source generative AI tools specifically designed for public education systems, low-resource schools, and nonprofit educational organizations. Commercial AI platforms often remain inaccessible to underserved communities due to subscription costs, data privacy concerns, or a lack of customization. Open-access, community-driven tools can bridge the digital divide and democratize AI-enhanced ELT, especially when supported by multilingual interfaces and offline functionalities. Equally crucial is the integration of ethical AI education into teacher training programs. As AI becomes a staple in digital pedagogy, teachers must be equipped not only with technical know-how but also with ethical frameworks to evaluate AI's role in their classrooms. Including AI ethics, data privacy, and critical digital literacy in teacher education will ensure that educators act as informed mediators, capable of guiding students in reflective and responsible AI use.

Finally, there is a growing need for longitudinal research on the effects of generative AI on language acquisition. While short-term gains in fluency and writing accuracy have been observed, little is known about how AI impacts deeper cognitive skills such as creativity, critical thinking, and long-term retention. Future studies should assess how AI influences learner identity, intercultural competence, and the development of communicative autonomy over time.

With equitable infrastructure, transparent data policies, and a commitment to inclusive and human-centered design, generative AI can become a transformative tool in the pursuit of global language equity. By shaping policies, technologies, and pedagogical practices in tandem, we can harness AI not as a substitute for educators, but as a partner in delivering meaningful, just, and accessible English education for all.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.7

REFERENCES

[1] CAST. *Universal Design for Learning Guidelines version 2.2*. 2018. www.cast.org

[2] García, Ofelia, and Li Wei. *Translanguaging: Language, Bilingualism and Education*. Palgrave Macmillan, 2014.

[3] Gay, Geneva. *Culturally Responsive Teaching: Theory, Research, and Practice*. Teachers College Press, 2010.

[4] Luckin, Rose, et al. *Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations*. Department for Education (UK), 2023.

[5] Selwyn, Neil. "Should Robots Replace Teachers?" *British Journal of Educational Technology*, vol. 52, no. 3, 2021, pp. 1205–1219.

[6] UNESCO. *AI and Education: Guidance for Policy-makers*. 2021. unesdoc.unesco.org

[7] Warschauer, Mark, and James Gee. "AI for Language Learning: Beyond Grammar and Vocabulary." *TESOL Quarterly*, vol. 57, no. 1, 2023, pp. 19–33.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.5.7