ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Constructing and Validating an AI-Empowered Evaluation Framework for Industry–Education Integration

Xiaoyan Yu[1,*] *and Jun Shen*[1]

[1] ZheJiang Institute of Communications, Hangzhou, 311112, Zhejiang, China

SUMMARY: *In response to the problems of scattered indicators, insufficient result verification, and weak interpretability in the evaluation of industry education integration, this article constructs an artificial intelligence (AI) empowerment evaluation framework. Based on observations from 126 universities and 3010 questionnaires from 42 universities from 2022 to 2024, this study integrates five dimensions: collaborative governance, curriculum co construction, practical platform, process quality, and outcome effectiveness. Research combines Analytic Hierarchy Process (AHP), Entropy Weight Method, and Light Gradient Boosting Machine (LightGBM) modeling. The results showed that the mean absolute error (MAE) and root mean square error (RMSE) of the LightGBM model were 0.138 and 0.189, respectively, with a coefficient of determination (R^2) of 0.814; The framework score has a significant positive impact on the corresponding employment rate, enterprise satisfaction, and collaborative innovation output. The key driving factors for high-quality industry education integration are the participation intensity of enterprise mentors, the coverage rate of curriculum co construction, and the proportion of dual teacher teachers, indicating that the deep participation of enterprises in curriculum and practice is more important.*

KEYWORDS: *Integration evaluation of industry and education; artificial intelligence empowerment; LightGBM; external validation of efficacy standards; interpretability analysis*

1 Introduction

Against the backdrop of continuously promoting high-quality development in vocational education and applied higher education, the integration of industry and education has gradually shifted from establishing cooperative relationships to identifying the effectiveness of cooperation and evaluating the quality level. For a considerable period of time in the past, many universities focused more on visibility issues such as whether to establish a cooperation platform, whether to sign a cooperation agreement, and whether to have a training base. However, with the increasing connection between talent cultivation quality, employment adaptability, and regional industrial upgrading, judging the level of industry education integration solely based on the existence of cooperation is no longer sufficient to meet the practical needs of educational governance and school improvement. At the same time, artificial intelligence is profoundly influencing the organization of educational activities, learning support mechanisms, and quality evaluation logic. According to relevant review studies, the application of artificial intelligence in education has gradually extended from early auxiliary teaching tools to deeper fields such as learning analysis, educational decision-making, personalized support, and evaluation optimization. This indicates that educational evaluation

*yushen202604@163.com
https://doi.org/10.65102/is20261068 Published: 10 June, 2026

itself is entering a new stage that places more emphasis on data integration and model recognition. In the context of higher education, the discussion on empowering learning with artificial intelligence also shows that the digitization of the educational process not only changes the way resources are supplied, but also reshapes the relationship between teacher support, learning behavior, and management collaboration. Therefore, the evaluation framework for complex educational objects also needs to shift from static indicator accumulation to a more adaptive comprehensive identification mechanism.

From the development trend of measurement research, the education sector has accumulated mature experience in tool development in areas such as artificial intelligence literacy, acceptance, and behavioral intention. Hornberger et al. constructed and validated assessment tools around the level of artificial intelligence knowledge among college students, demonstrating that abstract ability is not unquantifiable. As long as it is defined by clear dimensions and rigorously tested for validity, complex educational concepts can also be transformed into stable evaluation objects. Laupichler et al. used Delphi method to develop artificial intelligence literacy measurement items for non professional groups, further demonstrating the important role of expert consultation, item iteration, and preliminary validation in improving the credibility of evaluation tools [4]. When evaluating artificial intelligence literacy courses, Kong et al. included conceptual understanding, empowerment, and ethical awareness in their analysis, reflecting the increasing emphasis on collaborative assessment of multidimensional goals in educational evaluation. Chai et al. proposed and validated the Artificial Intelligence Learning Intention Scale, demonstrating that learner attitudes and behavioral tendencies can be entered into model analysis through structured measurements [6]. The study by Cabero Almenara et al. on the acceptance of artificial intelligence in teacher education also indicates that there is a recognizable correlation between teacher beliefs, technological attitudes, and educational contexts, which provides methodological implications for linking subjective cognition with objective outcomes [7]. Although these achievements mainly focus on artificial intelligence education scenarios, they collectively demonstrate that the evaluation of complex educational phenomena should not be limited to empirical judgment, but should rely on clear indicator structures, repeatable weighting processes, and verifiable result testing.

Compared to this, although research on the evaluation of industry education integration has begun to move towards quantitative analysis, there is still significant room for expansion overall. On the one hand, some studies have recognized the importance of information systems and digital platforms for the integration of industry and education. The research conducted by He et al. on the integrated information management system of industrial education shows that collaborative activities between industry and education have already exhibited strong data-driven characteristics, and related evaluation work should shift from decentralized recording to systematic integration [8]. On the other hand, the integration of industry and education is not confined to internal school management affairs, but is closely linked to changes in industry demand, motivation for enterprise participation, and multi-party cooperation relationships. Zheng et al. explored the willingness of stakeholders to participate in vocational education integration from the perspective of the hydrogen energy industry, indicating that the quality of industry education integration largely depends on whether a stable and sustainable collaborative mechanism is formed between schools and enterprises. In addition, Gong conducted empirical analysis on the performance of the integration of industry and education in higher education from the perspective of coupling coordination, providing empirical evidence for quantitative evaluation at the regional level and reflecting that current research has begun to focus on the empirical path of evaluation [10]. However, existing achievements still mostly remain at the level of macro coordination, project quantity, or comprehensive index. There is a lack of deeper

answers to questions such as which dimensions constitute the integration of industry and education, whether evaluation results can correspond to real educational performance, and how key driving factors are identified.

Specifically, existing research has at least three shortcomings. Firstly, the construction of the indicator system still appears scattered. Many studies tend to use single indicators such as the number of collaborative projects, the number of bases, cooperation agreements, or the frequency of enterprise participation, but rarely understand collaborative governance, curriculum co construction, practical platforms, process quality, and result performance in a unified framework. This approach can reflect the existence of cooperative activities, but it is difficult to judge the depth, stability, and educational effectiveness of cooperation. Secondly, the evaluation results often lack external criteria support. Although some studies can report internal consistency, factor structure, or comprehensive scores, they have not further examined whether high scoring universities truly perform better in terms of targeted employment rate, enterprise satisfaction, vocational certificate pass rate, joint project output, and other aspects. Without this layer of external verification, the evaluation results are easily stuck in the statistical sense of "self consistency" and cannot prove their explanatory power in reality. Thirdly, the application of artificial intelligence in the evaluation of industry education integration is not deep enough. There has been more discussion on understanding artificial intelligence as a teaching tool, platform function, or digital literacy content. However, research on using machine learning for evaluation modeling, weight optimization, result prediction, and explanatory analysis is still relatively limited. There is also a lack of sufficient exploration on why models are effective, which indicators are most critical, and whether there are significant differences among different types of universities.

Based on the above background, this article intends to construct and validate an AI powered evaluation framework for vocational education and applied higher education, which is used to identify key dimensions of the quality of industry education integration, compare its performance with traditional weighting methods, and test its applicability in different types of colleges, regional backgrounds, and outcome variables. The focus of this article is not only on how to rank schools, but more importantly, to form an evaluation tool that can support diagnosis, comparison, and improvement through multi-source data integration, mixed weighting, and interpretable modeling. This article proposes three research questions around this goal. RQ1, What dimensions and indicators can effectively characterize the quality of industry education integration. Can RQ2, a hybrid evaluation framework empowered by AI, outperform traditional weighting methods in terms of evaluation performance. RQ3, Can the framework demonstrate stable external validity across different types of institutions, regional differences, and outcome variables. The main contributions of this article are reflected in the following aspects. Firstly, this article constructs a hierarchical and operable evaluation framework for the quality of industry education integration, incorporating collaborative governance, curriculum co construction, practical platforms, process quality, and outcome performance into the same analysis system, thereby enhancing the overall evaluation. Secondly, this article combines the expert judgment formed by Delphi/AHP with the objective information reflected by entropy weight method, and further introduces gradient boosting to participate in result prediction, so that the evaluation process absorbs both institutional experience and data patterns. Again, this article does not consider the comprehensive score as the research endpoint, but introduces external results such as employment quality, enterprise satisfaction, and collaborative innovation output to test the practical explanatory power of the evaluation framework. Finally, this article combines SHAP, subgroup analysis, and robustness tests to analyze the key driving factors, heterogeneity characteristics, and conclusion stability in the model, in order to enhance the interpretability and application value of the evaluation results.

Overall, the evaluation of industry education integration is no longer suitable to remain at the level of single indicator listing or one-time experience judgment. As education digitization and artificial intelligence methods continue to enter quality governance practices, how to establish an evaluation framework that combines measurability, predictability, and interpretability has become a core issue worthy of further promotion in the research of industry education integration. This article is based on this judgment to conduct research, hoping to provide more empirical methodological support for the quality evaluation of industry education integration, as well as provide reference evidence for optimizing cooperation models for universities, improving participation methods for enterprises, and implementing classified governance by education management departments.

2 Methods

2.1 Data sources, sample construction, and variable definition

To enhance the ability of the evaluation framework to depict real educational situations, this article uses a combination of questionnaires, school public reports, and statistical bulletins as multi-source heterogeneous data. The sample covers 12 provinces in the eastern, central, and western regions, with a total of 42 colleges selected, including 20 vocational colleges and 22 applied undergraduate colleges. The research period is set from 2022 to 2024, and institutional level panels will be constructed annually, forming a total of 126 institution year observations. This sample arrangement takes into account regional differences, differences in educational levels, and temporal changes, which helps to identify horizontal differences between different universities and the changing characteristics of the same university in consecutive years in subsequent analysis.

Subjective data mainly comes from stratified questionnaire surveys. 2160 student samples, with a focus on reflecting course participation, practical involvement, learning acquisition, and job fit perception; 486 samples of full-time teachers are mainly used to describe curriculum co construction, practical teaching organization, collaborative participation of enterprises, and teaching feedback. 238 samples of enterprise mentors and managers were used to identify the evaluation of cooperation depth, talent matching, and cooperation output on the enterprise side. 126 samples of academic administrators and industry education integration managers were collected to supplement information on institutional arrangements, resource coordination, and school enterprise collaborative governance. The questionnaire adopts the Likert five point scale and combines the school's public system text and annual quality report to perform semantic calibration on key items, in order to reduce measurement bias caused by inconsistent understanding of the same concept among different respondents [11, 12]. The formation process of the research object and the integration process of multi-source data are shown in Figure 1.

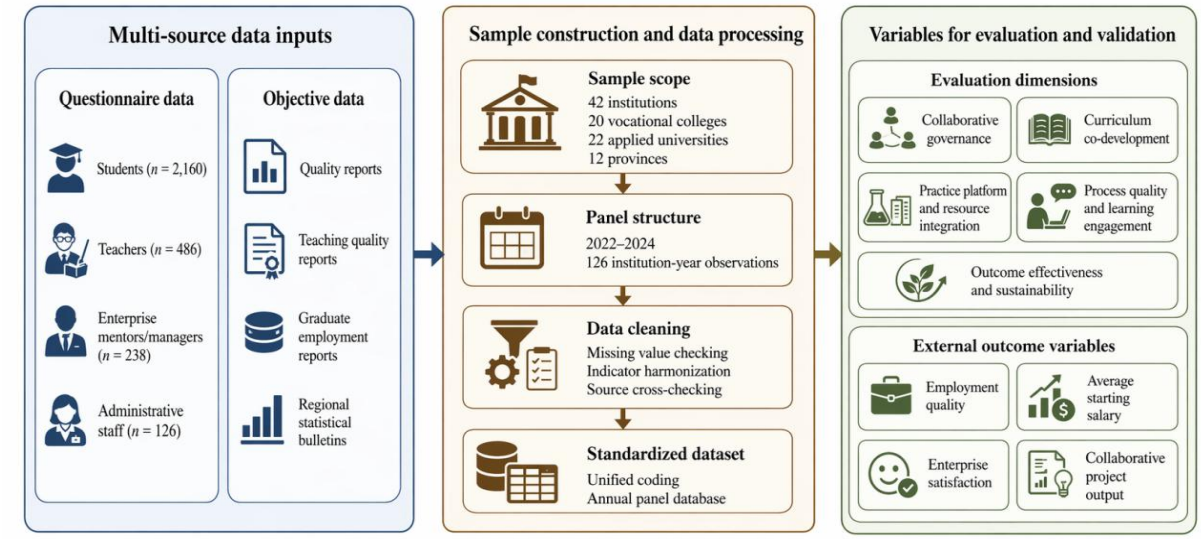

Figure 1: Overall research design and multi-source data integration process for the AI-empowered evaluation framework.

As shown in Figure 1, this article constructs an evaluation framework based on multi-source evidence. The objective data section is mainly compiled based on the undergraduate teaching quality report, higher vocational education quality annual report, and graduate employment quality annual report publicly released by the school, supplemented by regional control information from the National Bureau of Statistics and some provincial education development statistical bulletins [13]. Since 2013, the Ministry of Education has required universities to compile and publish annual reports on the employment quality of graduates. The list of university information disclosure also clearly includes undergraduate teaching quality reports and annual employment quality reports as timely public content. From the retrieved public materials of schools, it can be seen that the annual quality reports of vocational colleges, applied undergraduate teaching quality reports, and graduate employment quality annual reports do constitute the most operable sources of public data currently available.

The core outcome variables at the institutional level include the corresponding employment rate, average salary after six months of graduation, 1+X or vocational qualification certificate pass rate, enterprise satisfaction, number of school enterprise co built courses, proportion of dual teacher teachers, supply ratio of internship positions, as well as the number of joint projects, patents or horizontal projects [14]. These indicators correspond to the effectiveness of talent cultivation, the intensity of school enterprise collaboration, the support of practical resources, and the sustainable output of cooperation, which can reflect the main links from institutional investment to the presentation of results in the integration of industry and education more comprehensively [15]. Considering that there may be differences in the public statements of different schools, this article prioritizes the use of formal statistical values from annual reports when organizing data. For projects with inconsistent statements, a unified statement will be established based on the original public text, and variable cleaning and missing value verification will be completed before subsequent modeling [16].

In terms of variable structure, this article sets five primary indicators, namely Collaborative governance, Curriculum co-development, Practice platform and resource integration, Process quality and learning engagement, And the outcome effectiveness and sustainability. Further refine 20 secondary indicators around these five dimensions. The dimensions of collaborative governance mainly examine the stability of cooperation mechanisms, the degree of enterprise participation in professional construction, the frequency of joint governance meetings, and the

coordination of resources. The dimension of curriculum co construction focuses on the proportion of school enterprise co construction courses, the participation of enterprises in updating teaching content, the embedding degree of practical courses, and the coverage of project-based teaching. The dimensions of practical platform and resource integration measure the availability of on campus and off campus training platforms, job supply, dual teacher type teachers, and equipment resources. The dimensions of process quality and learning engagement reflect students' practical participation, classroom interaction, timely feedback, and sense of learning acquisition. The results and sustainability dimensions comprehensively examine employment adaptation, salary levels, certificate attainment, enterprise evaluation, and collaborative innovation output. To ensure the comparability of subsequent weighting and predictive analysis, all indicators are uniformly processed and standardized before entering the model, and missing items are corrected through a combination of report review and multi-source supplementation.

2.2 Indicator system and AI-empowered evaluation model

After completing data organization and variable standardization processing, this article further constructs a quality evaluation model for the integration of industry and education. Considering that this study involves both experts' judgments on the importance of indicators and the differences in information contained in the sample data itself, this article adopts a combination of subjective and objective weighting to improve the stability and interpretability of the comprehensive evaluation results. Specifically, first, subjective weights of each indicator are obtained based on Delphi interviews and AHP judgment matrices. Then, entropy weights are calculated based on the dispersion of the sample on each indicator, and combined weights are formed to calculate the annual comprehensive score of the institution [17]. On this basis, the evaluation results will be further introduced into external criterion models and machine learning models to test the explanatory power of the framework for employment quality, enterprise satisfaction, and collaborative innovation output. Due to the inconsistent dimensions and value ranges of different indicators, it is necessary to first standardize the original data [18]. For positive indicators, this article adopts the range standardization method to convert each indicator into a comparable interval, as shown in formula (1).

$$z_{ij}=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \tag{1}$$

In formula (1), z_{ij} represents the standardized value of the i-th evaluation object on the j-th indicator. x_{ij} represents the original observation value of the i-th evaluation object on the j-th indicator. $\min(x_j)$ and respectively represent the minimum and maximum values of the jth indicator in all samples. On the basis of standardized results, further calculate the information entropy of each indicator in the sample. If a certain indicator has significant differences between different universities, the amount of information it contains is relatively higher, and it should receive higher attention in objective weighting [19]. The calculation of information entropy is shown in formula (2).

$$\begin{cases} e_j=-k\sum_{i=1}^{n} p_{ij}\ln p_{ij} \\ p_{ij}=\frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}} \\ k=\frac{1}{\ln n} \end{cases} \tag{2}$$

In formula (2), e_j represents the information entropy of the jth indicator. p_{ij} represents the proportion of the i-th evaluation object under the j-th indicator. z_{ij} represents the standardized value. k represents the normalization constant in information entropy calculation. n represents the number of samples. ln represents the natural logarithm function. The smaller the information entropy, the more significant the differences in this indicator between different samples, and the relatively larger the amount of information it provides [20]. Based on this, the objective weights of each indicator can be obtained, as shown in formula (3).

$$w_j^{ent}=\frac{1-e_j}{\sum_{j=1}^{m}(1-e_j)} \tag{3}$$

In formula (3), w_j^{ent} represents the objective weight of the jth indicator obtained based on the entropy weight method. e_j represents the information entropy of j indicators. m represents the total number of indicators. In order to balance institutional experience and data distribution characteristics, this article combines the two into a combined weight, sets the weight coefficient $\lambda=0.5$, and examines whether there is a significant change in the results under different λ values in subsequent robustness tests, as shown in formula (4).

$$w_j=\lambda w_j^{sub}+(1-\lambda)w_j^{ent} \tag{4}$$

In formula (4), w_j represents the combined weight of the jth indicator. w_j^{sub} represents the subjective weight of the jth indicator obtained through Delphi consultation and AHP method. After obtaining the combination weights, the comprehensive score of industry education integration for college i in year t can be calculated, as shown in formula (5).

$$S_{it}=\sum_{j=1}^{m} w_j z_{ijt} \tag{5}$$

In formula (5), S_{it} represents the comprehensive evaluation score of industry education integration for the i-th institution in the t-th year. z_{ijt} represents the standardized value of the i-th institution on the j-th indicator in the t-th year. The higher the score, the better the quality of industry education integration performance of the institution in the corresponding year. To test whether the score has practical explanatory power, this article further constructs an external criterion regression model to link the comprehensive score with key outcome variables. The external criterion regression model is shown in formula (6).

$$Y_{it}=\alpha+\beta S_{it}+\gamma X_{it}+\mu_i+\tau_t+\varepsilon_{it} \tag{6}$$

In formula (6), Y_{it} represents the external outcome variable of the i-th institution in the t-th year, which can be replaced by corresponding employment rate, average salary after graduation for six months, enterprise satisfaction, or joint innovation output. α represents a constant term. β represents the coefficient of influence of S_{it} on external outcome variables. X_{it} represents the control variable vector. γ represents the regression coefficient vector. μ_i represents the fixed effect of institutions, used to control for individual differences that do not change over time. τ_t represents the fixed year effect, used to control for external shocks that occur in different years. ε_{it} represents the random perturbation term.

In the AI empowerment evaluation section, this article takes five first level dimension scores and comprehensive scores as the core input features, uses LightGBM and XGBoost to predict external outcome variables, and compares them with the results of AHP only, entropy only, PCA, random forest, and linear regression. The reason for using boosting models is that they

can handle non-linear relationships and feature interactions well, while maintaining stable fitting ability for medium-sized samples [21]. In model training, 10 fold cross validation is used to reduce the occasional fluctuations caused by single partitioning, and MAE, RMSE, and R2 are uniformly reported as comparison indicators [22]. To further explain the basis for model judgment, this article introduces SHAP value analysis to analyze the marginal contributions of various dimensions and key indicators, in order to identify which factors are most closely related to the performance of high-quality industry education integration. With this design, the evaluation framework no longer remains at the level of comprehensive score ranking, but can further explain the main sources behind score differences and provide a basis for subsequent grouping tests and policy discussions.

2.3 Reliability, validity, and robustness strategy

To ensure the measurability, interpretability, and generalizability of the evaluation framework, this paper conducts tests from three levels: measurement quality, external criterion validation, and robustness testing [23]. This approach can avoid research being limited to internal consistency within the scale or model fitting results, making the comprehensive score both a measurement basis and a response to real educational performance. The AI powered evaluation model and three-level validation framework are shown in Figure 2.

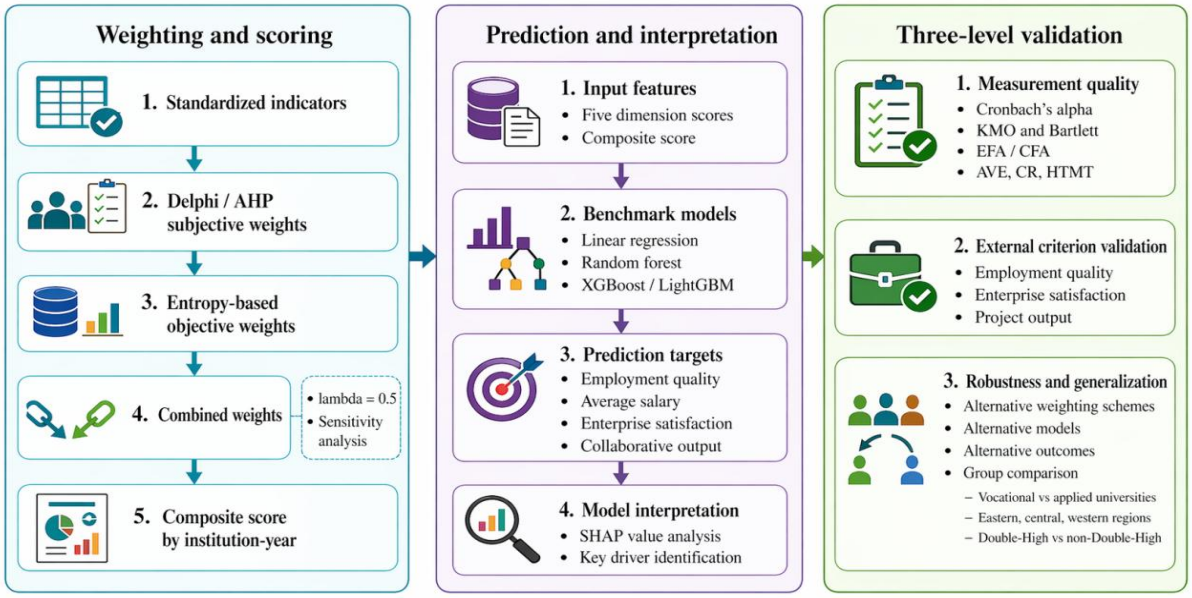

Figure 2: Hybrid weighting, AI prediction, and three-level validation framework.

As shown in Figure 2, the study first forms a comprehensive evaluation score based on subjective and objective weighting, and then inputs the dimension score and total score into traditional and machine learning models to predict external outcome variables, and tests the effectiveness of the framework from three levels: measurement quality, external criteria, and robustness. The first level of inspection focuses on measuring quality. This article first uses Cronbach's alpha coefficient to check the internal consistency of each dimension and the overall scale, and combines corrected item total correlation to screen the items. Subsequently, Kaiser Meyer Olkin measure and Bartlett's test of sphericity were used to determine whether the sample was suitable for factor analysis [24]. In the structure recognition stage, Exploratory Factor Analysis is first used to explore the aggregation relationship of indicators, and then Confirmatory Factor Analysis is used to test the fitting of the five dimensional structure.

Furthermore, this article reports Average Variance Extracted, Composite Reliability, and Heterotrait Monotrait Ratio to examine aggregation validity, combination reliability, and discriminant validity. Through the above program, it can be determined whether each observed variable can stably reflect potential dimensions such as collaborative governance, curriculum co construction, integration of practical platforms and resources, process quality and learning investment, outcome effectiveness and sustainable development.

The second level of testing focuses on whether the comprehensive evaluation score can correspond to the true results. This article links the annual comprehensive score of universities with external variables such as the corresponding employment rate, average salary in six months after graduation, enterprise satisfaction, and output of joint projects and horizontal projects, to examine whether high scoring universities perform better in key educational outcomes. If the evaluation results only perform well at the internal measurement level but cannot explain the differences in employment quality and cooperative output, their application value will be limited. Therefore, this article simultaneously tests the effectiveness of the composite score in both fixed effects regression and predictive model analyses, and compares the explanatory differences between the primary dimension score and the total score on external results to identify which dimensions are more closely related to outcome variables.

The third level of testing is used to examine the stability and generalization ability of the conclusions. Firstly, in terms of weighting methods, Analytic Hierarchy Process, entropy weighting, and equal weighting were used to recalculate the scores of universities, and the ranking changes and estimation results under different weight settings were compared [25]. Secondly, at the model level, linear regression, random forest, and Extreme Gradient Boosting are used for prediction, and the model performance is evaluated through Mean Absolute Error, Root Mean Squared Error, and coefficient of determination. Again, at the level of outcome variables, replace them with employment rate, salary level, and enterprise satisfaction in sequence, and observe whether the core conclusions remain consistent. Finally, this article conducts grouping tests based on vocational colleges and applied undergraduate programs, the eastern, central, and western regions, as well as "double high" colleges and non double high "colleges to determine the applicability of the evaluation framework in different types of educational institutions and regional contexts. If the main conclusions maintain similar direction and significance in alternative weights, alternative models, and different sample groups, it indicates that the framework has good robustness and cross contextual applicability.

3 Results and Discussion

3.1 Descriptive characteristics and measurement quality

Before testing the predictive ability of the framework, two fundamental questions need to be confirmed: whether the sample structure can support comparative analysis, and whether the measurement quality of the evaluation framework has reached an acceptable level. This article ultimately includes 42 colleges, including 20 vocational colleges and 22 applied undergraduate colleges, covering 15 in the eastern region, 13 in the central region, and 14 in the western region, forming a total of 126 institution year observations from 2022 to 2024. In terms of professional categories, there are 11 intelligent manufacturing colleges, 9 electronic information colleges, 8 modern service colleges, 7 finance and commerce colleges, and 7 medical education and public service colleges. Overall, the sample retained necessary differences in terms of region, educational level, and professional direction, providing a stable foundation for subsequent comparisons. The sample composition results are shown in Figure 3.

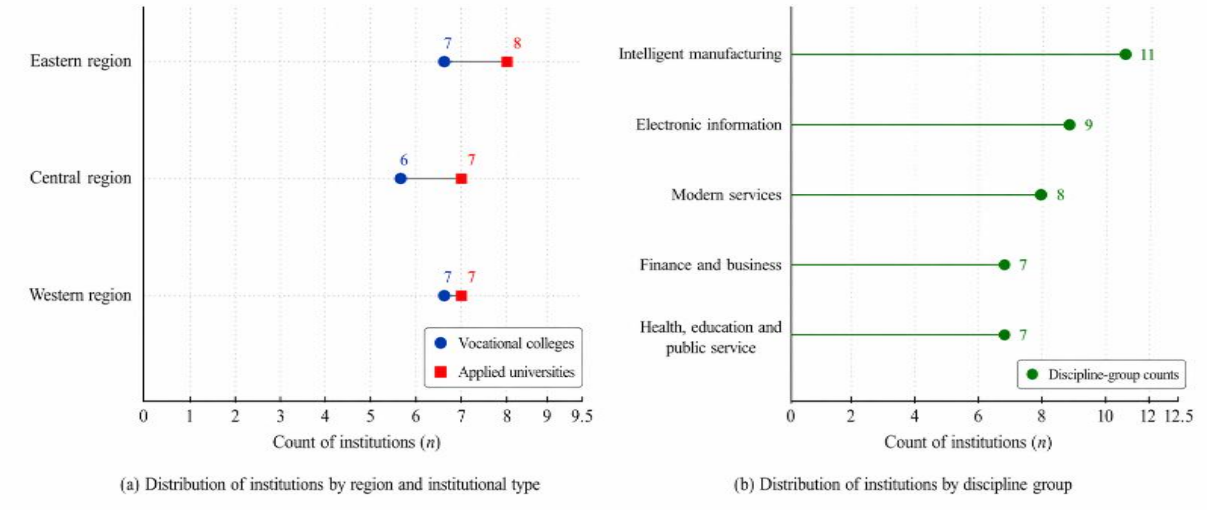

Figure 3: Sample structure of the study dataset.

Fig. 3 (a) shows that there are 7 and 8 vocational colleges and applied undergraduate programs in the eastern region, 6 and 7 in the central region, and 7 in the western region. Both types of colleges are distributed in the three major regions, indicating that the sample has not been overly concentrated towards any specific region or level of education. Fig. 3 (b) further indicates that the number of colleges and universities in the fields of intelligent manufacturing and electronic information is the highest, reaching 11 and 9 respectively, with 8 in the field of modern services, and 7 in the fields of finance, commerce, medical education, and public services. This distribution is basically consistent with the current active practice of industry education integration in professional fields. From this, it can be seen that the sample in this article has both regional coverage and retained differences in disciplinary structure, thus meeting the basic conditions for conducting horizontal comparisons and grouping tests between universities. After confirming the sample structure, it is also necessary to observe the descriptive features of the first level dimension to determine the distribution level and fluctuation amplitude of different dimensions. The relevant results are shown in Figure 4.

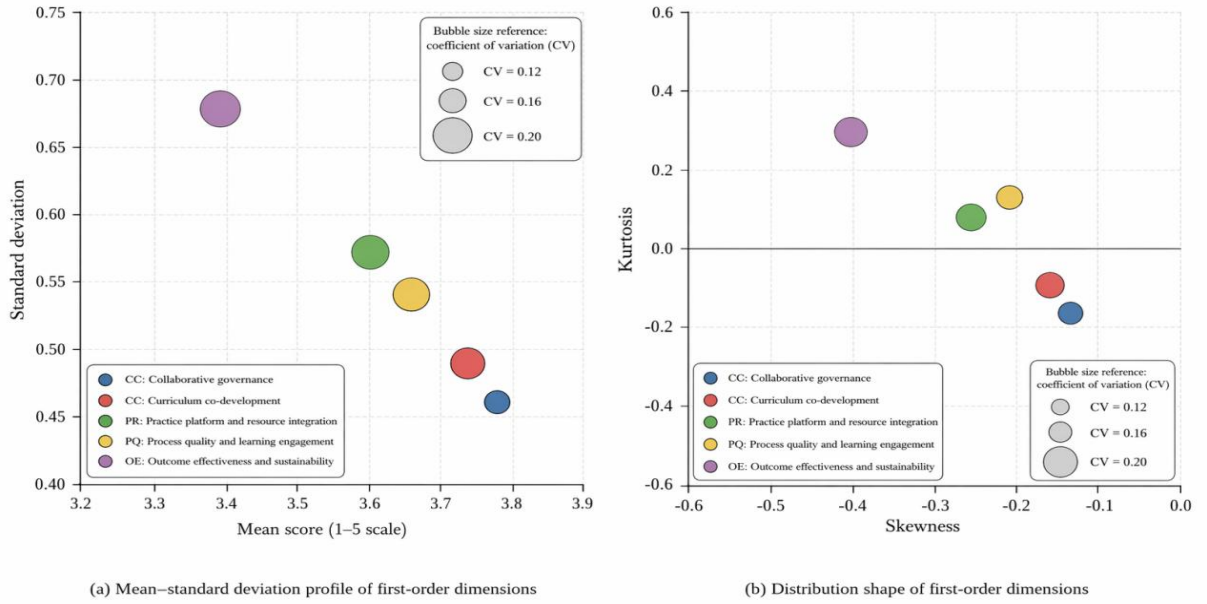

Figure 4: Descriptive characteristics of first-order dimensions.

Fig. 4 (a) shows that the mean of Collaborative governance is the highest, at 3.78, with a standard deviation of 0.46. The Curriculum co-development is 3.74 ± 0.49. The process quality and learning engagement is 3.66 ± 0.54. The practice platform and resource integration is 3.61 ± 0.57. The mean of Outcome Effectiveness and Sustainability is the lowest, only 3.42, but the standard deviation reaches 0.68, which is the highest among the five dimensions. Fig. 4 (b) shows that the skewness of each dimension ranges from -0.48 to -0.21, the kurtosis ranges from -0.18 to 0.34, and the overall distribution is within an acceptable range. Figure 4 reflects two important phenomena. Firstly, the average of collaborative governance and curriculum co construction ranks among the top, indicating that most universities have formed a certain foundation in institutional and curriculum collaboration. Secondly, the degree of dispersion in the outcome dimension is the highest, indicating that more significant differences between universities mainly occur in the outcome aspects such as employment quality, enterprise satisfaction, and sustained cooperation output, rather than staying at the level of resource allocation. Merely describing statistics is not enough to prove the validity of the framework, further investigation is needed to assess its reliability, validity, and ability to distinguish groups. The corresponding results are shown in Figure 5.

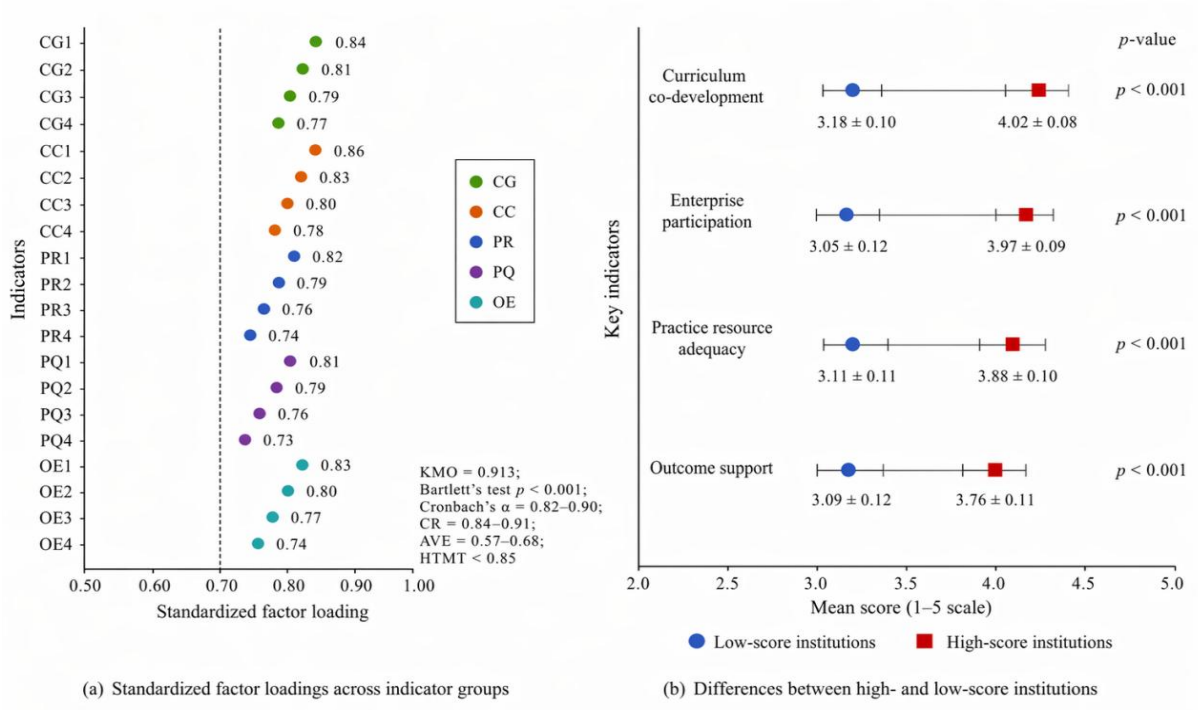

Figure 5: Measurement quality evidence and group-based validation results.

Fig. 5 (a) shows that the Cronbach's alpha of the five constructs ranges from 0.82 to 0.90, Composite Reliability ranges from 0.84 to 0.91, Average Variance Extracted ranges from 0.57 to 0.68, and the standardized factor loadings are distributed between 0.73 and 0.86; At the same time, the Kaiser Meyer Olkin value was 0.913, Bartlett's test of sphericity reached a significant level ($p<0.001$), and HTMT was less than 0.85, indicating that the framework is stable in terms of internal consistency, convergent validity, and discriminant validity. More noteworthy is that the overall load of indicators related to collaborative governance and curriculum co construction is higher, indicating that the core support of this evaluation framework comes from institutionalized collaboration, joint curriculum development, and sustained enterprise participation, rather than simply relying on equipment investment or platform quantity. Furthermore, Fig. 5 (b) shows the differences in key indicators between the high and low groups.

The mean values of Curriculum co-development were 4.02 ± 0.08 and 3.18 ± 0.10, Enterprise participation was 3.97 ± 0.09 and 3.05 ± 0.12, Practice resource adequacy was 3.88 ± 0.10 and 3.11 ± 0.11, and Outcome support was 3.76 ± 0.11 and 3.09 ± 0.12, respectively. All four comparisons reached a significant level ($p<0.001$). This result indicates that the evaluation framework constructed in this article is not only stable at the measurement level, but also can effectively identify the substantive differences between high-quality and low-quality colleges in terms of curriculum co construction, enterprise participation, and practical support.

Overall, the results of this section support two judgments. Firstly, the sample structure has a good comparative foundation and can support subsequent model estimation. Secondly, the evaluation framework has achieved a good level of reliability, validity, and inter group discrimination ability. Especially with the high load of collaborative governance and curriculum co construction, as well as greater fluctuations in the dimension of results and effectiveness, it indicates that the key differences in the quality of industry education integration are mainly reflected in the depth of institutional collaboration and its ability to transform into results. This also provides a necessary prerequisite for testing the predictive performance of the framework and the effectiveness of external metrics in the next section.

3.2 Predictive performance and external validity of the framework

To test whether the framework is truly superior to traditional scoring methods, this paper first compared the performance of six types of models in external outcome prediction tasks, including AHP only, entropy only, PCA score, linear regression, random forest, and LightGBM based framework. The core results are shown in Figure 6.

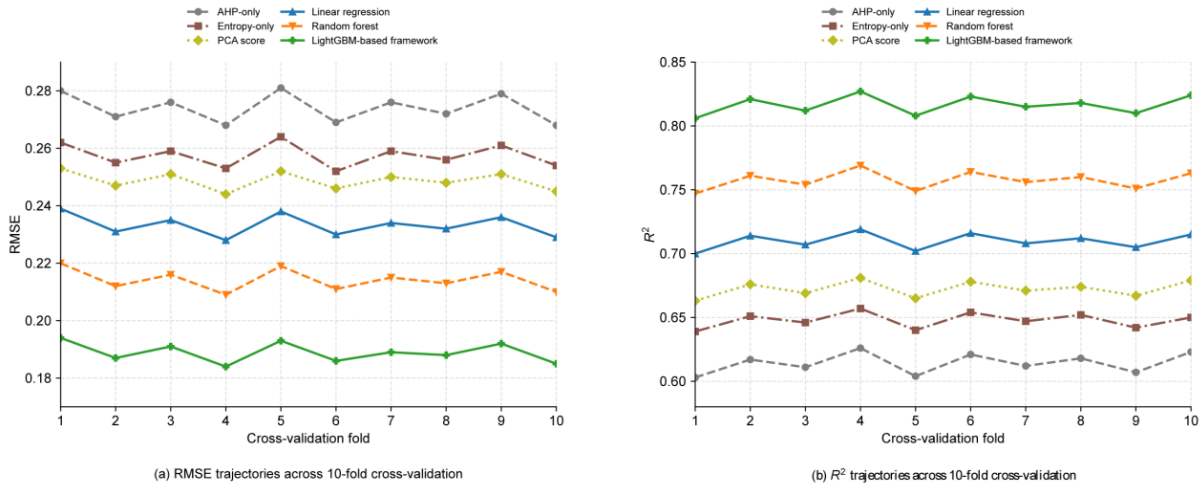

Figure 6: Predictive performance of alternative evaluation approaches.

Figure 6 (a) shows the RMSE variation trajectories of each model under 10 fold cross validation, while Figure 6 (b) displays the corresponding R2 variation. Overall, the LightGBM based framework maintains the lowest error and highest interpretability among the ten folds, with minimal curve fluctuations, indicating that it not only has higher accuracy but also better stability. According to the average results, the MAE, RMSE, and R2 of AHP only are 0.212, 0.274, and 0.612, respectively. Entropy only values are 0.198, 0.258, and 0.648; The PCA scores are 0.191, 0.249, and 0.671. Linear regression values are 0.179, 0.233, and 0.708; The random forest values are 0.162, 0.214, and 0.756. The LightGBM based framework further reduces to 0.138 and 0.189, and increases R2 to 0.814. In the extended classification task of high and low-quality universities, the AUC of the LightGBM based framework is 0.912, and the F1 is 0.846,

which is significantly higher than the 0.873 and 0.801 of the random forest. From this, it can be seen that the comprehensive framework empowered by AI can more fully characterize the nonlinear relationships and interaction effects between indicators, thus achieving stronger generalization ability in external result prediction. The prediction accuracy alone is not enough to demonstrate the practical significance of the evaluation framework, and it is necessary to verify whether there is a stable relationship between it and key external indicators. The relevant results are shown in Figure 7.

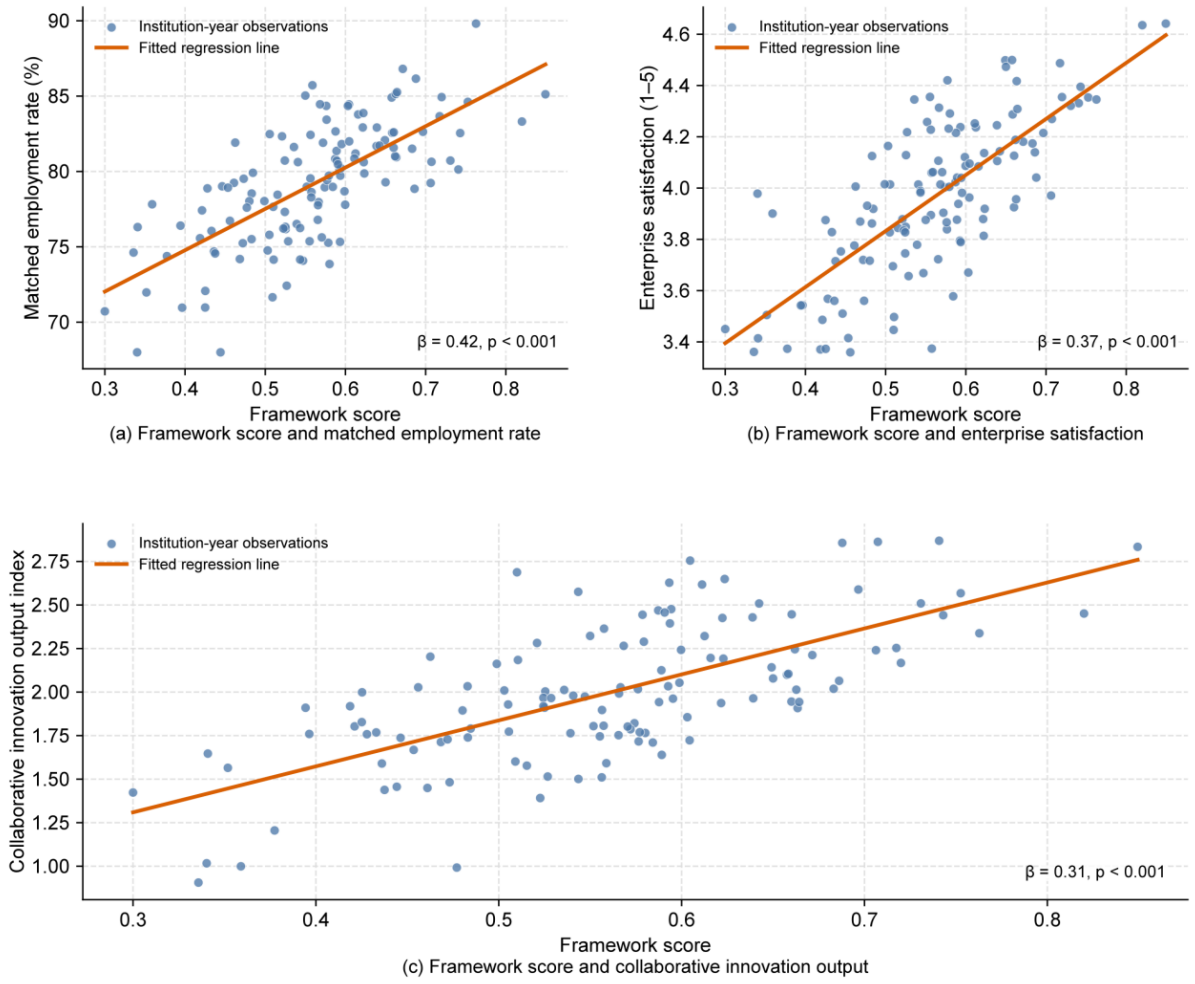

Figure 7: External validity of the AI-empowered evaluation framework.

Figure 7 (a) shows a significant positive relationship between the evaluation score and the corresponding employment rate. Figure 7 (b) shows that the higher the evaluation score, the higher the satisfaction of the enterprise. Figure 7 (c) shows that the evaluation score and collaborative innovation output also exhibit significant changes in the same direction. After controlling for regional economic level, university scale, financial investment, and professional structure, the regression coefficient of the framework score for the corresponding employment rate was 0.42 ($p<0.001$), for enterprise satisfaction was 0.37 ($p<0.001$), and for joint projects and horizontal output index was 0.31 ($p<0.001$). These results indicate that the framework can not only predict 'whose fusion quality is higher', but also explain 'why high scoring universities perform better on the outcome side'.

The mechanism behind this relationship is also relatively clear. The reason why curriculum co construction can improve the quality of employment is that the needs of enterprises are

embedded in the curriculum system earlier, and the ability structure obtained by students is closer to the job requirements, thus shortening the distance from learning to job adaptation. The continuous participation of dual qualified teachers and enterprise mentors will enhance the authenticity of case studies, the pertinence of training, and the timeliness of evaluation feedback, which will make enterprises have a more positive perception of the talent cultivation process and thus improve satisfaction. The more abundant the practical platform and training resources are, the easier it is for students to transform classroom knowledge into actionable abilities, and thus demonstrate higher conversion efficiency in competition results, joint projects, and horizontal cooperation. Combining Figures 6 and 7, it can be concluded that the advantage of this framework is not only reflected in the improvement of statistical fitting, but also in its ability to connect intermediate mechanisms such as institutional collaboration, curriculum collaboration, and practical support with employment, satisfaction, and innovation output, thereby enhancing the explanatory power and application value of evaluation results.

3.3 Interpretability, heterogeneity, and robustness analysis

After confirming the framework's good predictive accuracy and external effectiveness, three further questions need to be answered: what key factors does the model rely on to make judgments, whether different types of universities exhibit consistent patterns, and whether the conclusion still holds after replacing weights, models, and outcome variables. The interpretability analysis results are shown in Figure 8.

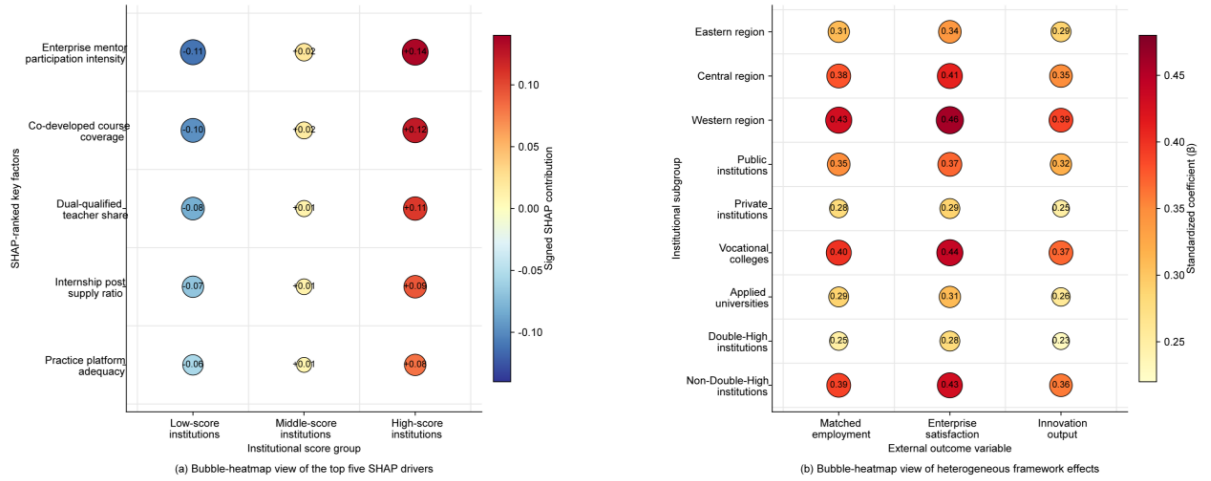

Figure 8: Interpretability and heterogeneous effects of the AI-empowered framework.

Figure 8 (a) shows the key driving factors for the top five SHAP rankings. The average absolute SHAP value of Enterprise mentor participation intensity is the highest, at 0.184, Co developed course coverage is 0.167, Dual qualified teacher share is 0.151, Internship post supply ratio is 0.139, and Practice platform proficiency is 0.126. This ranking indicates that the core variables determining the level of high-quality industry education integration are concentrated in the deep participation of enterprises in the curriculum and practical process, rather than relying solely on platform and equipment conditions. It is particularly noteworthy that the participation intensity of enterprise mentors and the coverage of course co construction are significantly higher than the adequacy of the platform, which is not entirely consistent with the traditional research that emphasizes hardware investment. This result indicates that the AI model is more sensitive in identifying the transmission effect of process collaborative variables on the performance of the outcome side, which also constitutes an important highlight of this

paper. The heterogeneity results are shown in Figure 8 (b). From a regional perspective, the standardized coefficients of the framework score on the outcome variables were 0.34, 0.41, and 0.46 in the eastern, central, and western regions, respectively, with the most significant marginal improvement observed in western universities; From the perspective of educational attributes, public institutions and private institutions have respective values of 0.37 and 0.29. From the perspective of college types, vocational colleges and applied undergraduate programs are 0.44 and 0.31, respectively. From the perspective of policy attributes, "double high" universities have a score of 0.28, while non double high "universities have a score of 0.43. It can be seen that the benefits brought by the improvement of the quality of industry education integration are not exactly the same among all universities. For colleges and universities with weak foundations and immature governance mechanisms, institutional collaboration and enterprise participation bring greater incremental benefits. In colleges with better resource foundations, the room for improvement is relatively narrow. Therefore, regional and institutional differences mean that policy tools should not be completely homogenized. Western regions, vocational colleges, and non "double high" colleges need to focus on targeted governance around curriculum co construction, participation of enterprise mentors, and job supply capabilities. The robustness test results are shown in Figure 9.

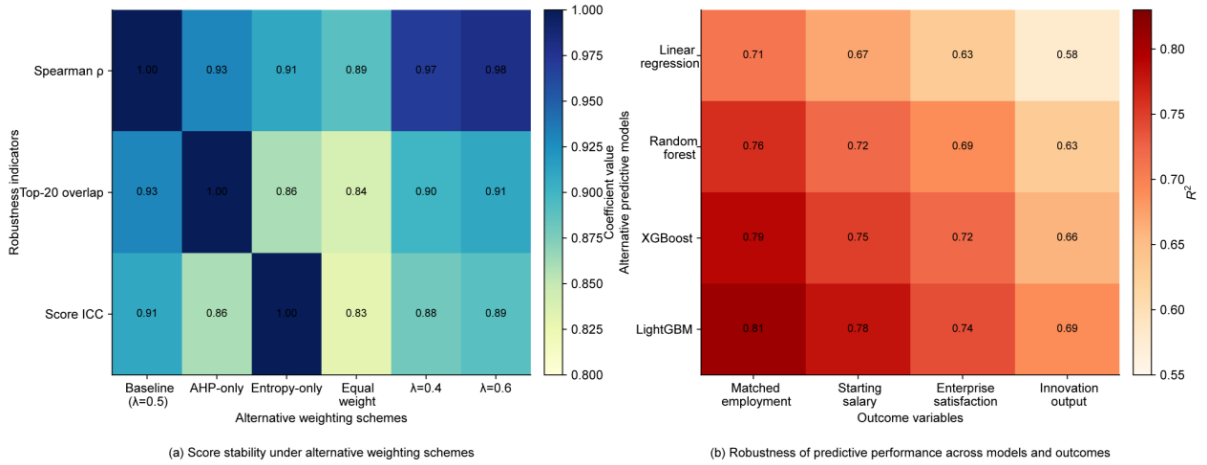

Figure 9: Robustness of the evaluation framework under alternative specifications.

Figure 9 (a) shows that when using AHP only, entropy only, equal weight, and different combination weight parameters, the Spearman correlation coefficient between the benchmark score and the alternative solution remains above 0.89, the Top-20 overlap is not less than 0.84, and the Score ICC also remains above 0.88, indicating that the overall ranking of colleges and universities is relatively stable. Figure 9 (b) further shows that after replacing the predictive model and outcome variable, the explanatory power of the framework remains at a good level. Taking R2R ^ 2R2 as an example, LightGBM achieved R2R ^ 2R2 values of 0.81, 0.78, 0.74, and 0.69 on the four outcome variables of matched employment, starting salary, enterprise satisfaction, and innovation output, respectively, which are higher than linear regression and random forest. The results of XGBoost are also similar, indicating that the conclusion does not rely on a single algorithm setting.

From the comprehensive analysis of Figures 8 and 9, it can be seen that the advantages of this framework are mainly reflected in three points. Firstly, the AI model identifies the deep involvement of enterprises in curriculum and practical processes as the most critical driving factor, which gives the evaluation results strong explanatory power. Secondly, the heterogeneity results indicate that the improvement paths corresponding to different regions and types of

universities are not the same. Again, after replacing the weighting scheme, prediction model, and outcome variables, the main conclusion remains stable. It can be seen that this framework is not only stable at the statistical level, but also provides more targeted evidence support for differentiated governance.

4 Conclusion

Overall, this article demonstrates that AI enabled evaluation frameworks can more accurately characterize the quality of industry education integration and enhance the interpretability of results.

(1) The framework consists of five dimensions: collaborative governance, curriculum co construction, integration of practical platforms and resources, process quality and learning investment, and outcome effectiveness and sustainability. It systematically answers the question of "which dimensions characterize the quality of industry education integration".

(2) AI has significantly improved the effectiveness of evaluations. The MAE and RMSE of the LightGBM based framework have been reduced to 0.138 and 0.189, respectively, while R2R ^ 2R2 has been improved to 0.814, which is superior to traditional weighted and conventional models.

(3) In theory, this article will advance the evaluation of industry education integration from static indicator aggregation to an interpretable and externally verifiable analytical framework. In practice, it can provide multi-dimensional diagnosis and differentiated improvement basis for colleges, enterprises, and education authorities.

However, there are still certain limitations to this study, as the research sample mainly consists of Chinese universities, and its cross-border applicability still needs to be tested. In addition, some outcome variables still rely on school annual reports. Therefore, in future work, text mining, enterprise visits and job expansion records, course platform behavior data, and graduate tracking panels can be introduced to further improve the dynamic and external validity of the framework.

About the Author

Xiaoyan Yu was born in 1979 in Lin'an District, Hangzhou City, Zhejiang Province. She obtained her bachelor's degree from Hangzhou Normal University and her master's degree from East China Normal University. Currently, she is employed at ZheJiang Institute of Communications. Her primary research focus includes industry-education integration management in vocational education.

Jun Shen was born in 1975 in Xiaoshan District, Hangzhou City, Zhejiang Province. He obtained his bachelor's degree from Shanghai University of Sport and his master's degree from East China Normal University. Currently, he is employed at ZheJiang Institute of Communications. His primary research focuses on vocational education, physical education, and training.

References

[1] Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252, 124167. DOI: 10.1016/j.eswa.2024.124167.

[2] Rahiman, H. U., & Kodikal, R. (2024). Revolutionizing education: Artificial intelligence empowered learning in higher education. Cogent Education, 11(1), 2293431. DOI: 10.1080/2331186X.2023.2293431.

[3] Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. Computers and Education: Artificial Intelligence, 5, 100165. DOI: 10.1016/j.caeai.2023.100165.

[4] Laupichler, M. C., Aster, A., & Raupach, T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. Computers and Education: Artificial Intelligence, 4, 100126, 1–10. DOI: 10.1016/j.caeai.2023.100126.

[5] Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2023). Evaluating an artificial intelligence literacy programme for developing university students' conceptual understanding, literacy, empowerment and ethical awareness. Educational Technology & Society, 26(1), 16–30. DOI: 10.30191/ETS.202301_26(1).0002.

[6] Chai, C. S., Yu, D., King, R. B., & Zhou, Y. (2024). Development and validation of the Artificial Intelligence Learning Intention Scale (AILIS) for university students. SAGE Open, 14(2), Article 21582440241242188. DOI: 10.1177/21582440241242188.

[7] Cabero-Almenara, J., Palacios-Rodríguez, A., Loaiza-Aguirre, M. I., & Rivas-Manzano, M. R. de. (2024). Acceptance of Educational Artificial Intelligence by Teachers and Its Relationship with Some Variables and Pedagogical Beliefs. Education Sciences, 14(7), 740. DOI: 10.3390/educsci14070740.

[8] He, Z., Chen, L., & Zhu, L. (2023). A study of Inter-Technology Information Management (ITIM) system for industry-education integration. Heliyon, 9(9), e19928. DOI: 10.1016/j.heliyon.2023.e19928.

[9] Zheng, W., Zheng, X., & Zhu, X. (2024). Promoting integration of industry and vocational education: Exploring stakeholder intentions of hydrogen energy industry. International Journal of Hydrogen Energy, 52, 454–464. DOI: 10.1016/j.ijhydene.2023.06.072.

[10] Gong, X. (2024). Performance evaluation of industry-education integration in higher education from the perspective of coupling coordination—an empirical study based on Chongqing. PLOS ONE, 19(9), e0308572. DOI: 10.1371/journal.pone.0308572.

[11] Zhao, X. (2024). Study on the matching degree of major groups and industrial groups in higher vocational colleges. Heliyon, 10(8), e29945. DOI: 10.1016/j.heliyon.2024.e29945.

[12] Martín-Martín, A. O., Bañuls, V. A., & Ruiz-Benítez, R. (2023). Technology Transfer Assessment in Regional Business Contexts. Sustainability, 15(15), 11680. DOI: 10.3390/su151511680.

[13] Chen, Y., Li, S., & Chen, R. (2025). Impact of Industry and Education Integration on Employment Quality in Higher Vocational Colleges: Moderating Role of Faculty Qualifications and Curriculum Development Capacity. Education Sciences, 15(10), 1316. DOI: 10.3390/educsci15101316.

[14] Zhu, Z. (2025). Building a Digital-Enhanced I&E Curriculum Through Industry-Education Integration. Information Resources Management Journal, 38(1), 1–14. DOI: 10.4018/IRMJ.395340.

[15] Chen, C., Zhang, J., Du, A. M., & Li, Z. (2025). University-industry collaboration and enterprise total factor productivity. International Review of Economics & Finance, 102, 104311. DOI: 10.1016/j.iref.2025.104311.

[16] Zhuang, T., Oh, M., & Kimura, K. (2025). Modernizing higher education with industrial forces in Asia: a comparative study of discourse of university-industry collaboration in China, Japan and Singapore. Asia Pacific Education Review, 26(1), 195–210. DOI: 10.1007/s12564-024-10033-y.

[17] Ranieri, M., Biagini, G., & Cuomo, S. (2025). AI Literacy in Higher Education: A Systematic Approach to Questionnaire Development and Validation. International Journal of Digital Literacy and Digital Competence, 16(1), 1–25. DOI: 10.4018/IJDLDC.388469.

[18] Lan, M., & Zhou, X. (2025). A qualitative systematic review on AI empowered self-regulated learning in higher education. npj Science of Learning, 10, Article 21. DOI: 10.1038/s41539-025-00319-0.

[19] Schmidt, D. A., Alboloushi, B., Thomas, A., & Magalhaes, R. (2025). Integrating artificial intelligence in higher education: perceptions, challenges, and strategies for academic innovation. Computers and Education Open, 9, 100274. DOI: 10.1016/j.caeo.2025.100274.

[20] Setiyawan, A., Soeharto, S., Wijaya, T. T., Korenova, L., & Lavicza, Z. (2025). Measuring Teachers' competencies for AI integration: Development and validation of the AI-TPACK in vocational education. Computers and Education Open, 9, 100319. DOI: 10.1016/j.caeo.2025.100319.

[21] Tan, X., Cheng, K. S., & Ling, M. H. A. (2025). Enhancing teachers' AI competency: A professional development intervention study based on intelligent-TPACK framework. Computers and Education: Artificial Intelligence, 9, 100521. DOI: 10.1016/j.caeai.2025.100521.

[22] Ahmad, Z., Sultana, A., Abdul Latheef, N., Siby, N., Sellami, A., & Abbasi, S. A. (2025). Measuring students' AI competence: Development and validation of a multidimensional scale integrating educational psychology perspectives. Acta Psychologica, 259, 105446. DOI: 10.1016/j.actpsy.2025.105446.

[23] Xu, L., Zhang, L., Ou, L., & Wang, D. (2025). The development and validation of a scale on student AI literacy in L2 writing: A domain-specific perspective. Journal of Second Language Writing, 69, 101227. DOI: 10.1016/j.jslw.2025.101227.

[24] Faraon, M., Rönkkö, K., Milrad, M., & Tsui, E. (2025). International perspectives on artificial intelligence in higher education: An explorative study of students' intention to use ChatGPT across the Nordic countries and the USA. Education and Information Technologies, 30, 17835–17880. DOI: 10.1007/s10639-025-13492-x.

[25] Madanchian, M., & Taherdoost, H. (2025). Decision-making criteria for AI tools in digital education. Digital Engineering, 7, 100069. DOI: 10.1016/j.dte.2025.100069.