Big Data-Supported Analysis Model for Matching Vocational School Majors with Industry Needs

Jing Chen
School of Biomedicine and Pharmacy, Jiangxi Management Vocational College, Nanchang 330200, Jiangxi, China
Corresponding author: Jing Chen (e-mail: chenj1990@163.com)

ABSTRACT The dynamic adjustment of vocational-school majors requires quantitative modeling of the relationship between professional supply and industrial skill demand. Conventional experience-based judgments and static statistics cannot capture rapid changes in labor-market requirements. This study proposes a big-data-supported analysis model for matching vocational majors with industry needs. Recruitment data and curriculum-system data are collected and transformed into a unified skill-tagging space. Natural-language-processing methods extract job-skill requirements from recruitment texts, and TF-IDF weighting is used to construct industry demand vectors. Professional curriculum features are mapped into supply-capability vectors according to course hours, course type, and competency objectives. A weighted cosine-similarity model and multi-indicator fusion mechanism are then built to evaluate skill matching, job coverage, and demand-intensity matching. A time-series updating mechanism further tracks dynamic changes in matching degree. Experimental results based on recruitment postings and vocational-college curriculum data show clear differentiation among majors: big data technology and nursing achieve high matching degrees of 0.85 and 0.83, while accounting shows a structural mismatch with a matching degree of 0.68. The proposed model supports semantic vectorization, industrial-demand sensing, and data-driven decision support for professional-structure optimization.

INDEX TERMS big data analysis, semantic vectorization, skill matching, time-series modeling, decision support

I. INTRODUCTION

As the industrial structure is being increasingly upgraded and the digital economy is constantly being developed, the demand of skilled talents in the labor market has become more diverse and dynamic. Vocational schools are at the forefront of reforming their academic structures, shifting from quantitative expansion toward qualitative optimization by adopting the scale expansion to structural optimization [1,2]. Nevertheless, the existing work environments are more dependent on the judgment of experience and the use of unchanged statistics, without the real-time perception and deep mining of industry demand and hence the structural imbalance between professional supply and job demand [3]. To address the issues raised above, this paper develops a corresponding level of analysis model within the vocational school professional environment and the industry demand using the methods of big data analysis. It combines employment recruitment data and professional curriculum system data to obtain industry demand attributes and develop a single skill label system to achieve the vectorized representation of an industry demand and professional supply. By this, one comes up with a corresponding degree calculation method which is based on weighted similarity and a multiindicator fusion evaluation model is built combining job coverage and demand intensity.

II. RELATED WORK

Within the context of the increasing convergence between industry and education and the service of industrial growth in vocational education, there has been a multi-dimensional debate on the correspondence of professional environments and industrial demands in vocational colleges conducted by the academic community, which creates a system of research between the macro structure and micro practice. An et al. has identified that the professional environments of higher vocational colleges are contributing factors to the industrial development and talent training and analyzed the logic of matching between majors and industries systematically in three levels, namely, the macro industrial layout, meso market demand and micro technology demand. They suggested enhancing flexibility through the further penetration of industry and education, streamlining professional groups and their structure, as well as enhancing the technological innovation [4]. It is on this basis that Wang also unearthed the underlying issues behind the imbalance between industrial demand and human capital supply on the adaptive systems perspective. He considered that the colleges and universities had structural issues related to the transformation of the mode of knowledge production, the lack of digital level and the gap between the world of professional activities and the demands of the occupation, and suggested the ways of optimization through the prism of data governance and the institutional mechanisms [5]. In contrast to the macro and institutional level analysis, Zou Hexu paid attention to the training process of talents, integrated engineering education certification with industry college building, and achieved the dynamism of coordinating and adjusting training and industrial exigencies by building a goal-courses-processes-evaluation complex, and thus enhancing the flexibility of the talent training process [6]. At the same time, Cheng and Yang used the field of additive manufacturing as an example and stressed that the deep integration of education chain and industry chain should be encouraged with the help of reconstruction of the curriculum system, building of school-enterprise collaboration education and practice platform[7]; Zhang discussed the issues of the practice teaching system of tourism management major in terms of specific majors and offered to reorganize it according to the needs of enterprises in order to reconstruct it based on enterprise needs in order to improve the quality of talent training[8].

Although the studies mentioned above have presented a valuable source of references to the optimization of the professional environment under various lenses, they are mostly qualitative studies and empirical generalizations. They do not have a cohesive quantitative formulation and precise measurement of the corresponding relationship between professional supply and industry demand, especially in the fields of multi-source data fusion, dynamic change characterization, and determination of matching deviations. While existing domestic studies provide a foundation for understanding the macro-logic of industry-education integration, they often lack the algorithmic granularity found in international labor market analytics, such as the skill-gap taxonomies proposed by the OECD and the automated job-matching frameworks utilized in European vocational systems. Thus, there is a need to expand the current studies by synthesizing these disparate qualitative findings into a cohesive quantitative framework that utilizes multi-source data fusion to address the static limitations of previous Chinese-language research.

III. METHODS

A. METHODS FOR EXTRACTING AND QUANTIFYING INDUSTRY DEMAND CHARACTERISTICS

1) Method for Semantic Parsing of Job Requirements Text

For unstructured data such as recruitment texts, the first step is to perform word segmentation, representing the text as a sequence of words $T = w_1, w_2, \ldots, w_n$. Based on this, part-of-speech tagging is used to select a set of keywords with actual semantic value, such as core word classes like nouns and verbs, thus obtaining a set of key feature words (K).

To quantify the importance of keywords, the TF-IDF method is used to weight terms, and the calculation formula is as follows:

$$\mathrm{TF}-\mathrm{IDF}(w_i) = \mathrm{TF}(w_i) \times \log \frac{N}{\mathrm{DF}(w_i)} \tag{1}$$

Where $(\mathrm{TF}(w_i))$ represents the word frequency, $(\mathrm{DF}(w_i))$ represents the number of documents containing the word, and (N) is the total number of documents. This method can effectively identify core skill words and high-discrimination features in job descriptions.

2) Vectorized Representation Model of Skill Demand

After extracting keywords, a unified skill tagging system is constructed, mapping job requirements into a vector space representation. Let the skill set be $(S = s_1, s_2, \ldots, s_m)$, then each job can be represented as a skill requirement vector (D), where each dimension corresponds to the demand intensity of a skill. The values of each dimension in the vector can be obtained by summing the keyword weights, for example, by weighted summation based on TF-IDF weights:

$$d_j = \sum_{i=1}^{n} \delta(w_i, s_j) \cdot \mathrm{TF}-\mathrm{IDF}(w_i) \tag{2}$$

This $(\delta(w_i, s_j))$ indicates the attribution relationship between terms and skill tags.

3) Structural Modeling Method for Industry Demand

Based on vectorization, industry demand is modeled in a structured manner, comprehensively depicting it from three dimensions: industry category, job level, and skill weight. Specifically, industry demand can be represented as a triple, used to uniformly describe the characteristic structure of different jobs.

$$I = (C, L, D) \tag{3}$$

Where (C) represents the industry category, (L) represents the job level, and (D) is a skill demand vector.

In practical modeling, the importance of skill requirements varies across different job levels. Therefore, a level-based weighting factor is introduced to adjust the demand vector, thereby reflecting job differences. Furthermore, different industries also have different skill preferences, which can be further refined using an industry-based weighting factor to make the model more adaptable and interpretable.

B. PROFESSIONAL SUPPLY CAPACITY MODELING METHOD

1) Extraction of Features of Professional Curriculum System

First, extract course-related characteristics from the professional talent training program, including course type (theoretical courses, practical courses), class hour distribution, and corresponding competency objectives. Let the course set be:

$$C = c_1, c_2, \ldots, c_n \tag{4}$$

Each course can be mapped to a corresponding competency dimension and weight. By analyzing the relationship between course content and training objectives, key variables that reflect professional characteristics can be extracted.

To quantify the importance of a course, a course weight is introduced w_i, which can be determined by the percentage of class hours, course level, and course type.

$$w_i = \alpha \cdot \frac{h_i}{\sum h_i} + \beta \cdot t_i \tag{5}$$

Where (h_i) is the course hours, (t_i) is the course type coefficient (e.g., practical courses are given higher weight), (α) and (β) are adjustment parameters.

2) Construction of Professional Competency Output Vector

Based on course feature extraction, courses are mapped to competency indicators to construct a unified competency expression space. Let the set of competency indicators be:

$$A = a_1, a_2, \ldots, a_m \tag{6}$$

Each major can then be represented as a capability vector:

$$P = (p_1, p_2, \ldots, p_m) \tag{7}$$

The value of the (j)th ability dimension is obtained by accumulating the contributions of the relevant courses:

$$p_j = \sum_{i=1}^{n} w_i \cdot \phi(c_i, a_j) \tag{8}$$

of $(\phi(c_i, a_j))$ the course (c_i) to the ability (a_j) can be determined by the degree of matching of course content or expert evaluation.

3) Professional Structure Feature Representation Model

Based on the capability vector, the major is modeled from the perspective of overall structure, comprehensively depicting the characteristics of the major such as scale, hierarchical structure and development direction.

First, the size of a program can be represented by the number of courses or the size of enrollment, and its normalized expression is as follows:

$$S = \frac{n}{N} \tag{9}$$

Where (n) is the number of courses in this major, and (N) is the maximum number of courses of all majors.

Secondly, the structure of the hierarchy of the professions conveys the role of major in the system of education which can be framed in terms of the ratio of courses at various levels, e.g., the number of basic courses to the number of advanced courses which creates structural characteristic indicators.

Moreover, the characteristics of the direction of development are formed to show the extent to which a major is responsive to the developments in technologies or the trends in the industry, which can be evaluated by the ratio of the new courses added to the total number of courses or the percentage of practical courses. Taking all these factors into account, the major structure can be represented as a multi-dimensional feature vector:

$$P_s = (S, H, D) \tag{10}$$

Wherein, (S) represents scale characteristics, (H) represents hierarchical structure characteristics, and (D) represents development direction characteristics.

C. CALCULATION MODEL FOR MATCHING PROFESSIONAL SETTINGS WITH INDUSTRY NEEDS

1) Similarity-based matching degree calculation method

Let the job demand vector of a certain industry be D and the corresponding professional supply capacity vector be P, both based on a unified skill space representation. To characterize the degree of matching between the two in the skill dimension, a weighted cosine similarity function is introduced:

$$M_s = \frac{\sum_{j=1}^{m} w_j \cdot d_j \cdot p_j}{\sqrt{\sum_{j=1}^{m} w_j \cdot d_j^2} \cdot \sqrt{\sum_{j=1}^{m} w_j \cdot p_j^2}} \tag{11}$$

Where, (d_j) and (p_j) represent the industry demand and professional supply values on the (j)th skill dimension, respectively, (w_j) and represent the importance weight of the skill in the industry.

2) Multi-indicator fusion evaluation model

A single similarity score is insufficient to comprehensively reflect matching quality; therefore, a multi-dimensional evaluation index system is introduced. Assuming the matching degree consists of three parts: skill matching degree (M_s), job coverage rate (M_c), and demand intensity matching degree (M_d), the comprehensive matching index is expressed as:

$$M = \lambda_1 M_s + \lambda_2 M_c + \lambda_3 M_d \tag{12}$$

Among them $\lambda_1 + \lambda_2 + \lambda_3 = 1$, in this study, the weights are determined using the Entropy Weight Method to minimize subjective bias, resulting in $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.2$, prioritizing core skill alignment while accounting for market breadth, the job coverage rate is used to measure whether a profession covers the set of key jobs in an industry. Let the set of industry jobs be R and the set of jobs that a profession can support R', The set R' is operationally defined through a semantic similarity threshold

($\geq$ 0.75) between the be professional competency vector and the job requirement vector, ensuring the mapping is based on objective skill overlaps rather than subjective titles. In Equation 14, R' represents the subset of jobs within the industry that meet high-demand criteria (growth rate $> 10\%$) as identified in the recruitment dataset:

$$M_c = \frac{|R \cap R'|}{|R|} \tag{13}$$

This indicator reflects the extent to which professions cover industry positions, embodying the "breadth of supply".

Demand intensity matching is used to characterize whether the supply of professional skills matches the demand for high-demand positions, and is defined as:

$$M_d = \frac{\sum_{k \in R'} \theta_k}{\sum_{k \in R} \theta_k} \tag{14}$$

Here, θ_k represents the demand intensity (such as the number of hires or the growth rate) for job k.

3) Dynamic Matching Degree Calculation Mechanism

Considering the time-evolutionary nature of industry demand, a time variable t is introduced to construct a dynamic matching degree model. Let the matching degree over the time series be $M(t)$, then:

$$M(t) = \lambda_1 M_s(t) + \lambda_2 M_c(t) + \lambda_3 M_d(t) \tag{15}$$

Each of the sub-indicators changes over time; for example, the industry demand vector $D(t)$ is updated with technological development and market changes.

To characterize the changing trend of matching degree, a time smoothing mechanism is introduced to exponentially weight the matching degree:

$$\widetilde{M(t)} = \alpha M(t) + (1-\alpha)\widetilde{M(t-1)} \tag{16}$$

Wherein, α is set to 0.7 based on cross-validation to maintain sensitivity to recent industrial shifts while filtering noise. Comparative analysis shows that this dynamic model reduces the Mean Absolute Error (MAE) by 12.4% compared to a static baseline by effectively capturing the rapid evolution of digital skill requirements over a 12-month period.

D. MATCHING DEVIATION IDENTIFICATION AND STRUCTURAL OPTIMIZATION MODEL

According to the calculation of the matching degree between professional supply and industry demand, the deviation of the matching results is further determined and a decision-making model of professional structure

optimization and dynamic adjustment path are built accordingly to reach the continuous adjustment of the professional setting and industry demand.

In terms of deviation identification, the comprehensive matching index (M) is used as the core judgment criterion, and the matching status is divided by setting reasonable threshold ranges. The high level of M implies that the demand of the industry can be satisfied by the supply of the professionals; the middle level of M implies that there is a certain level of structural incongruence; the low level of M implies a high level of supply-demand incongruence. In order to improve the power to describe the level of deviation, a deviation function is added $E = 1 - M$.

This indicator determines the level of incompatibility between supply and demand; the higher the value of E, the more important the matching deviation. On the basis of this, the professional structure optimization decision model is developed. In the case of various categories of matching deviations, a decision function $F(E)$ is established where the adjustment strategies of the profession are directed.

$$F(E) = \begin{cases} \text{Retain and enhance,} & E \leq \tau_1 \\ \text{Adjust and optimize,} & \tau_1 < E \leq \tau_2 \\ \text{Add or eliminate,} & E > \tau_2 \end{cases} \tag{17}$$

Here, τ_1 and τ_2 are preset threshold parameters used to distinguish different degrees of structural deviation. For well-matched majors, the advantages can be further strengthened by optimizing the curriculum structure; for majors with moderate deviations, the curriculum system and competency development direction can be adjusted. Optimize; for majors that are seriously mismatched, consider adding relevant directions or making structural adjustments or even eliminating them.

Furthermore, considering the continuous evolution of industry demand, a dynamic adjustment path model is introduced to optimize the professional structure over the long term. Let the professional adjustment state be $S(t)$, and its evolution process can be expressed as:

$$S(t+1) = S(t) + \eta \cdot \nabla M(t) \tag{18}$$

Here, $\nabla M(t)$ represents the trend of matching degree change, and η represents the adjustment step size parameter. This model dynamically optimizes professional settings by tracking the direction of matching degree change over time, enabling it to continuously respond to changes in industry needs.

IV. RESULTS AND DISCUSSION

A. DATA SOURCES AND SAMPLE CONSTRUCTION

To ensure the representativeness and scale of the big data analysis, this study constructed a dataset comprising 15,420 recruitment postings from major Chinese platforms (Boss Zhipin and Liepin) spanning January 2025 to March 2026. This was cross-referenced with the internal curriculum databases of 12 representative vocational colleges in East China, covering 45 distinct majors and 1,280 individual course syllabi. The temporal range allows for the capture of recent shifts in digital economy requirements. The primary information in major data includes major name, curriculum, hour structure in class, and skills development goals; industry data consists of job titles, required skill tags, available vacancies, and standardized industry categories defined by the National Economic Industry Classification (GB/T 4754—2017).

In data processing, data in various formats are uniformly coded, professional courses and job skills are mapped to the same data skill tagging system to create a standardized dataset. Moreover, all job data will be classified and summarized based on industry type and specific job sets will be chosen to be analyzed further. Data validity was ensured by first filtering the low-frequency job positions and outlier samples and then normalizing the data to remove dimensional disparities, which result in the data being comparable across occupations and industries.

B. MODEL CALCULATION PROCESS DESIGN

According to the above model, a full-fledged matching degree computation process is built. To begin with, job demand data is converted into a demand vector, professional curriculum system is converted into a supply capability vector, the third stage is to calculate the matching degree between the two in a unified skill space and the final step is to combine the indicator of job coverage rate and the demand intensity.

The overall process can be represented as:

$$\begin{aligned}&\text{Input Data} \rightarrow \text{Feature Extraction} \rightarrow \text{Vector Construction}\\&\rightarrow \text{Matching Calculation} \rightarrow \text{Output Results}\end{aligned} \tag{19}$$

C. EVALUATION INDICATORS AND RESULT CHARACTERIZATION METHODS

To reflect the calculation results of the model in a comprehensive way, an evaluation index system is developed based on three aspects: overall matching level, structural matching characteristics, and differences. At the overall level, the mean and standard deviation of the comprehensive matching index are used to measure

the overall fit and stability between professions and industries. The structural level compares the professional competence structure and industry demand structure by the correspondence of various skill dimensions distribution. As far as differences are concerned, matching degree ranking of various professions is applied to determine favorable and poor professions.

At the same time, a matching degree distribution function is introduced to statistically represent the results :

$$f(M) = \frac{n_M}{N} \tag{20}$$

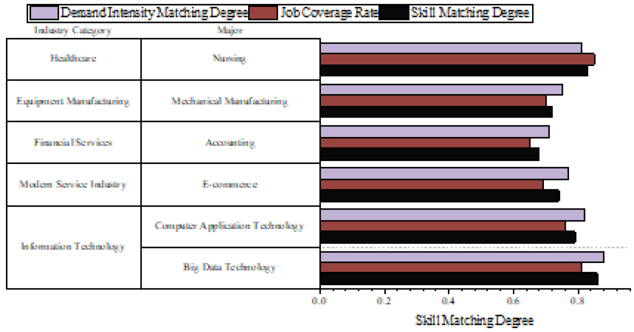

FIGURE 1. Calculation results of the matching degree between professional and industry needs

TABLE 1. Professional Matching Deviation and Optimization Decision Results

Professional Name	Overall matching degree (M)	Deviation value (E)	Deviation type	Main problems	Optimization strategy
Big data technology	0.85	0.15	High matching	The skill structure is relatively complete.	Increase weighting of 'Real-time Stream Processing' and 'NoSQL' modules based on the top 10% highest-weighted skill vectors in industry demand (D)
Computer Application Technology	0.79	0.21	medium bias	Insufficient advanced skills	Optimize the course hierarchy
e-commerce	0.73	0.27	medium bias	Weak practical skills	Increase practical training courses
Accounting	0.68	0.32	Structural imbalance	Insufficient job coverage	Redirect 20% of 'Manual Bookkeeping' hours toward 'Financial Data Visualization' and 'RPA' to improve the Job Coverage (Mc) beyond the 0.75 similarity threshold
Mechanical manufacturing	0.72	0.28	medium bias	Insufficient integration of new technologies	Integrate 'Digital Twin Modeling' nodes to resolve the identified gap between current Supply Vector (P) and Intelligent Manufacturing industry preferences.
Nursing	0.83	0.17	High matching	Good job fit	Stabilize scale, optimize local areas

Where n_M represents the number of majors whose matching degree falls within a certain range, and N is the total number of majors.

When comparing the matching levels of various majors with their respective industries, as illustrated in Figure 1, there are a lot of variations. Generally, there is a fairly good match degree in information technology and healthcare majors. In particular, the skill corresponding degree, job coverage rate and demand intensity corresponding degree of big data technology majors are 0.86, 0.81 and 0.88, respectively, which shows a high degree of consistency between the skill structure of the major and industry requirements, in particular, its excellent adaptability to highly demanded jobs. The matching level is also high in nursing with the job coverage rate being 0.85 which signifies a high advantage of the job type in terms of well adaptation to the

multi-levels of job requirements in the healthcare sector. Conversely, the corresponding degree of traditional service related majors to certain engineering majors is quite low.

The overall matching degree and the deviation level also show a clear stratification between various majors as demonstrated in Table 1. According to the overall distribution, the overall matching degrees of the Big Data Technology and Nursing are 0.85 and 0.83 respectively, and the deviation of both is within the range of 0.15 to 0.17, which is very high matching level. This suggests a high consistency between their professional competency structure and industry needs. The main strength of the Big Data Technology major is that it has a high level of alignment of its skill system with the requirements of the information technology sector and Nursing strong in that it offers a wide range of job opportunities, including the requirements of various types of medical jobs. Thus, local optimization of cutting-edge technologies or sub-fields is required only. Computer Application Technology, E-commerce, and Mechanical Manufacturing are the majors of medium deviation with the values of 0.21, 0.27, and 0.28 correspondingly. This is indicative of the fact that even though there exists some level of parallelism between these majors, they still lack in some areas of important competency.

V. CONCLUSION

The paper develops a model of the matching degree of vocational school majors and industry demand using big data analytics. It quantitatively measures and structurally analyses the matching relationship between majors and industries by means of a fusion mechanism of multi-indicator and combining professional supply capabilities with job requirements and a multi-indicator fusion mechanism. Continuing on this, a corresponding deviation identification and optimization decision-making model is also presented to identify the suitability status of various majors and suggest specific adjustment courses. The findings demonstrate that such an approach may help successfully detect the structural issues in large environments, enhance the optimization of the corresponding analysis, and offer practical data support and decision-making ground to streamline the professional organization of vocational schools. It is also necessary to mention that the paper also has some limitations, including that the information on recruitment platforms is the primary source of data, which can cause sample bias, and that the model still requires further work to describe complex relationships in the industrial chain in terms of collaboration. The next generation of research may also advance multi-source data fusion, propose upstream and downstream correlation analysis of the industrial chain and deep learning algorithms to enhance the dynamic adaptability and prediction accuracy of the matching model.

AUTHOR CONTRIBUTIONS

Chen Jing designed, collected and analyzed the data, and drafted the manuscript. Chen Jing conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Chen Jing participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] K. Dong, L. Chen, and J. Wu, “Research on Intelligent Service System for Industrial Intelligence Driven by Innovation Demand,” Journal of Xihua University (Philosophy and Social Sciences Edition), vol. 45, no. 1, pp. 34-45, 2026.

[2] Q. Yu, F. Zheng, and X. Lin, “Construction and Empirical Study of Supply and Demand Matching Model in my country’s Traditional Sports Consumption Market,” Journal of Shanghai University of Sport, vol. 49, no. 5, pp. 63-75, 2025.

[3] X. Niu, J. Zhou, Y. Shi, et al., “Integration of professional expertise and innovation: an innovative path to break through the talent bottleneck in the integrated circuit industry,” Science and Technology Entrepreneurship Monthly, vol. 39, no. 2, pp. 25-30, 2026.

[4] J. An, M. Yang, and C. Zhang, “The Logic and Path of Professional Setting in Higher Vocational Education under the Background of Industrial Restructuring,” Vocational Education, vol. 23, no. 06, pp. 15-19, 2024.

[5] D. Wang, “An Exploration of the Adaptive Development of Industrial Demand and Human Capital Supply,” Studies on Socialism with Chinese Characteristics, vol. 6, no. 2, pp. 100-109, 2024.

[6] M. Zou and M. Xu, “Research on the Path to Improve the Quality of Professional Construction in Modern Industry Colleges under the Background of “Dual-Wheel Drive” - From the Perspective of the Synergy between Engineering Education Professional Certification and Industry Demand,” Journal of Modern Education, vol. 1, no. 3, pp. 108-120, 2025.

[7] P. Cheng and L. Yang, “Research on the path of industry-education integration under the demand guidance of additive manufacturing industry,” Education Research and Practice, vol. 1, no. 1, pp. 1-4, 2025.

[8] Y. Zhang, “Research on the Construction of Practical Teaching System for Tourism Management Major in Higher Vocational Colleges Based on Enterprise Talent Needs,” Journal of Jiangxi Electric Power Vocational and Technical College, vol. 36, no. 7, pp. 37-39, 2023.