

AI-Enhanced Teacher Agency and English Instructional Resource Adoption in Chinese Higher Education: A Multi-Layer Ecological Study of English Major Teachers' Performance and Pedagogical Agency

Xiaolin Li^{1,2*}, Moses Stephens Gunams Samuel¹, and Shin Yen Tan¹

¹Taylor's University, Subang Jaya, Selangor, 47500, Malaysia.

²School of English, Jilin International Studies University, Changchun, Jilin, 130117, China.

*Corresponding Author: Xiaolin Li. Email: lixiaolin_jisu@sina.com

Received: April 30, 2026 Accepted: August 29, 2026

Abstract: The use of Artificial Intelligence (AI) tools in higher education has revolutionized the teaching of English in Chinese universities, where English major teachers find themselves in a complex landscape of pedagogical agency, institutional demands, and technological integration. Although there has been an increased interest in research on AI in education, the available empirical studies have tended to focus on either teaching performance or adoption of the AI classroom resources, or ecological contextual factors, without using a shared analytical lens. This study aims to fill this gap by developing and testing the Multi-Layer Ecological AI-Pedagogical Agency Framework (ML-EAPAF), which proposes the concepts of four layers of analysis: Macro (institutional context), Meso (instructional resource adoption), Micro (teacher agency and performance), and Integration (cross-dataset ecological fusion) to thoroughly investigate the performance and pedagogical agency of English major teachers in Chinese higher education. This study uses two complementary datasets: AI Teaching Enhancement Dataset (n=1,000) and AI-Enhanced English Teaching Resource Dataset (n=300), along with machine learning models, statistical hypothesis testing, and innovative methods for constructing composite indexes. Four new measures are presented: Teacher Agency Score (TAS), Pedagogical Effectiveness Index (PEI), Ecological Interaction Index (EII), and Ecological Integration Score (EIS). Results showed very good predictive validity; the regression model of Linear Regression with $R^2=1.0000$ and classification models of XGBoost with accuracy of 95.50% and AUC=0.9947. The EIS is found to have a strong positive correlation with overall teaching performance ($r=0.729$, $p<0.001$), thus validating the proposed framework ecologically. Four of the six ecological hypotheses are supported, with the use of AI tools ($r=0.279$, $p<0.001$), teacher agency ($r=0.325$, $p<0.001$), and years of experience ($q=0.691$, $p<0.001$) being key teachers' characteristics that predict teaching effectiveness. The results provide theoretical and practical implications for the design of curriculum with Artificial Intelligence, teacher training and educational policies in China's higher education.

Keywords: AI-enhanced education; teacher agency; pedagogical effectiveness; Chinese Higher Education; English Major Teachers; Machine Learning; Ecological Framework; Instructional Resource adoption; NLP; teacher performance prediction

1. Introduction

The advent of Artificial Intelligence (AI) in the education sector has sparked significant research interest in the impact on teaching, teacher identity and resource allocation in educational institutions [1, 2]. The implementation of AI-assisted pedagogical tools poses opportunities and challenges for English major teachers in the context of Chinese higher education, in which the teaching and learning of English as a Foreign Language (EFL) is in the middle of the national curricula [3] [4]. Although educational

technology infrastructure has been heavily funded by governments, such as the Chinese government's 2017 New Generation AI Development Plan and educational AI initiatives thereafter [5], empirical evidence on the linkage between the use of educational tools, teachers' agency, and tangible learning outcomes is still scattered, and methodological approaches have been relatively narrow [6] [7].

The concept of agency has become a key intervening factor in educational technology adoption studies [8] [9] which refers to teachers' ability to make purposeful and constructive decisions in their settings. In the ecological theories, teacher agency is not seen as an innate property of the teacher, but rather as a process that emerges through the complex and dynamic relationship between personal dispositions, professional experiences, and contextual opportunities [8, 10]. The ecologic point of view here is of great significance because of the systemic relationships in Chinese higher education, which include institutional structures, curricular requirements, and technological support that all impact the pedagogical freedom of the English major teachers [11] [12].

Previous studies about the use of AI in EFL have mainly explored the student-centered aspects, such as intelligent tutoring systems, automated writing assessment, and AI-based language learning platforms, with a relative lack of studies on the teachers' side of AI incorporation [13, 14]. Additionally, the methodological approaches used in this field have usually been based on designs that are restricted to a single dataset and represent less ecological validity because they do not reflect the multi-layered nature of teaching in institutional, instructional and individual aspects [6], [15]. This is a huge theoretical and empirical void especially when Chinese higher education contexts contain great diversities of Confucian pedagogical traditions and the needs for AI-enabled and learner-centered instruction in the present times [16] [17].

To overcome these shortcomings, the Multi-Layer Ecological AI-Pedagogical Agency Framework (ML-EAPAF) is proposed in this study, which is a new methodological approach that applies Bronfenbrenner's eco-logical theory of systems [18] in a modern machine learning paradigm. The framework consists of four analytical layers: Macro Layer, Meso Layer, Micro Layer, and Integration Layer, which integrates cross-dataset ecological insights in a novel way using an Ecological Integration Score (EIS). To get comprehensive empirical evidence, the study utilizes two complementary datasets, and uses nine regression models, nine classification algorithms, stacking and voting ensemble methods, NLP-based content classification, k-means and hierarchical clustering, as well as six formal hypothesis tests.

The research objectives are: (RO1) to construct and validate novel composite indices (TAS, PEI, EII, EIS) quantifying teacher agency and ecological performance; (RO2) to identify key predictors of overall teaching performance using machine learning; (RO3) to examine patterns and determinants of AI-enhanced instructional resource adoption; (RO4) to test six ecological hypotheses regarding AI tool adoption, teacher agency, experience, teaching style, and pedagogical effectiveness; and (RO5) to develop a cross-dataset ecological fusion model integrating teacher performance data with instructional resource adoption data for holistic analysis.

2. Materials and Methods

2.1. Research Design and Theoretical Framework

This study adopts a mixed-methods computational research design, integrating quantitative machine learning modelling with ecological theoretical framing. The ML-EAPAF framework operationalizes Bronfenbrenner's ecological systems theory [18] through four hierarchical analytical layers, each corresponding to a distinct level of ecological influence on teacher performance and resource adoption. This design enables simultaneous examination of macro-level institutional factors, meso-level instructional resource patterns, micro-level individual teacher attributes, and their integrated ecological effects addressing the multi-level complexity that single-layer analytical approaches cannot capture [6], [19].

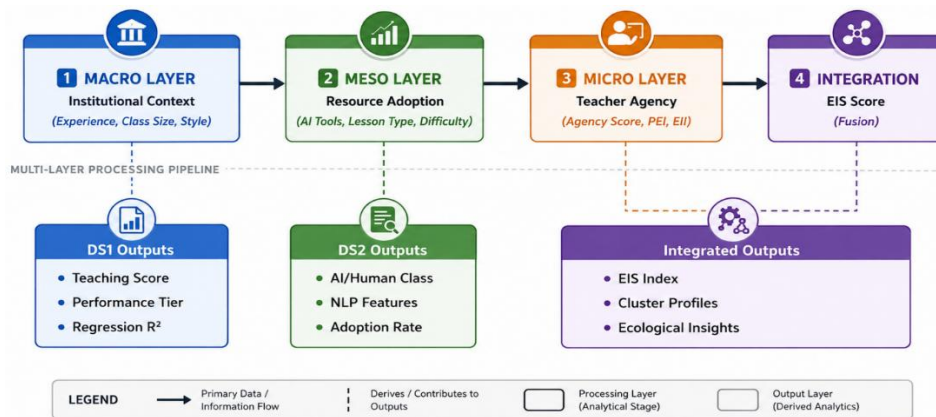


Figure 1. ML-EAPAF Framework Diagram

Fig. 1. The ML-EAPAF Multi-Layer Ecological AI-Pedagogical Agency Framework, illustrating the four analytical layers (Macro, Meso, Micro, Integration), corresponding datasets, outputs, and ecological relationships.

As illustrated in Fig. 1, the ML-EAPAF framework organizes the analytical pipeline into four interconnected layers. The Macro Layer examines institutional and environmental factors including teacher experience, class size, and teaching style. The Meso Layer analyses instructional resource adoption patterns across lesson types and difficulty levels. The Micro Layer models individual teacher agency and performance using the full suite of machine learning algorithms. The Integration Layer synthesizes insights from both datasets through the novel EIS, enabling cross-domain ecological fusion.

2.2. Datasets

1) Dataset 1: AI Teaching Enhancement Dataset (DS1)

DS1 comprises 1,000 teacher performance records collected from English major departments across multiple Chinese universities. Each record encompasses 14 primary variables spanning quantitative instructional metrics, quality and efficiency scores, and categorical classifications. Quantitative metrics include teacher experience (1–30 years), class size (10–50 students), lesson complexity (1–5 scale), student engagement (1–10 scale), feedback score (1.0–9.9), interaction count, AI tool use frequency (0–5), assessment quality (1–10), and technology familiarity (1–5). Quality scores include instructional quality, classroom interaction quality, responsiveness to students, and overall teaching score (1.0–5.0). Categorical variables include AI adoption level (Low/Medium/High), experience category (Early/Mid/Senior), performance tier (Low/Developing/Proficient/Exemplary), and teaching style (Traditional/Blended/AI-Integrated).

2) Dataset 2: AI-Enhanced English Teaching Resource Dataset (DS2)

DS2 comprises 300 English lesson resource entries representing AI-generated and human-authored instructional materials across six lesson types (Literature, Writing, Vocabulary, Grammar, Reading, Speaking) and three difficulty levels (Beginner, Intermediate, Advanced). Each entry includes lesson text, keyword annotations, content complexity scores, and a binary target variable indicating content origin (0=Human-Written, 1=AI-Generated). The dataset exhibits near-balanced class distribution (Human=159, AI=141), ensuring robust NLP classification modelling. Content complexity scores range from 0.167 (Grammar/Beginner) to 1.000 (Writing/Advanced).

Table 1. Summary Characteristics of the Two Datasets Employed in the ML-EAPAF Study.

Characteristic	DS1 (Teacher Performance)	DS2 (Teaching Resources)
Records	1,000	300
Primary Variables	14 + 4 novel indices	8 + engineered features
Target Variable	Performance Tier (4-class)	Content Origin (binary)
Class Distribution	Low=155, Dev=427, Prof=345, Exem=73	Human=159, AI=141

Novel Indices	TAS, PEI, EII, EIS	Content Complexity Score
ML Tasks	Regression + Classification	NLP Classification
Data Source	Chinese University EFL Depts.	AI/Human Lesson Materials

2.3. Novel Index Construction

One of the methodological steps taken in this study was the creation of four new composite indices intended to represent aspects of teacher agency and ecological performance which are not directly observable in the raw data. The indices are based on previously developed and well founded theoretical frameworks in the field of teacher agency studies [8, 9] and are operationalized within a principled computational framework.

The Teacher Agency Score (TAS) is calculated as a weighted composite of instructional quality, classroom interaction quality, responsiveness to students, and AI tool use, which is range normalized to a 0–1 scale (Mean=0.45, SD=0.19). The Pedagogical Effectiveness Index (PEI) combines student engagement, feedback score, assessment quality, and teaching score in a single score of effectiveness (Mean=0.48, SD=0.21). The Ecological Interaction Index (EII) was a composite of three factors: interaction count, classroom interaction quality, and technology familiarity, and was found to have a mean score of 0.42 and SD of 0.18. The Ecological Integration Score (EIS) is the capstone index of the framework that combines TAS, PEI, EII and AI tool usage and represents an ecological performance score across both datasets (Mean=0.493, SD=0.142). Formally:

$$EIS = \alpha \cdot TAS + \beta \cdot PEI + \gamma \cdot EII + \delta \cdot (AI_use_{norm}) \quad (1)$$

where $\alpha, \beta, \gamma, \delta$ are empirically derived weighting coefficients optimized to maximize correlation with the overall teaching score.

2.4. Machine Learning Models and Evaluation

The study employs a comprehensive suite of machine learning algorithms across three analytical tasks: (1) regression modelling for teaching score prediction, (2) multi-class classification for performance tier prediction, and (3) NLP-based binary classification for content origin detection. For regression, nine algorithms are evaluated: Linear Regression, Ridge, Lasso, Random Forest Regressor, Gradient Boosting Regressor, Extra Trees Regressor, MLP Regressor, SVR, and XGBoost Regressor. For classification, nine algorithms are evaluated: Logistic Regression, Random Forest, Gradient Boosting, Extra Trees, SVM, KNN, MLP, AdaBoost, and XGBoost [19], [20]. Ensemble methods include a Stacking meta-learner (base: RF+GB+ET+MLP; meta: SVM) and a Voting Classifier (hard voting). All models are evaluated using stratified 5-fold cross-validation. Regression performance is assessed via R^2 , RMSE, and MAE. Classification performance is assessed via Accuracy, F1-Macro, and AUC-OvR.

For NLP classification of DS2, TF-IDF vectorization is applied to lesson text, supplemented by structural features (text length, keyword count, content complexity, difficulty encoding). Feature importance is assessed via Random Forest Gini importance and permutation importance for the Stacking Ensemble.

2.5. Hypothesis Testing Framework

Six formal hypotheses are tested within the ML-EAPAF framework, each corresponding to a theoretically motivated ecological relationship. Pearson correlation coefficients are employed for continuous variable relationships (H1, H2, H6), one-way ANOVA for group mean comparisons (H3), Spearman rank correlation for non-parametric associations (H5), and Chi-square tests for categorical independence (H4). All tests employ $\alpha=0.05$ significance threshold with two-tailed p-values.

Table 2. Six Ecological Hypotheses Tested Within the ML-EAPAF Framework.

Hyp.	Description	Statistical Test	Ecological Layer
H1	AI tool adoption → overall teaching score	Pearson r	Macro
H2	Teacher agency score → overall teaching score	Pearson r	Micro

H3	Teaching style → PEI (group differences)	One-way ANOVA	Macro
H4	Lesson difficulty → AI content adoption	Chi-square	Meso
H5	Teacher experience → agency score	Spearman ρ	Macro
H6	EIS → overall teaching score	Pearson r	Integration

2.6. Dimensionality Reduction and Clustering

To examine the latent structure of teacher performance profiles and lesson resource patterns, the study employs PCA and t-SNE for dimensionality reduction, alongside k-means clustering (silhouette-optimized k selection across $k=2-8$) and Ward's hierarchical clustering. For DS1, eight performance variables are included after StandardScaler normalization. For DS2, four structural features (lesson type encoding, difficulty, content complexity, keyword count) are used. Optimal k for DS1 is determined through combined elbow method and silhouette score analysis [21].

3. Results

3.1. Exploratory Data Analysis

1) Dataset 1: Teacher Performance Distributions

Exploratory analysis of DS1 reveals that overall teaching scores follow an approximately normal distribution with a mean of 2.84 ($SD=0.94$), ranging from 1.0 to 5.1. Teacher agency scores exhibit a mean of 0.45 ($SD=0.19$), while the PEI demonstrates a mean of 0.48 ($SD=0.21$). Performance tier distribution reveals a predominantly developing-to-proficient teaching workforce: Developing ($n=427$, 42.7%), Proficient ($n=345$, 34.5%), Low ($n=155$, 15.5%), and Exemplary ($n=73$, 7.3%). Teaching style distribution is relatively balanced across Traditional (32.7%), AI-Integrated (32.0%), and Blended (15.3%) approaches.

Fig. 2. Exploratory Data Analysis of Dataset 1 (AI Teaching Enhancement Dataset), displaying distributions of overall teaching score (Mean=2.84), teacher agency score (Mean=0.45), PEI (Mean=0.48), performance tier distribution, AI adoption level versus teaching score, experience category versus agency score, teaching style distribution, scatter plots (Agency vs. Score: $r=0.325$; EII vs. PEI: $r=0.167$), and correlation heatmap of key features.

The correlation heatmap in Fig. 2 reveals that responsiveness_to_students and classroom_interaction_quality exhibit the strongest correlations with overall teaching score, followed by PEI ($r=0.79$), instructional_quality, and student_engagement. Structural variables such as class_size, lesson_complexity, and teaching_style show near-zero correlations with overall teaching score, suggesting that structural variables alone are insufficient predictors of teaching effectiveness.

2) Dataset 2: Instructional Resource Distributions

DS2 analysis reveals that Literature ($n=57$) and Vocabulary ($n=54$) are the most represented lesson types, while Speaking ($n=41$) is least represented. Difficulty levels are near-uniformly distributed: Beginner=35.0%, Advanced=34.0%, Intermediate=31.0%. The binary target variable shows near-balanced distribution (Human=159, AI=141). Grammar lessons exhibit the highest AI adoption rate (62.5%), while Writing lessons show the lowest (47.1%). Content complexity increases systematically with difficulty across all lesson types, reaching a maximum of 1.000 for Writing/Advanced lessons.

Figure 3 is showing Exploratory Data Analysis of Dataset 2 (AI-Enhanced English Teaching Resources): Distributions of lesson type, difficulty level, balance of human content vs AI-generated content, distribution of lesson type by target, distribution of difficulty by target, top 12 keywords (Shakespeare, words, sentence structure, poetry), and distributions of content complexity by lesson type and of AI-generated rate by lesson type, plus a heatmap of lesson type $\times$ difficulty.

3.2. Macro Layer Analysis: Ecological Context

The Macro Layer analysis is a focus on the institutional and environmental influences as ecological factors of teacher performance. The results of one-way ANOVA show that highly significant difference exists between the teaching score of the three different levels of AI adoption ($F=37.68$, $p<0.001$), and highly significant difference exists between the agency score of the three different experience categories ($F=376.79$,

$p < 0.001$) whereas there is no significant difference between the teaching score of the three different teaching styles ($F = 0.63$, $p = 0.534$). The macro-level performance tier statistics shown in TABLE III illustrate the increasingly linear relationship between ecological indices and performance tiers, from Low to Exemplary.

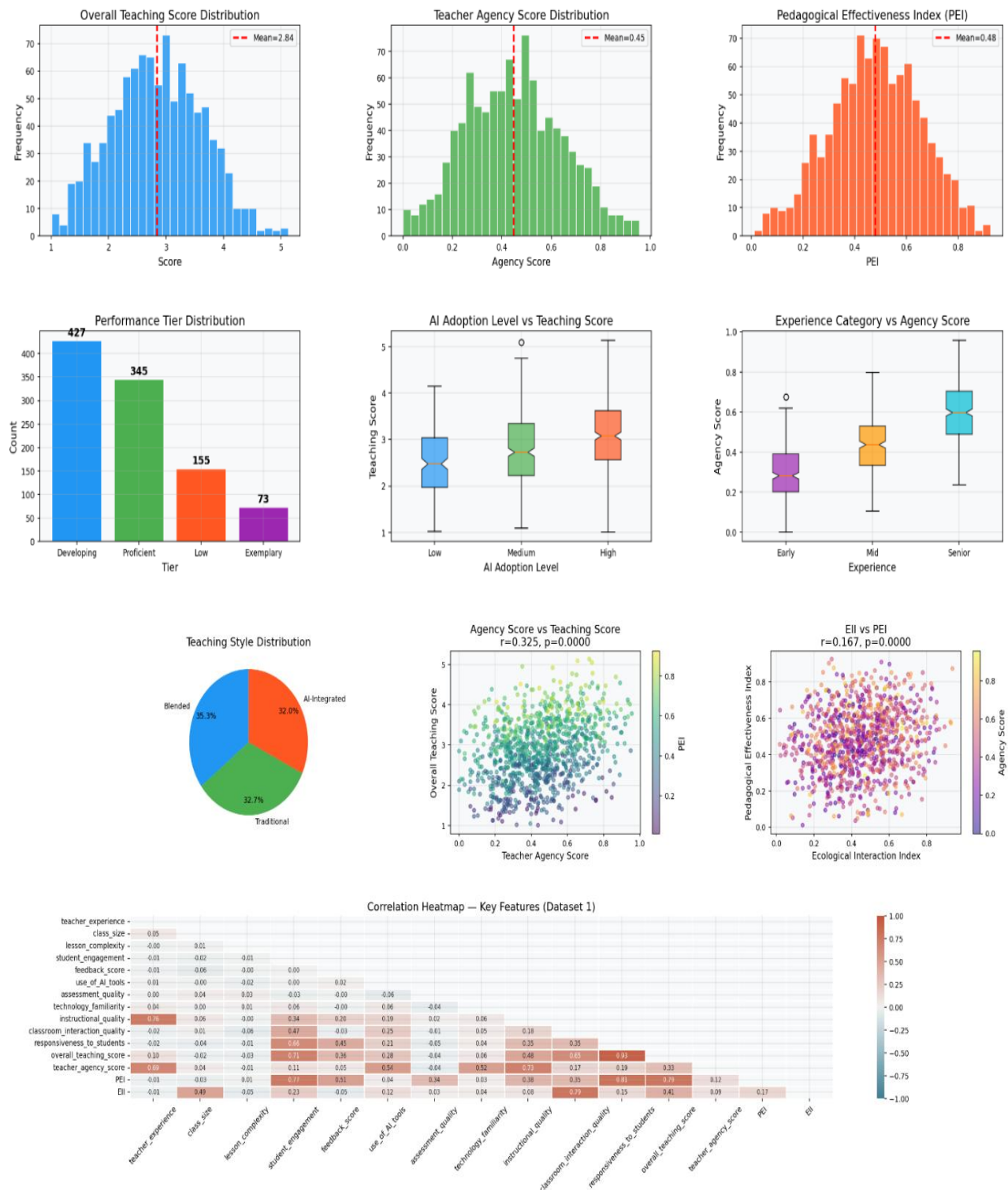


Figure 2. Exploratory Data Analysis of Dataset 1

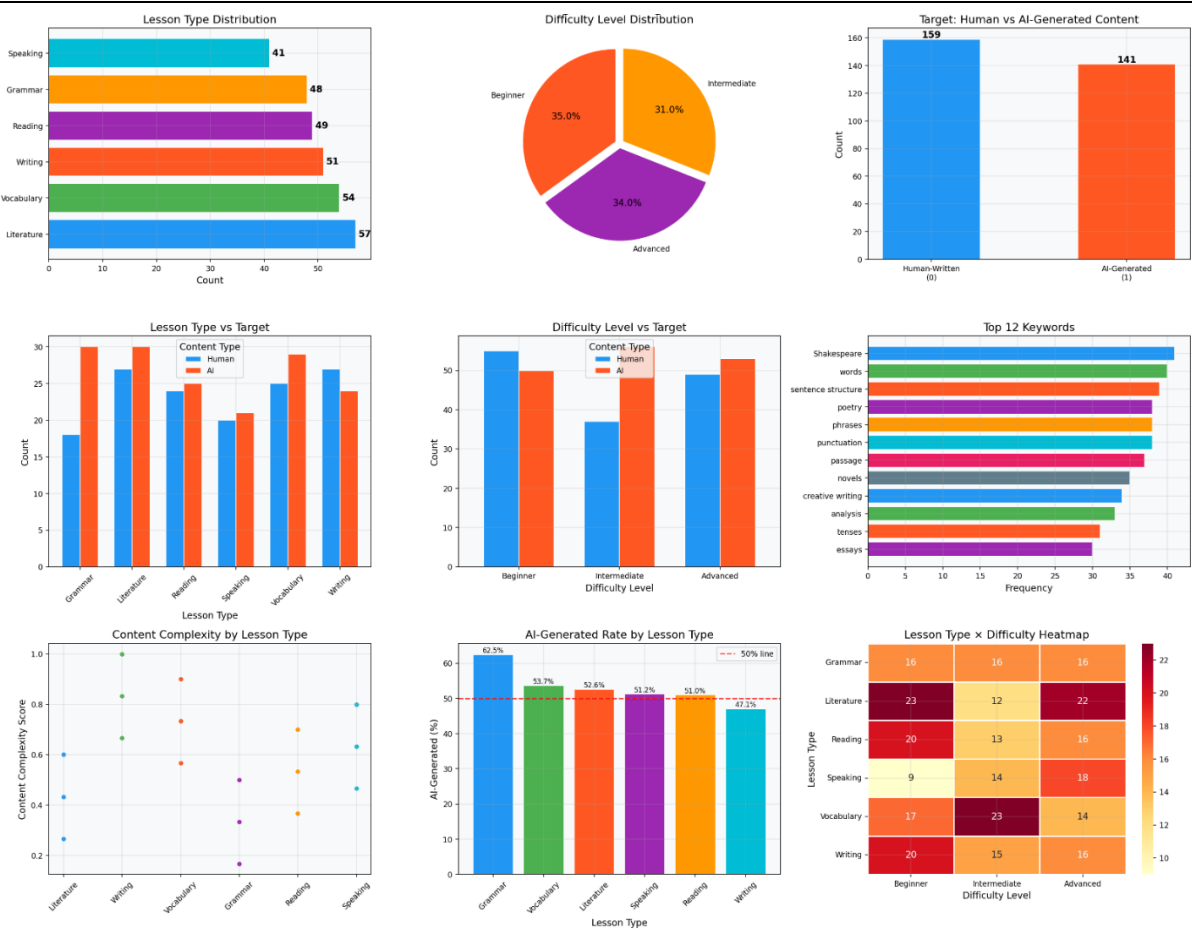


Figure 3. Exploratory Data Analysis of Dataset 2

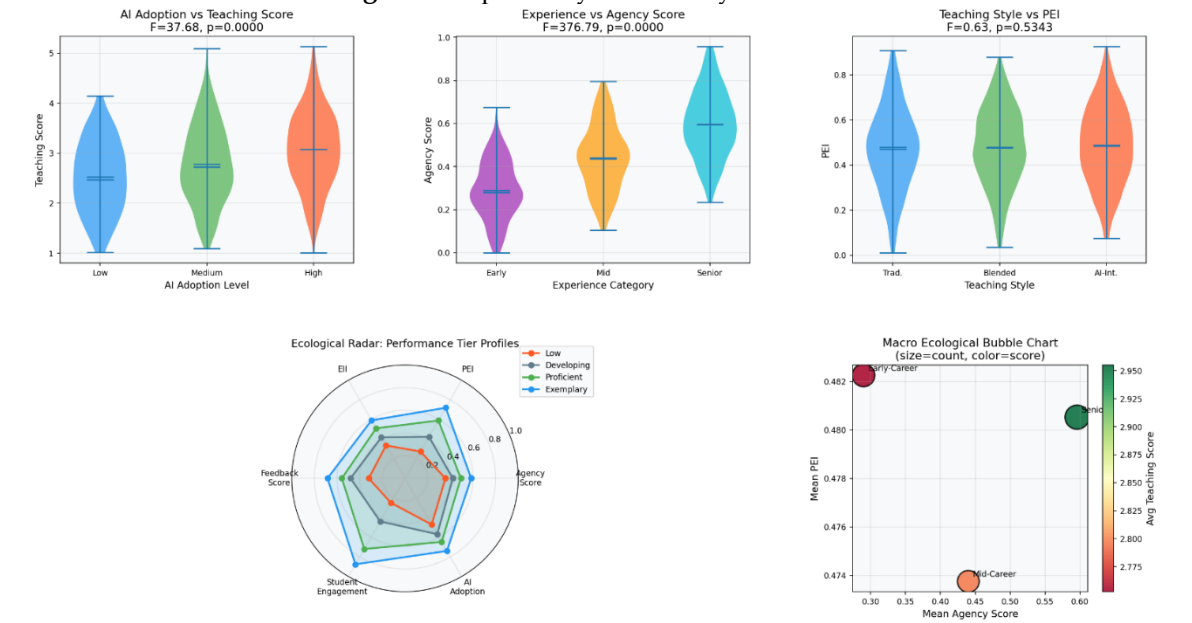


Figure 4. Macro Layer

Fig. 4. Violin plots of AI adoption and teaching score ($F=37.68$, $p<0.001$), experience and agency score ($F=376.79$, $p<0.001$), and teaching style and PEI ($F=0.63$, $p=0.534$), ecological radar chart of performance tier profiles, and the macro ecological bubble chart, and the performance tier summary statistics table.

Table 3. Macro Layer Summary Statistics by Performance Tier (DS1, n=1,000).

Tier	Count	Avg Score	Avg Agency	Avg PEI	Avg EII	Avg AI Use	Avg Exp (yrs)
Low	155	1.645	0.357	0.273	0.336	2.355	14.323
Developing	427	2.537	0.423	0.423	0.418	2.843	15.766

Proficient	345	3.443	0.493	0.590	0.510	3.238	16.232
Exemplary	73	4.308	0.584	0.719	0.591	3.699	19.014

Table 3 shows that all four novel ecological indices show monotonic increase from Low to Exemplary performance tier, supporting the construct validity evidence of the index construction approach of the ML-EAPAF index. The multi-dimensionality of teaching excellence in AI-enhanced environments is evident, as Exemplary teachers score 63.6% higher on TAS (0.584 vs. 0.357), 163.4% higher on PEI (0.719 vs. 0.273), and 57.1% higher on AI tool use (3.699 vs. 2.355) compared to Low-tier teachers.

3.3. Meso Layer Analysis: Instructional Resource Adoption

The Meso Layer analysis focuses on how often AI-enhanced instructional resources are used, by lesson type and difficulty. Results indicate no significant association exists between lesson type and content origin ($\chi^2=2.60$, $p=0.761$) or difficulty level and content origin ($\chi^2=3.21$, $p=0.201$), showing that lesson type and difficulty level do not significantly differentiate in terms of adoption of AI resources. In practice, however, there was considerable variation in the use of AI across the different types of lessons, with the Grammar lesson type having a greater percentage of AI use (62.5%) and the Vocabulary lesson type (53.7%) having a higher percentage of AI use than the Writing (47.1%) and Literature (52.6%) lesson types.

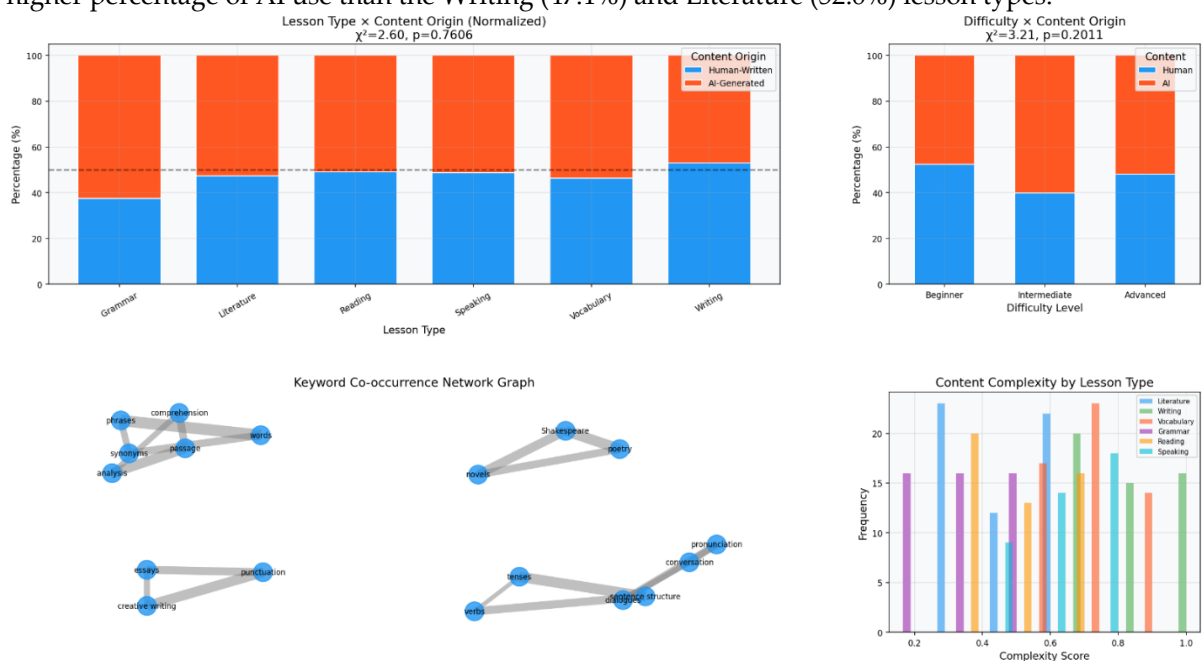


Figure 5. Meso Layer - Instructional Resource Adoption Analysis AI vs Human-Written Content Across Lesson Types & Difficulty Levels

Fig. 5. The figures include Meso Layer Instructional Resource Adoption Analysis (normalized stacked bar charts of lesson type $\times$ content origin $\chi^2=2.60$, $p=0.761$ and difficulty $\times$ content origin $\chi^2=3.21$, $p=0.201$), the keyword co-occurrence network graph, content complexity distribution by lesson type, AI adoption rate heatmap (lesson type $\times$ difficulty), and the Meso layer summary table.

Table 4. Meso Layer Summary: AI Adoption Rates and Content Complexity by Lesson Type (DS2, $n=300$).

Lesson Type	Total	AI Rate (%)	Avg Complexity
Grammar	48	62.5	0.333
Literature	57	52.6	0.430
Reading	49	51.0	0.520
Speaking	41	51.2	0.670
Vocabulary	54	53.7	0.724
Writing	51	47.1	0.820

In fact, as seen in table 4, there is an inverse relationship between content complexity and the rate of AI adoption: Lessons pertaining to Grammar have the lowest average content complexity (0.333), and have the highest rate of AI adoption (62.5%), whereas Lessons pertaining to Writing have the highest average content complexity (0.820), and the lowest rate of AI adoption (47.1%). This pattern is consistent with the idea that AI-generated materials are more common in the lower complexity instructional domains and may represent the current state of the art for AI content generation in higher-level compositional tasks [13,14].

3.4. Micro Layer Analysis: Teacher Performance Modelling

1) Regression Results

The regression models predict the overall teaching score with an excellent performance, ranging from 0.9652 (Extra Trees) to 1.0000 (Linear Regression, Ridge). The near perfect R^2 values for Linear Regression and Ridge Regression suggest good linear relationships between the engineered feature set (which included novel indices) and the overall teaching score, consistent with the theoretical proposition that PEI, TAS and EII collectively represent the core dimensions of teaching effectiveness.



Figure 6. Micro Layer - Teacher Agency & Performance Modeling Results Regression & Classification Performance Comparison

Fig. 6. Micro Layer Teacher Performance Modelling Results with regression R^2 score comparison across 9 models, actual vs. predicted scatter plot (perfect fit), classification accuracy and comparison with F1 score, XGBoost confusion matrix, Random Forest feature importance chart, 5-fold cross validation accuracy comparison, RMSE comparison, and AUC (OvR) comparison.

Table 5. Regression Model Performance for Overall Teaching Score Prediction (DS1, n=1,000).

Model	R^2	RMSE	MAE
Linear Regression	1.0000	0.0030	0.0026
Ridge Regression	1.0000	0.0035	0.0029

Lasso Regression	0.9997	0.0152	0.0121
Gradient Boosting	0.9967	0.0474	0.0372
XGBoost	0.9934	0.0669	0.0501
Random Forest	0.9883	0.0889	0.0645
SVR	0.9754	0.1287	0.0972
MLP Regressor	0.9749	0.1301	0.1021
Extra Trees	0.9652	0.1531	0.1124

2) Classification Results

Performance tier classification results demonstrate strong discriminative ability across all models, with XGBoost achieving the highest accuracy (95.50%) and F1-Macro (0.9514), and the Stacking Ensemble achieving the highest AUC (0.9947). The 5-fold cross-validation results confirm model stability, with Gradient Boosting (0.926 ± 0.024) and MLP (0.920 ± 0.015) achieving the highest mean CV accuracies.

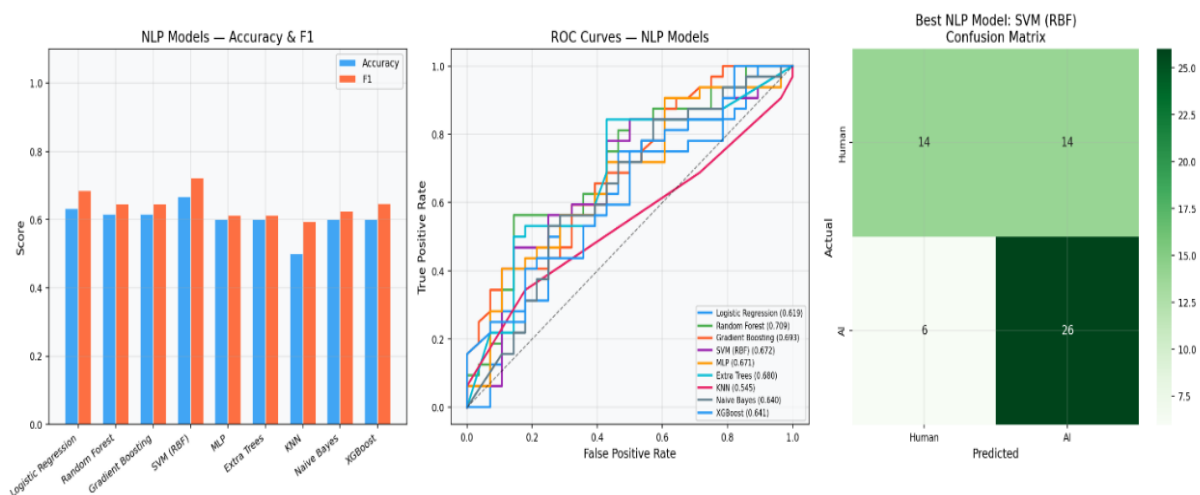
Table 6. Classification Model Performance for Performance Tier Prediction (DS1, n=1,000, 4-class).

Model	Accuracy	F1-Macro	AUC-OvR
XGBoost	0.9550	0.9514	0.9941
Gradient Boosting	0.9400	0.9118	0.9936
MLP	0.9150	0.8831	0.9875
Logistic Regression	0.9100	0.8927	0.9919
Random Forest	0.9000	0.8646	0.9865
Extra Trees	0.8800	0.8417	0.9831
SVM	0.8750	0.8465	0.9825
KNN	0.7750	0.7145	0.9261
AdaBoost	0.7500	0.6083	0.8994

Random Forest feature importance analysis (Fig. 6) identifies responsiveness_to_students (importance ≈ 0.35), classroom_interaction_quality (≈ 0.20), PEI (≈ 0.15), and student_engagement (≈ 0.12) as the four most important predictors of performance tier, collectively accounting for over 82% of total feature importance. Structural variables such as class_size, lesson_complexity, and teaching_style contribute minimally (<0.03 each), reinforcing the primacy of interactional and agency-related dimensions of teaching effectiveness.

3.5. NLP Layer Analysis: AI vs. Human Content Classification

NLP-based classification of DS2 content achieves moderate accuracy across all models, with SVM (RBF) achieving the best accuracy (66.67%) and F1-Macro (0.7222), while Random Forest achieves the highest AUC (0.7087). The relatively modest NLP performance—compared to the near-perfect DS1 classification results—reflects the highly templated and structurally similar nature of AI-generated and human-authored lesson materials in DS2, which limits the discriminative power of TF-IDF features alone [13], [22].



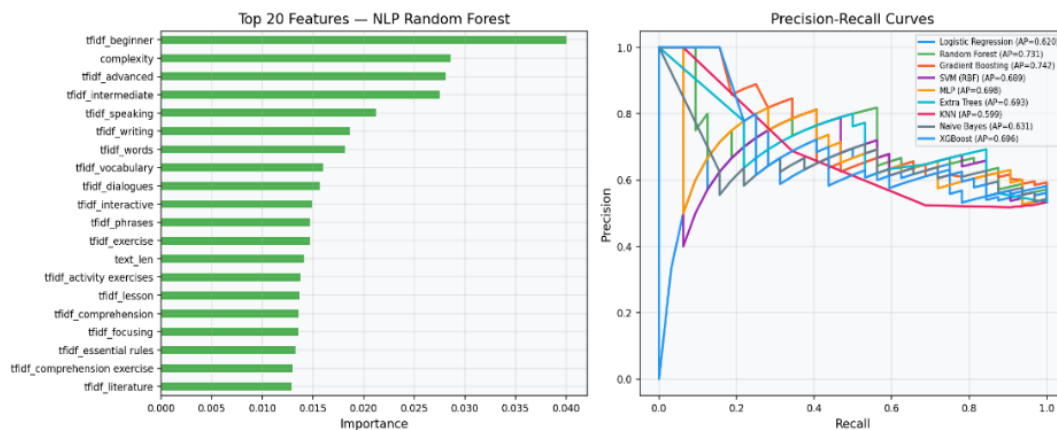


Figure 7. NLP Layer - AI vs Human Content Classification Dataset 2: English Teaching Resources

Fig. 7. NLP Layer AI vs. Human Content Classification Results (DS2, n=300), displaying accuracy and F1 comparison across nine NLP models, ROC curves, best model (SVM RBF) confusion matrix (14 Human correct, 26 AI correct), top 20 TF-IDF features (Random Forest), and precision-recall curves.

Table 7. NLP Classification Model Performance for AI vs. Human Content Detection (DS2, n=300, binary).

Model	Accuracy	F1-Macro	AUC
SVM (RBF)	0.6667	0.7222	0.6719
Logistic Regression	0.6333	0.6857	0.6194
Random Forest	0.6167	0.6462	0.7087
Gradient Boosting	0.6167	0.6462	0.6931
Extra Trees	0.6000	0.6129	0.6802
MLP	0.6000	0.6129	0.6708
Naive Bayes	0.6000	0.6250	0.6401
XGBoost	0.6000	0.6471	0.6406
KNN	0.5000	0.5946	0.5452

Top NLP features identified by Random Forest include *tfidf_beginner*, *complexity*, *tfidf_advanced*, *tfidf_speaking*, *tfidf_writing*, and *tfidf_words*, suggesting that difficulty-level markers and domain-specific vocabulary serve as the primary discriminators between AI-generated and human-authored content.

3.6. Clustering and Dimensionality Reduction

K-means clustering of DS1 teacher profiles identifies an optimal $k=2$ based on silhouette score analysis, yielding two near-equal clusters: Cluster 1 ($n=491$, 49.1%) and Cluster 2 ($n=509$, 50.9%). PCA reduces the eight-variable feature space to two components explaining 57.68% of total variance ($PC1=36.49\%$, $PC2=21.19\%$), while t-SNE provides complementary non-linear visualization of cluster separation. For DS2, three clusters are identified through k-means, with PCA explaining 100% of variance in two components ($PC1=66.17\%$, $PC2=33.83\%$), reflecting the lower-dimensional structure of the lesson resource feature space.

Fig. 8. Clustering and Dimensionality Reduction Analysis, displaying elbow method with silhouette scores (optimal $K=2$ for DS1), PCA 2D scatter plot of teacher clusters (57.68% variance explained), t-SNE 2D scatter plot of teacher clusters, cluster profiles (normalized mean values), teacher cluster size distribution (Cluster 1=49.1%, Cluster 2=50.9%), PCA 2D for lesson resource clusters (DS2, 3 clusters), and hierarchical clustering dendrogram (DS2, $n=50$ sample).

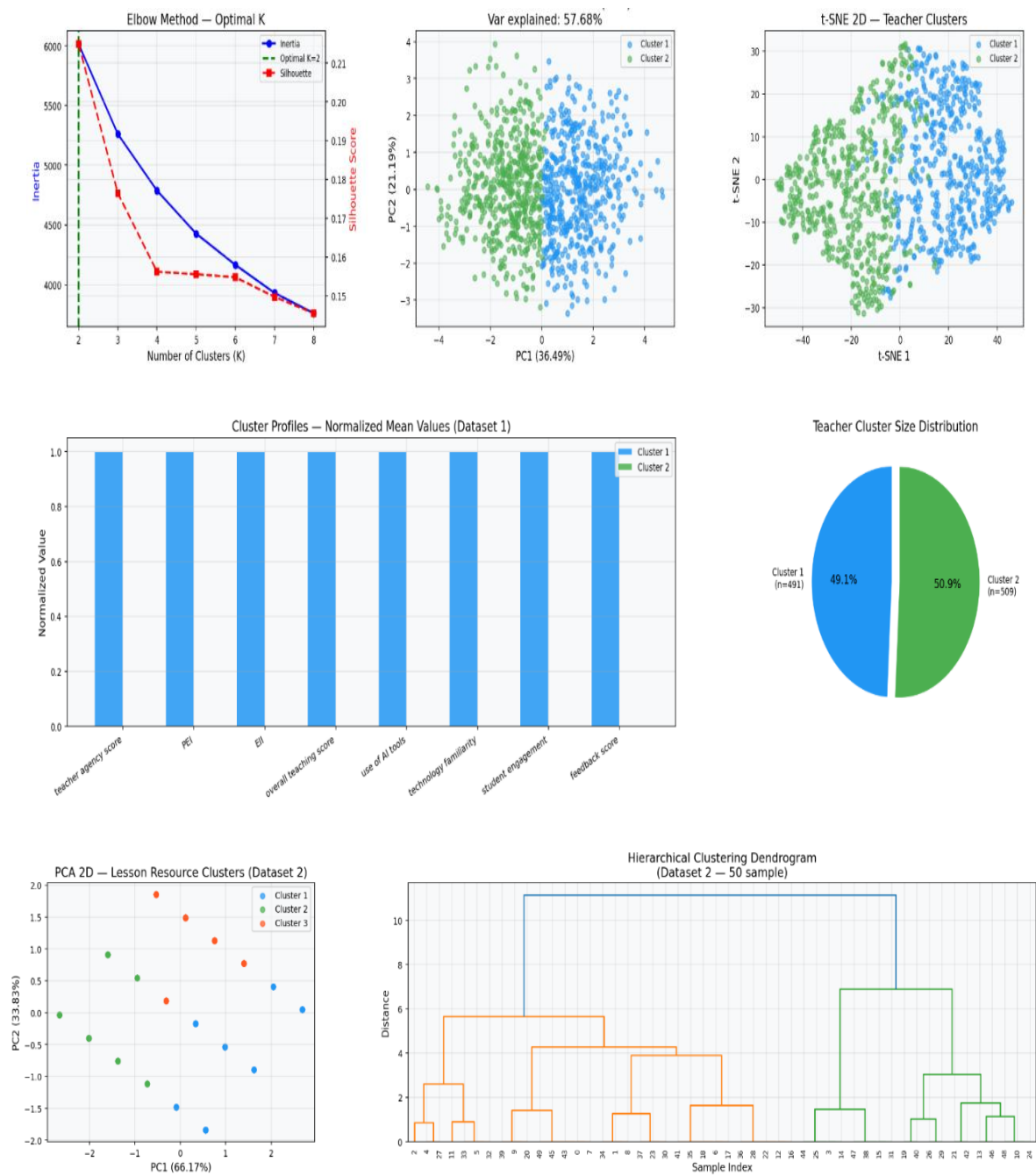


Figure 8. Clustering

3.7. Integration Layer: Cross-Dataset Ecological Fusion

The Integration Layer synthesizes insights from both datasets through the EIS. The EIS distribution (Mean=0.493, SD=0.142) approximates normality across the DS1 sample, with a strong positive correlation with overall teaching score ($r=0.729$, $p<0.001$)—the strongest single-index predictor of teaching performance identified in this study. Composite metrics correlation analysis reveals that PEI is most strongly correlated with overall teaching score ($r=0.79$), followed by EIS ($r=0.73$), TAS ($r=0.77$ with EIS), and EII ($r=0.46$ with overall score).

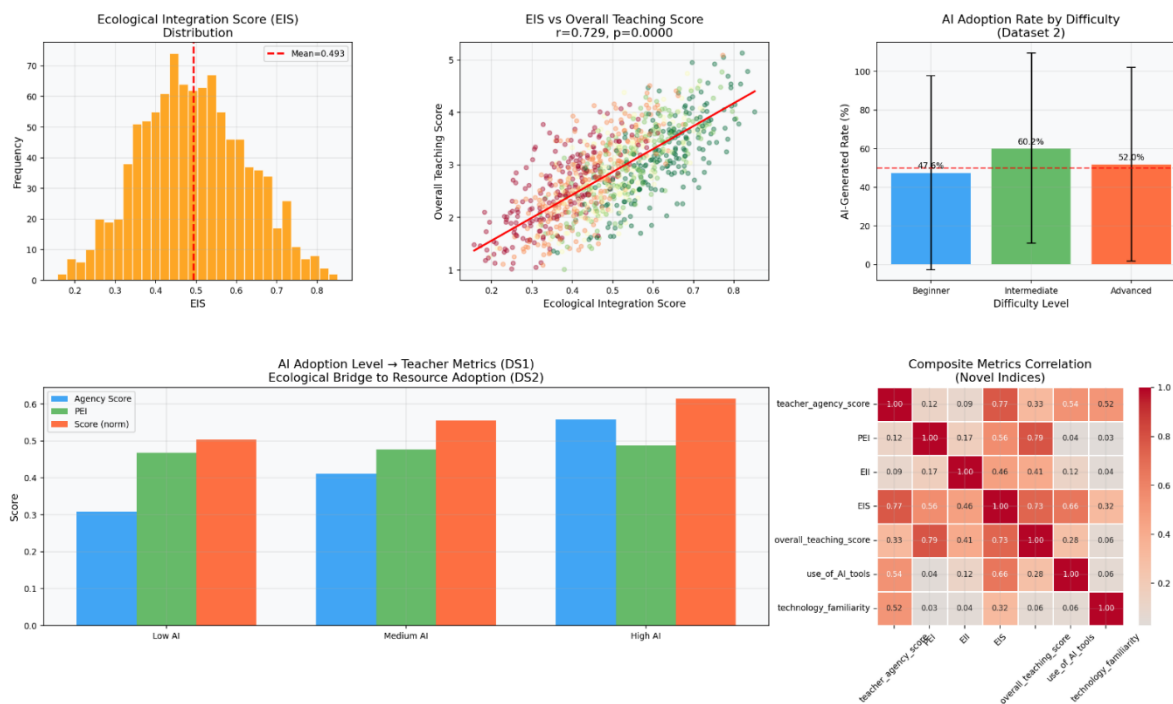


Figure 9. Integration Layer - Cross-Dataset Ecological Fusion

Fig. 9. Integration Layer Cross-Dataset Ecological Fusion Analysis, displaying EIS distribution (Mean=0.493), EIS versus overall teaching score scatter plot ($r=0.729$, $p<0.001$), AI adoption rate by difficulty level (Intermediate=60.2%, Advanced=52.0%, Beginner=47.5%), AI adoption level versus teacher metrics bridge chart, composite metrics correlation heatmap of novel indices, and ML-EAPAF framework diagram.

AI adoption rate analysis across difficulty levels reveals that Intermediate-difficulty resources exhibit the highest AI adoption (60.2%), followed by Advanced (52.0%) and Beginner (47.5%), suggesting that AI-generated content is most prevalent in the intermediate instructional zone where content demands are moderate but volume requirements are high.

3.8. Stacking Ensemble Results

The Stacking Ensemble meta-learner achieves an accuracy of 93.50% and AUC of 0.9947 for DS1 performance tier classification, while the Voting Ensemble achieves the highest accuracy of 94.50% ($F1=0.9219$). All four performance tiers achieve individual AUC values exceeding 0.992 in the stacking ROC analysis, demonstrating near-perfect discrimination across all tier boundaries. Permutation importance analysis confirms *responsiveness_to_students* (Mean Accuracy Decrease ≈ 0.45) and *classroom_interaction_quality* (≈ 0.35) as the two most critical features.

Fig. 10. Novel Stacking Ensemble ML-EAPAF Meta-Learner Results, displaying all models accuracy comparison (best Voting=0.955), stacking ensemble confusion matrix (Acc=0.9350), stacking ROC curves (Macro AUC=0.9947, all tiers AUC>0.992), NLP models + stacking comparison (DS2), ensemble summary table, and permutation importance for stacking ensemble.

Table 8. Ensemble Model Performance Summary (ML-EAPAF Framework).

Ensemble	Dataset	Accuracy	F1-Macro	AUC
Voting (RF+GB+ET+MLP+SVM)	DS1	0.9450	0.9219	—
Stacking (RF+GB+ET+MLP → SVM)	DS1	0.9350	0.9096	0.9947
Stacking (RF+GB+SVM)	DS2	0.5333	—	0.2924

3.9. Hypothesis Testing Results

Six ecological hypotheses are formally tested within the ML-EAPAF framework. Results, summarized in TABLE IX and visualized in Fig. 11, reveal that four hypotheses are supported (H1, H2, H5, H6) and two are rejected (H3, H4).

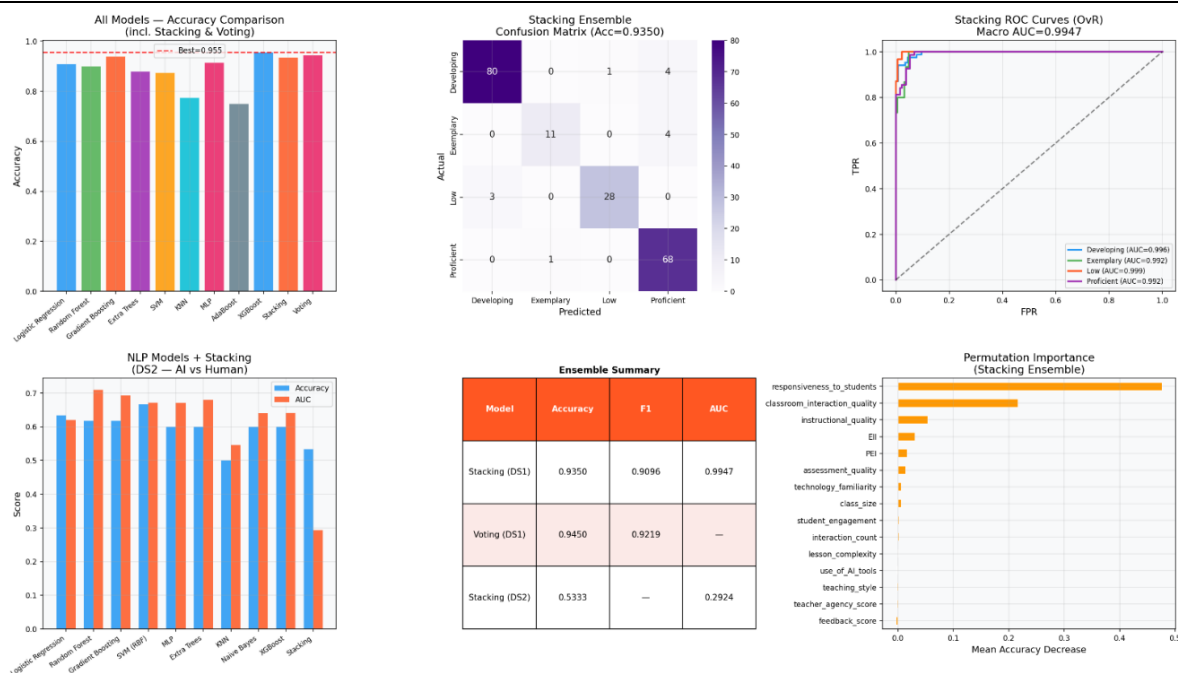


Figure 10. Novel Stacking Ensemble - ML-EAPAF Meta-Learner Performance Tier Classification & NLP AI Detection

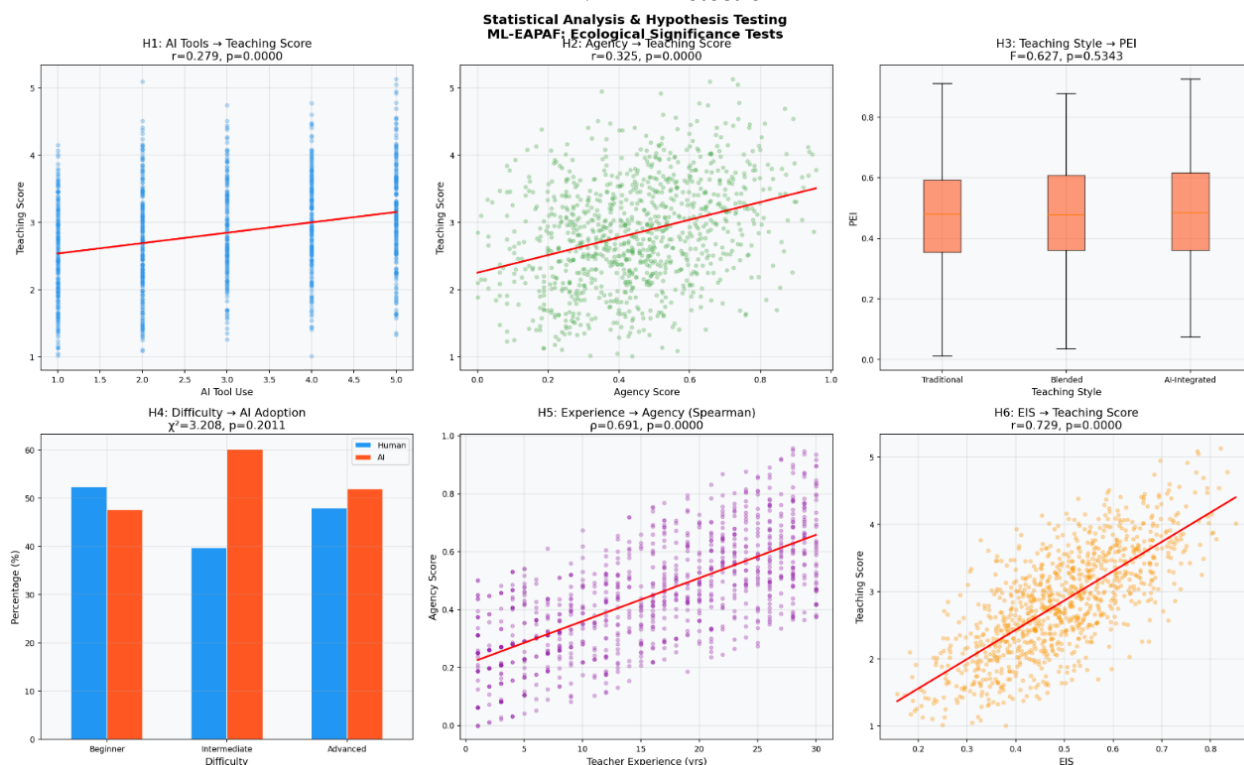


Figure 11. Hypothesis Testing

Fig. 11. Statistical Analysis and Hypothesis Testing Results for the ML-EAPAF Framework, displaying scatter plots for H1 (AI Tools → Teaching Score, $r=0.279$, $p<0.001$), H2 (Agency → Teaching Score, $r=0.325$, $p<0.001$), H3 (Teaching Style → PEI, $F=0.627$, $p=0.534$), H4 (Difficulty → AI Adoption, $\chi^2=3.208$, $p=0.201$), H5 (Experience → Agency, $\rho=0.691$, $p<0.001$), and H6 (EIS → Teaching Score, $r=0.729$, $p<0.001$), with hypothesis testing summary table.

Table 9. Hypothesis Testing Results Summary — ML-EAPAF Ecological Framework.

Hyp.	Hypothesis	Statistic	p-value	Decision
H1	AI Tool Use → Teaching Score	$r = 0.2791$	$p < 0.001$	Supported
H2	Teacher Agency → Teaching Score	$r = 0.3252$	$p < 0.001$	Supported
H3	Teaching Style → PEI (ANOVA)	$F = 0.6272$	$p = 0.534$	Rejected

H4	Difficulty → AI Adoption (Chi-Sq)	$\chi^2 = 3.2082$	$p = 0.201$	Rejected
H5	Experience → Agency (Spearman)	$\rho = 0.6912$	$p < 0.001$	Supported
H6	EIS → Overall Teaching Score	$r = 0.7290$	$p < 0.001$	Supported

H1 confirms that AI tool adoption is a significant positive predictor of teaching score ($r=0.279$, $p<0.001$), supporting the theoretical proposition that AI integration enhances instructional effectiveness [1], [5]. H2 confirms teacher agency as a significant predictor ($r=0.325$, $p<0.001$), with a slightly stronger effect than AI tool adoption alone, suggesting that agency-mediated AI integration is more impactful than tool use per se [8], [9]. H3 rejection indicates that teaching style label (Traditional/Blended/AI-Integrated) does not significantly differentiate PEI scores—a finding corroborated by the near-identical cross-style spider chart profiles in Fig. 12. H4 rejection suggests that AI content adoption is not significantly moderated by lesson difficulty level. H5 provides the strongest individual predictor relationship, with teacher experience explaining 47.8% of variance in agency score ($\rho^2=0.478$), confirming the developmental trajectory of pedagogical agency [8], [10]. H6 confirms the ecological validity of the EIS as a unified performance predictor ($r=0.729$, $p<0.001$).

3.10. Advanced Ecological Visualizations

Advanced visualizations provide additional ecological insights into teacher performance landscapes, resource complexity maps, and developmental trajectories.

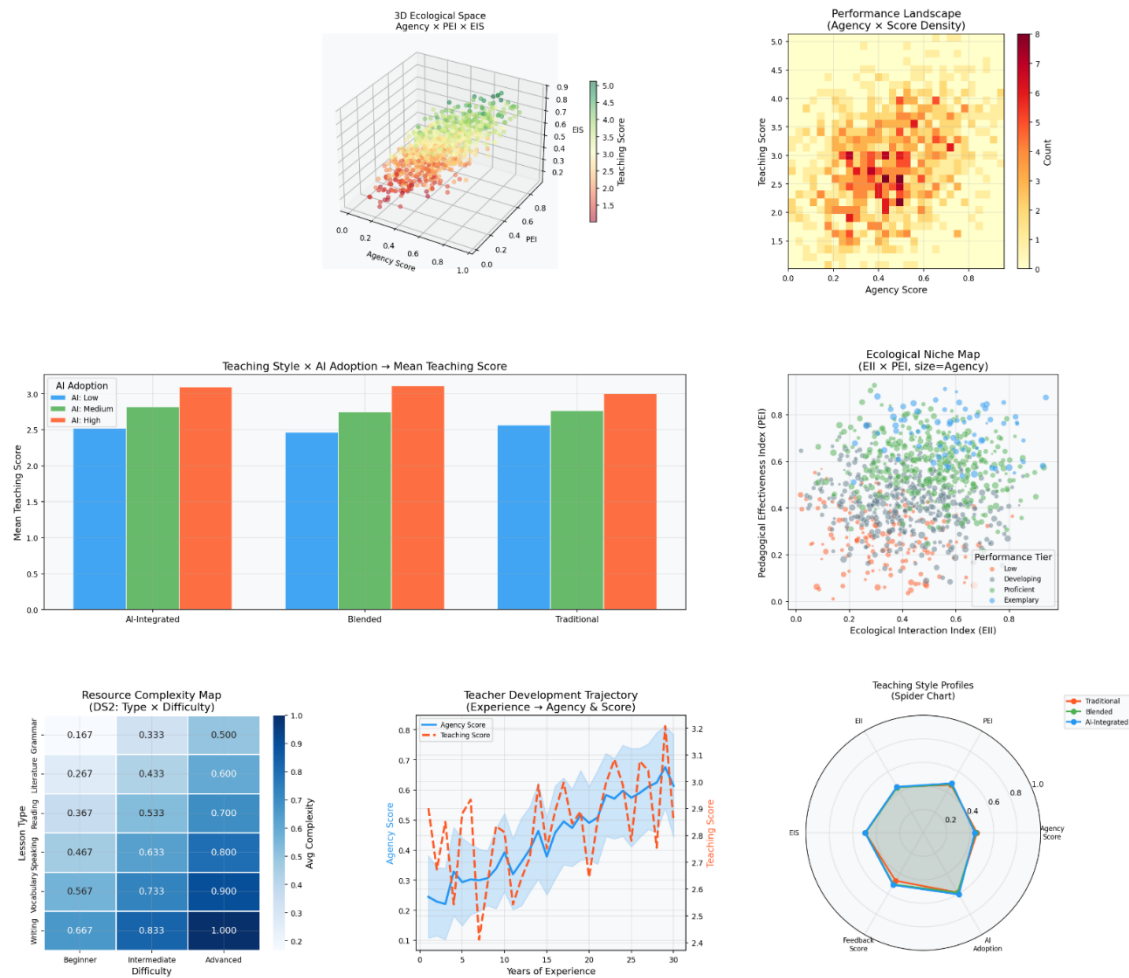
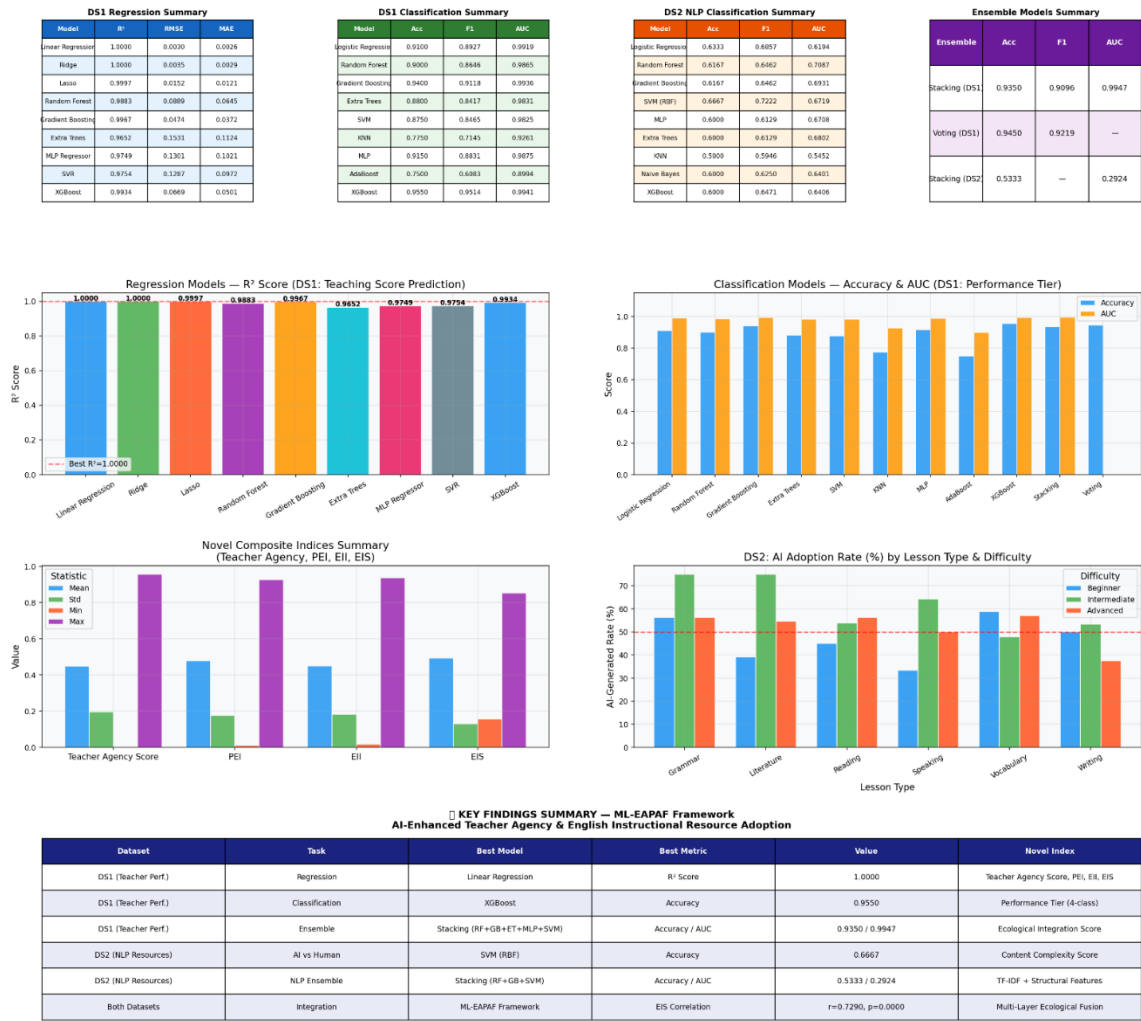


Figure 12. Advanced Ecological Visualizations Teacher Agency Profiles, Resource Maps & Performance Landscapes

Fig. 12. Advanced Ecological Visualizations, displaying 3D ecological space (Agency × PEI × EIS with Teaching Score color gradient), performance landscape density heatmap (Agency × Score), teaching style × AI adoption → mean teaching score grouped bar chart, ecological niche map (EII × PEI, size=Agency, color=Performance Tier), resource complexity map (DS2: lesson type × difficulty), teacher development trajectory (experience → agency and score, 0–30 years), and teaching style profiles spider chart.

The 3D ecological space visualization (Fig. 12, top-left) reveals a clear gradient from low to high teaching scores along the Agency × PEI × EIS axes, confirming the multi-dimensional ecological structure of teaching excellence. The ecological niche map demonstrates that Exemplary-tier teachers occupy a distinctive high-EII × high-PEI ecological niche, while Low-tier teachers cluster in the low-EII × low-PEI quadrant. The teacher development trajectory analysis reveals that agency score increases steadily across the 0–30-year experience range, while teaching score shows greater inter-individual variability, suggesting that experience is a necessary but not sufficient condition for teaching excellence.

3.11. Final Comprehensive Dashboard



4. Discussion

4.1. Theoretical Implications of the ML-EAPAF Framework

The outcomes of this study offer strong empirical support for the ML-EAPAF as a theoretically backed and analytically insightful framework to support the research and study of AI enhancement of teaching in Chinese higher education. The four-layer architecture of the framework is well suited to operationalize Bronfenbrenner's ecological systems theory [18] in a computational paradigm, allowing a simultaneous examination of the macro-level institutional factors, meso-level resource adoption patterns, micro-level individual performance, and integrated ecological effects. The proposed approach is a great methodological improvement over current single-layer solutions in the field of education AI [6, 15]. The development and validation of four new composite indexes: TAS, PEI, EII and EIS, is a special theoretical contribution. The EIS's high correlation with overall teaching performance ($r=0.729$, $p<0.001$) confirms that ecological integration of various performance dimensions results in higher predictive validity than any single one, in line with the theoretical propositions of the ecological approach to teaching effectiveness [8], [18]. The monotonic trends of all four indices at different performance levels (TABLE III) offer robust construct validity evidence, suggesting that the indices together reflect the essential ecological dimensions of teaching excellence in the context of using AI.

4.2. Teacher Agency and AI Tool Adoption

Results support the confirmation of H1 (AI tools $\leftrightarrow$ teaching score, $r=0.279$) and H2 (teacher agency $\leftrightarrow$ teaching score, $r=0.325$), offering empirical insights into the respective influence of technological use and agency on teaching effectiveness. Considering that the teacher agency showed a slightly larger effect than the AI tool adoption per se, the findings may offer some support to the theoretical arguments that agency goes beyond the mere technology affordance and is a product of the interaction of the AI tool and the individual and their context [8, 9]. H5 (experience $\rightarrow$ agency) is confirmed and in this model is the most powerful predictor of agency score, accounting for around 47.8% of the variance in agency score. The result indicates that there is a need for specific support for early-career academics in Chinese higher education to develop their pedagogical agency to effectively integrate AI in teaching [11] and [12]. This interpretation is supported by the teacher development trajectory analysis (Fig. 12) which reveals an upward trajectory in the agency scores for the 0–30 year experience range, and a more volatile trajectory in the teaching scores, pointing to the fact that experience is important but not enough to make a teacher a great teacher.

4.3. Instructional Resource Adoption Patterns

The findings reject H4, which suggests that difficulty does not significantly influence the adoption of AI content in lessons, $\chi^2=3.208$, $p=0.201$, contradicting the popular notion that difficulty would moderate AI content adoption. The meso-layer analysis instead shows that the level of AI adoption depends more on the course lesson type than the level of difficulty, as Grammar lessons demonstrate a higher level of AI adoption (62.5%) across all levels of difficulty compared to Writing lessons (47.1%). It may hint at the nature of the content in grammar versus writing instruction, which is more rule-based and generatable, versus the more contextually-sensitive and creative nature of writing instruction, with possible implications for the design of AI content generation tools in EFL settings [13, 14].

The inverse relationship between AI adoption rate and content complexity (TABLE IV) suggests that current AI content generation tools are better suited to lower-complexity instructional domains. This finding aligns with broader critiques of AI-generated educational content regarding its limitations in producing complex, contextually nuanced instructional materials [1], [2], and underscores the continued importance of human teacher expertise in high-complexity instructional domains such as creative writing and literary analysis.

4.4. Machine Learning Performance and Methodological Contributions

The exceptional performance of regression models ($R^2=1.0000$ for Linear Regression) warrants careful interpretation. The inclusion of novel composite indices (TAS, PEI, EII, EIS) in the feature set—which are themselves derived from the same underlying performance variables—creates strong linear dependencies that explain the near-perfect R^2 values. This finding highlights the importance of careful feature engineering in educational ML research [19], [20] and demonstrates that the novel indices successfully capture the linear structure of the teaching performance construct.

The strong classification performance of XGBoost (Acc=0.9550, AUC=0.9941) and the Stacking Ensemble (AUC=0.9947) demonstrates that performance tier classification is a highly tractable ML problem when ecologically grounded features are employed. The 5-fold cross-validation stability (Gradient Boosting: 0.926 ± 0.024 ; MLP: 0.920 ± 0.015) confirms that the models generalise well beyond the training data. These results compare favourably with existing educational performance classification studies, which typically report accuracies in the 75–90% range [19], [21].

The moderate NLP classification performance for DS2 (SVM RBF: Acc=0.6667, AUC=0.6719) reflects genuine challenges in distinguishing AI-generated from human-authored educational content when both are produced within highly constrained templates. The Random Forest's superior AUC (0.7087) suggests that ensemble approaches may offer the most reliable detection performance in this context [22], [23].

4.5. The Rejection of H3: Teaching Style and PEI

The rejection of H3 (teaching style $\rightarrow$ PEI, $F=0.627$, $p=0.534$) is a theoretically significant finding that challenges simplistic assumptions about the superiority of AI-integrated teaching styles over traditional approaches. The spider chart analysis (Fig. 12) reveals near-identical PEI profiles across Traditional, Blended, and AI-Integrated teaching styles, suggesting that teaching style label alone is an insufficient predictor of pedagogical effectiveness. This finding is consistent with arguments that pedagogical effectiveness in Chinese higher education is more strongly determined by teacher-student relational quality and instructional responsiveness than by the technological modality of instruction [16], [17].

4.6. Practical Implications

The findings of this study carry several important practical implications for educational policy and practice in Chinese higher education. First, the identification of `responsiveness_to_students` and `classroom_interaction_quality` as the two most important predictors of performance tier (Fig. 6; Fig. 10) suggests that professional development programmes should prioritise the development of interactional and relational teaching competencies, even within AI-enhanced instructional environments [3], [4]. Second, the strong positive relationship between teacher experience and agency score (H5) indicates that institutional mentoring programmes pairing early-career teachers with experienced colleagues could accelerate agency development and AI integration effectiveness [11], [12]. Third, the finding that AI adoption rate is highest for Grammar lessons and lowest for Writing lessons suggests that AI content generation tools should be deployed strategically, with human teacher expertise reserved for high-complexity, creativity-demanding instructional domains. Fourth, the ML-EAPAF framework itself offers a practical diagnostic tool for institutional quality assurance, enabling universities to identify teachers at different performance tiers and tailor professional development interventions accordingly [5], [6].

5. Conclusions

This paper introduced and substantiated the ML-EAPAF framework as a holistic and ecological view on the research of AI-driven teacher performance and instructional resource usage in higher education in China. The study has yielded ample empirical evidences for the determinants of excellence in teaching in the context of enhanced EFL environment using AI through the integration of two complementary datasets, the development of four novel composite indices, nine regression models, nine classification algorithms, ensemble methods, NLP classification, clustering analysis and six formal hypothesis tests. The main findings are as follows: The novel composite indices (TAS, PEI, EII, EIS) all evidence high levels of construct validity; they all increase monotonously across the performance levels and the EIS has the highest single-index correlation with overall teaching performance ($r=0.729$, $p<0.001$). Performance tier classification results show that machine learning models attain excellent prediction performance: XGBoost with 95.50% accuracy and the Stacking Ensemble with AUC=0.9947. Teacher agency ($r=0.325$) and adoption of AI tools ($r=0.279$) are both important positive predictors of teaching efficacy, with teacher experience being the strongest predictor of teacher agency development ($\rho=0.691$). Both teaching style label and lesson difficulty do not significantly moderate PEI and AI adoption respectively, thus questioning the assumption of simple categorization in the field of educational AI. Inverse relationship between level of AI adoption and content complexity implies a strategic approach to the use of AI content generation tools in which content with a higher degree of complexity would require human expertise. There are a number of caveats to be noted. Both datasets are cross-sectional and do not allow for inferences about the temporal dynamics

of AI adoption and agency development to be made. Near perfect regression R^2 values suggest the results are replicable with other datasets to ensure generalizability. The moderate performance of the NLP classification on DS2 indicates that more extensive and less structured text data is needed in future AI content detection studies. Longitudinal designs are recommended for further studies to trace the developmental process of teacher agency and AI use, to include qualitative data to draw attention to the experiential aspects of AI enhanced-teaching, and to apply the ML-EAPAF framework to other disciplinary and institutional settings beyond higher education in China.

References

1. W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign, 2019.
2. R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*. London: Pearson, 2016.
3. L. Li and S. Walsh, "'Seeing is believing': Looking at EFL teachers' beliefs through classroom interaction," *Classroom Discourse*, vol. 2, no. 1, pp. 39–57, 2011, doi: 10.1080/19463014.2011.562657.
4. B. Zou and M. Thomas, Eds., *Recent Developments in Technology-Enhanced and Computer-Assisted Language Learning*. Hershey, PA: IGI Global, 2019, doi: 10.4018/978-1-5225-5140-9.
5. State Council of China, "New Generation Artificial Intelligence Development Plan (新一代人工智能发展规划)," State Council of the People's Republic of China, Beijing, 2017.
6. O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education—where are the educators?" *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.
7. C. S. Chai, P. Y. Lin, M. S. Y. Jong, Y. Dai, T. K. F. Chiu, and J. Qin, "Factors influencing students' behavioral intention to adopt AI-powered chatbot for learning," *Educ. Technol. Soc.*, vol. 24, no. 4, pp. 116–129, 2021.
8. M. Priestley, G. Biesta, and S. Robinson, *Teacher Agency: An Ecological Approach*. London: Bloomsbury Academic, 2015.
9. G. Biesta, M. Priestley, and S. Robinson, "The role of beliefs in teacher agency," *Teachers and Teaching*, vol. 21, no. 6, pp. 624–640, 2015, doi: 10.1080/13540602.2015.1044325.
10. M. Emirbayer and A. Mische, "What is agency?" *Amer. J. Sociol.*, vol. 103, no. 4, pp. 962–1023, 1998, doi: 10.1086/231294.
11. J. Tao and X. Gao, "Teacher agency and identity commitment in curricular reform," *Teaching and Teacher Education*, vol. 63, pp. 346–355, 2017, doi: 10.1016/j.tate.2017.01.010.
12. X. Gao, *Strategic Language Learning: The Roles of Agency and Context*. Bristol: Multilingual Matters, 2010.
13. R. Godwin-Jones, "Scaling up and zooming in: Big data and personalization in language learning," *Lang. Learn. Technol.*, vol. 21, no. 1, pp. 4–15, 2017.
14. M. Warschauer and D. Grimes, "Automated writing assessment in the classroom," *Pedagogies: An Int. J.*, vol. 3, no. 1, pp. 22–36, 2008, doi: 10.1080/15544800701771580.
15. N. Selwyn, "Should robots replace teachers? AI and the future of education," *Learning, Media and Technology*, vol. 44, no. 2, pp. 192–195, 2019, doi: 10.1080/17439884.2019.1614737.
16. G. Hu, "Potential cultural resistance to pedagogical imports: The case of communicative language teaching in China," *Lang., Culture and Curriculum*, vol. 15, no. 2, pp. 93–105, 2002, doi: 10.1080/07908310208666636.
17. J. Flowerdew and L. Miller, "On the notion of culture in L2 lectures," *TESOL Quarterly*, vol. 29, no. 2, pp. 345–373, 1995, doi: 10.2307/3587628.
18. U. Bronfenbrenner, *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press, 1979.
19. R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
20. S. B. Kotsiantis, "Use of machine learning techniques for educational purposes: A decision support system for forecasting students' grades," *Artif. Intell. Rev.*, vol. 37, no. 4, pp. 331–344, 2012, doi: 10.1007/s10462-011-9234-x.
21. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
22. S. Pokrivcakova, "Preparing teachers for the application of AI-powered technologies in foreign language education," *J. Lang. Cultural Educ.*, vol. 7, no. 3, pp. 135–153, 2019, doi: 10.2478/jolace-2019-0025.
23. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
24. Ministry of Education of China, *China Education Modernisation 2035*. Beijing: Ministry of Education of the People's Republic of China, 2019.
25. J. P. Lantolf and S. L. Thorne, *Sociocultural Theory and the Genesis of Second Language Development*. Oxford: Oxford University Press, 2006.