ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

An Integrated Intelligent Teaching Evaluation System Empowered by Multimodal Large Models

Wanli Wen, Jiayu Li

School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Abstract

To address the structural contradiction between large-scale education and personalized assessment confronting the current "Four Evaluations" reform, and to overcome the drawbacks of traditional evaluation methods such as the separation of process and outcome, high subjectivity, and a lack of timeliness, this study proposes and designs an integrated intelligent teaching evaluation system empowered by Multimodal Large Models. The core of this system is designed around three main steps. First, it integrates multimodal data streams, including audio, vision, and text, to establish a unified data foundation for the in-depth perception and quantitative representation of all classroom elements. Second, grounded in core educational theories such as formative assessment, cognitive diagnosis, and multiple intelligences, a multidimensional evaluation indicator framework is proposed. Finally, by leveraging multi-task learning and joint inference algorithms, an integrated intelligent evaluation engine is conceptualized. This engine is capable of simultaneously producing outcome, process, value-added, and comprehensive evaluations within a unified computational framework. Theoretical analysis and architectural design suggest that this system will shift the evaluation focus from being solely score-oriented to being competency-oriented, and from summative assessment to formative empowerment. It aims to establish a real-time, closed-loop feedback mechanism that effectively integrates evaluation with teaching, thereby providing a practical technological pathway for implementing large-scale individualized instruction and fostering educational equity.

Keywords

Multimodal Large Models; Integrated Evaluation; Educational Artificial Intelligence; Evaluation-Teaching Closed Loop; Cognitive Diagnosis.

1. Introduction

The issuance of the Overall Plan for Deepening the Reform of Educational Evaluation in the New Era (hereinafter referred to as the Plan) in China marks the entry of the national educational evaluation reform into a systematic and in-depth phase [1]. The Plan places explicit emphasis on improving evaluation in higher education, requiring enhancements to undergraduate education assessment, reinforcing the central position of talent cultivation, and highlighting key metrics such as ideological and political education and graduate development. However, the practical implementation of these reforms, particularly in large-scale science and engineering programs, faces significant structural challenges.

These challenges manifest primarily as three long-standing structural contradictions within traditional evaluation paradigms:

Data Fragmentation and Task Overload: Process-based behavioral data collected during instruction is often disconnected from summative data (e.g., exams and assignments). This

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

separation forces teachers to spend extensive effort on manual integration for comprehensive assessment, leading to high administrative task loads.

Reliability vs. Validity Dilemma [2]: Traditional classroom observation methods rely heavily on the subjective judgment of the instructor, which compromises evaluation reliability. Conversely, standardized testing often fails to effectively measure the high-order cognitive abilities crucial for engineers, such as critical thinking, communication, and collaboration, thereby limiting evaluation validity.

Disconnection between Feedback and Teaching Improvement [3]: Evaluation results are frequently delayed, offering "summative" rather than "formative" feedback. This time lag impedes timely and precise diagnosis and improvement suggestions, resulting in a distinct separation between the "evaluation" and "teaching" processes [4].

To systematically address these structural dilemmas, Multimodal Large Model (MMLM) technology offers a powerful new technical possibility [5]. MMLMs excel in cross-modal data understanding, content generation, and sophisticated human-computer interaction.

Therefore, the core research problem addressed by this study is: How can MMLM technology be leveraged to construct an intelligent evaluation system that deeply integrates data, indicators, and algorithms?

The proposed solution involves developing an integrated computational framework that breaks down data silos, organically unifies the "Four Evaluations" (Outcome, Process, Value-Added, and Comprehensive), and establishes a data-driven closed-loop mechanism to guide teaching improvement. This paper's primary contribution is to propose this technical framework and implementation path, offering an operational solution to the inherent conflict between delivering large-scale education and providing personalized assessment.

2. Overall Architecture Design

The technical premise for achieving the organic integration of the "Four Evaluations" is the construction of a unified technical architecture capable of breaking through multimodal data barriers and integrating cross-domain information. The system architecture designed in this study adheres to the principle of layered decoupling and is ensured to be scientifically sound and effective through a deep fusion of theory and design. The core design philosophy of this system is to realize the unification of the evaluation process and outcome, evaluation and teaching, and multi-dimensional abilities and comprehensive literacy.

To realize this philosophy, the system architecture is deeply rooted in a "Scholarly Core" based on educational and algorithmic theories. Specifically, Formative Assessment Theory [6] emphasizes that evaluation's core function is to promote learning by continuously adjusting teaching strategies through feedback throughout the instruction, directly guiding the design of the "Evaluation-Teaching Closed Loop." Building upon this, Cognitive Diagnostic Theory [7,8] provides robust mathematical modeling support for the precise and dynamic assessment of student ability levels. Furthermore, the Multiple Intelligences Theory [9] offers the theoretical framework necessary for constructing a comprehensive, multi-dimensional evaluation indicator system that transcends singular academic scores, thereby necessitating the capture of students' comprehensive performance in areas like collaboration, expression, and innovation. Finally, the Multi-Task Learning Theory [10, 11] serves as the key algorithmic support, enabling the synchronous processing of all four evaluation types (outcome, process, value-added, and comprehensive) by sharing model parameters across related tasks.

Under the guidance of these theories, the system architecture (Fig. 1) is constructed bottom-up, starting with the Perception and Collection Layer, which digitizes the entire classroom environment using devices like HD cameras and microphone arrays. The data then flows to the Data Processing and Feature Layer, the core for data fusion, where raw input is converted into

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

a spatiotemporally synchronized, unified feature space by utilizing techniques such as Automatic Speech Recognition (ASR), Computer Vision (CV), and Natural Language Processing (NLP), thereby providing structured evidence. This evidence then feeds into the Model Inference and Evaluation Engine Layer, which acts as the system's "brain," building an integrated evaluation model based on the multi-task learning framework to simultaneously output the results for the four evaluation types within a unified deep neural network. Finally, the Application Services and Feedback Layer, the Application Services and Feedback Layer, translates these quantitative results into visualized services, such as a teaching cockpit and student digital portraits, thereby constructing an integrated closed loop from data collection and intelligent analysis to teaching intervention and effect tracking. The core design principle guiding this entire structure is the systemic conversion of discrete, heterogeneous classroom behavior information into a unified, computable digital representation space, upon which the integrated evaluation model is established.

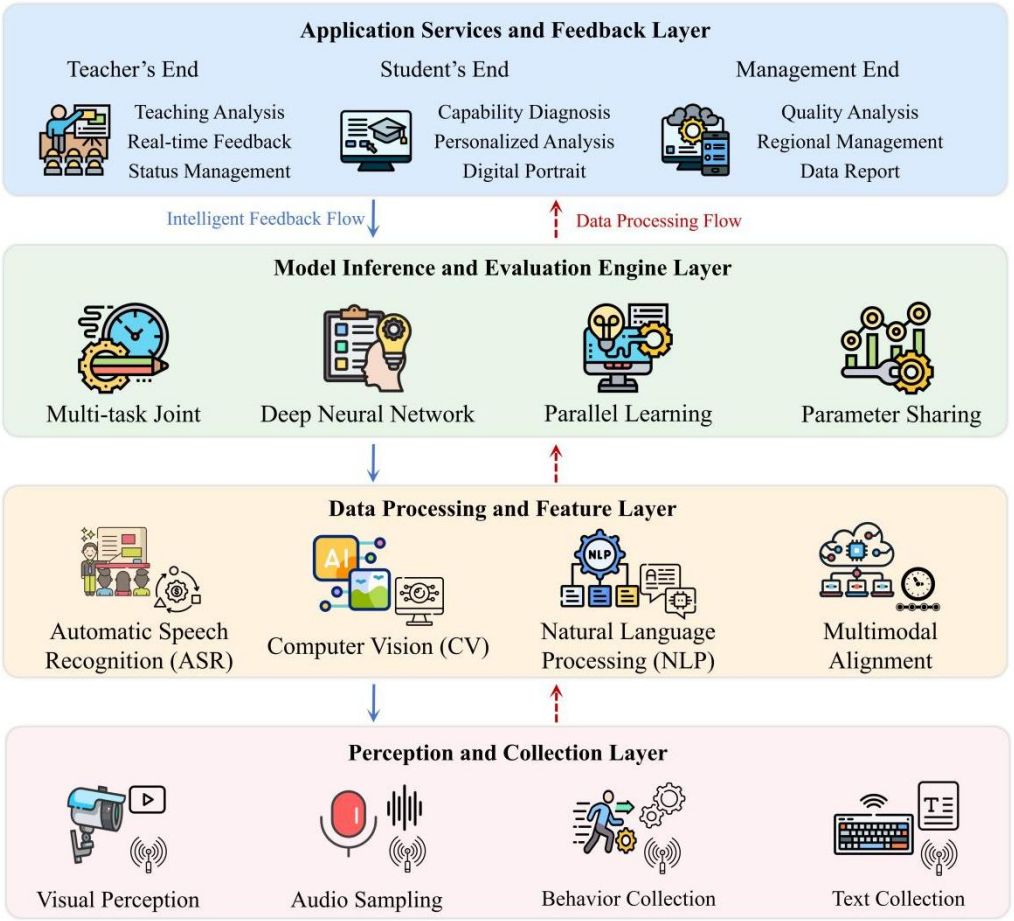

Figure 1. Layered Technical Architecture of Integrated Evaluation System

3. Implementation Path

Based on the unified data architecture, this research proposes a "four-step" integrated implementation path, progressing from data to evidence, then to scientific evaluation, and finally to a continuous improvement closed loop.

3.1. Step 1: Constructing the Unified Evidence Space

This step focuses on abstracting low-level multimodal features into intermediate-layer "Evidence" that is interpretable, time-stamped, context-aware, and subject-labeled (e.g., "Student A's frown frequency was 80% higher than average during the explanation of knowledge point K"). This evidence is organized and stored in a "Learner-Task-Context" graph

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

structure, allowing the system to query and correlate student performance freely across time and task boundaries, which is crucial for integrated evaluation.

3.2. Step 2: Designing the Scientific "Shared-Exclusive" Indicator System

The four evaluation dimensions share many underlying metrics. A hierarchical indicator library is therefore constructed, comprising core indicators shared by multiple dimensions (e.g., peer collaboration quality) and exclusive key indicators specific to each dimension (e.g., "knowledge application transfer ability" for outcome evaluation).

To ensure the scientific rigor of this system, a comprehensive validation path is designed, including:

Content Validity: Ensured through expert interviews.

Construct Validity: Tested using Confirmatory Factor Analysis (CFA) [12].

Reliability: Assessed using Inter-rater Consistency (ICC) as a calibration metric, targeting a score of ≥0.80.

Fairness: Guaranteed using techniques such as Differential Item Functioning (DIF) analysis.

3.3. Step 3: Running the Multi-task Joint Inference Model

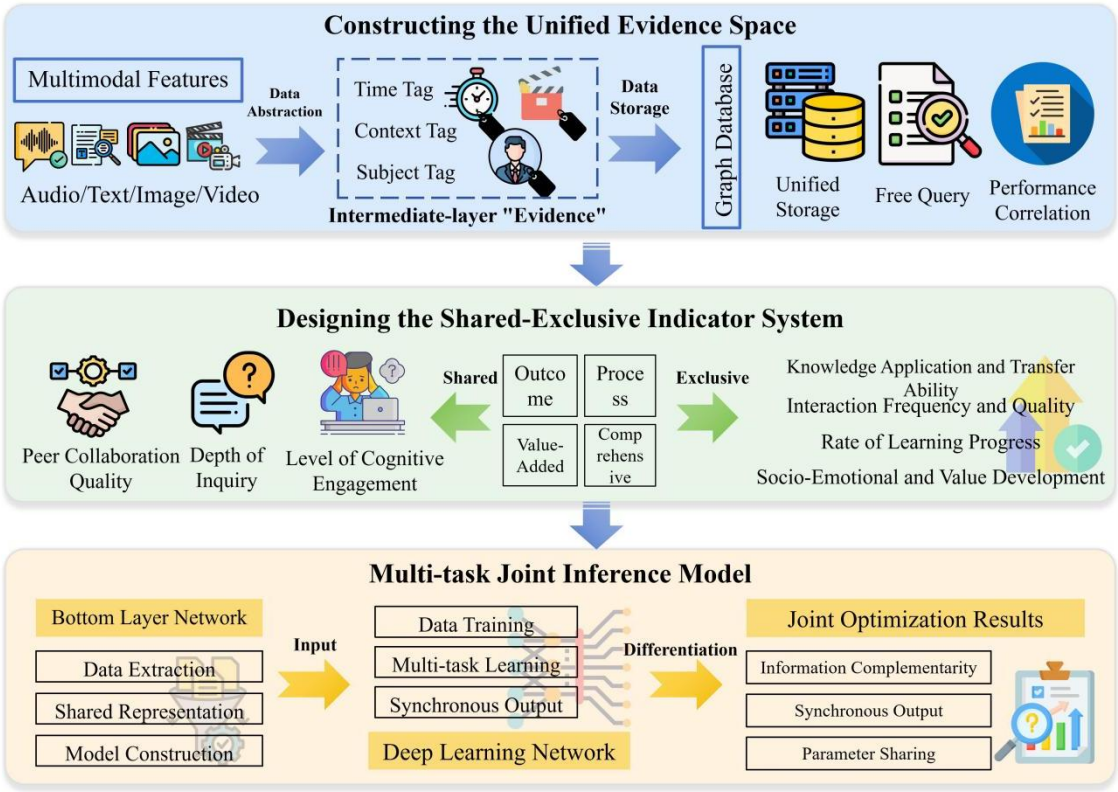

Figure 2. Integrated Implementation Logic Model for the "Four Evaluations"

This represents the technical core for integrated evaluation. As illustrated in Fig. 2, the unified evidence space serves as the input to an end-to-end deep learning network, which simultaneously outputs the results for the four evaluation types. The bottom network layer learns the shared representation of all evidence, while the top layer differentiates into four "Task Heads" corresponding to the calculation of each evaluation type. The loss functions of the four tasks are weighted and combined for joint optimization, ensuring that the different evaluation tasks can mutually reinforce each other, thereby enhancing overall accuracy and robustness.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

3.4. Step 4: Evaluation Closed Loop

The final step, which actualizes the ultimate value of evaluation, is the formation of the evaluation-driven continuous improvement closed loop [14]. This mechanism relies on two core, actionable functionalities (Fig. 3). First, the system implements Threshold-triggered Human-Computer Collaborative Intervention where, upon precisely monitoring widespread learning difficulties (such as a majority of students exhibiting confusion), it immediately pushes low-interference, instant teaching suggestions to the instructor. Second, it utilizes Data-driven Personalized Learning Resource Recommendation, automatically recommending digital educational resources that match each student's knowledge weak points and individual learning styles after class, based on their fine-grained, multi-dimensional ability portrait. By integrating these in-class instant interventions with personalized support outside of class, the system seamlessly translates evaluation data into actionable teaching steps, thereby constituting a complete and continuously improving personalized learning closed loop.

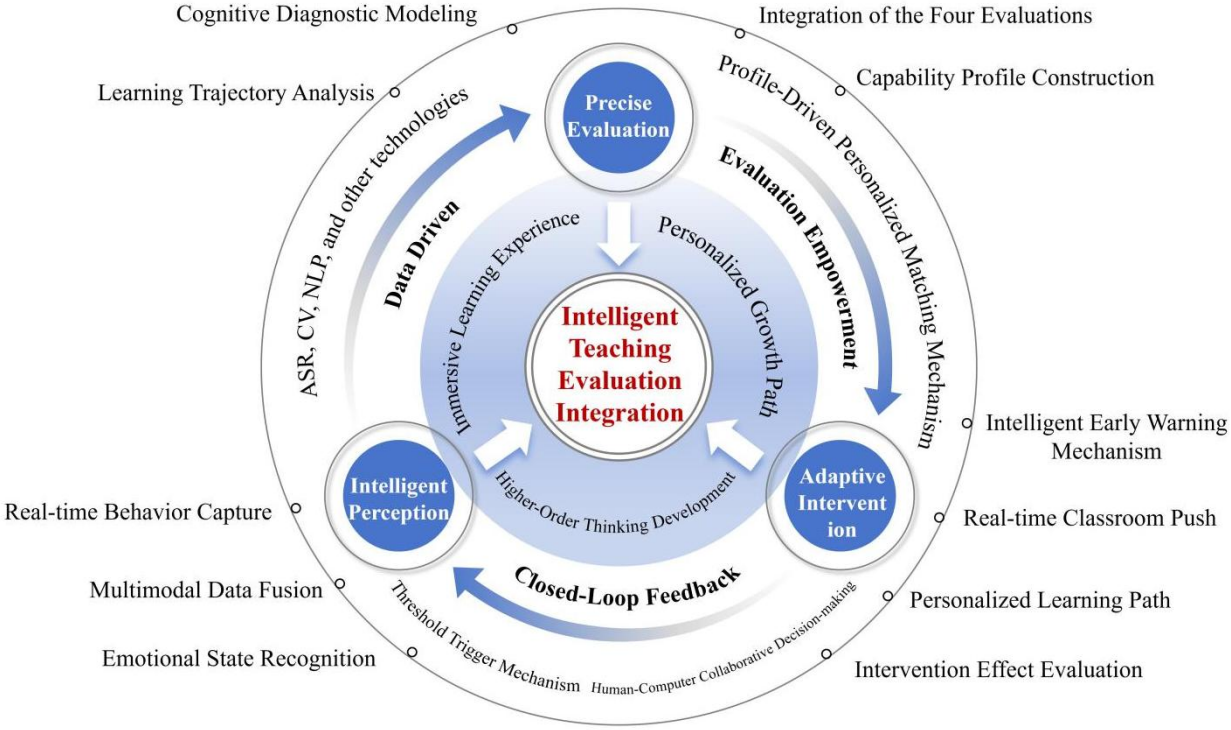

Figure 3. Schematic Diagram of the "Evaluation-Teaching Integration" Feedback Closed-Loop Mechanism

4. Deployment and Ethical Implications

The deep integration of Artificial Intelligence into educational evaluation necessitates a prudent approach to the ethical challenges concerning educational equity and data security [5, 15]. A responsible system must promote, rather than undermine, fairness. We propose the following three recommendations for a comprehensive implementation framework.

4.1. Flexible Deployment Architecture for Equity and Efficiency

To prevent smart technology from exacerbating the "digital divide," we propose a "Cloud-Edge-End" collaborative deployment model (Fig. 4). Locally deployed edge computing nodes can handle most real-time data processing in areas with weak information infrastructure, significantly reducing bandwidth requirements. Furthermore, offering free basic core functionalities and optimizing algorithms for mid-to-low-end hardware ensures wider accessibility and promotes educational equity.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

4.2. Full Life-cycle Data Security and Privacy Protection

Strict adherence to legal frameworks is necessary, requiring de-identification during data collection, encryption during transmission, and stringent access control during storage. Crucially, the system is strongly recommended to support the Federated Learning framework. This "data stays put, model moves" paradigm allows raw data to remain on local servers, uploading only encrypted model update parameters, effectively eliminating privacy risks associated with centralized sensitive data storage [13-15].

4.3. Algorithmic Ethics and Supervision Mechanism

To ensure fairness and transparency, all core evaluation algorithms must undergo rigorous ethical assessment prior to deployment [16]. The system should implement an Algorithm Transparency Mechanism, providing explainable reports for key evaluation results, granting users the right to know and the right to appeal. We recommend forming an Ethics Committee—comprising education, technology, and legal experts, along with teacher and student representatives—to regularly review the algorithm's fairness and handle user appeals, ensuring the system operates ethically [17, 18].

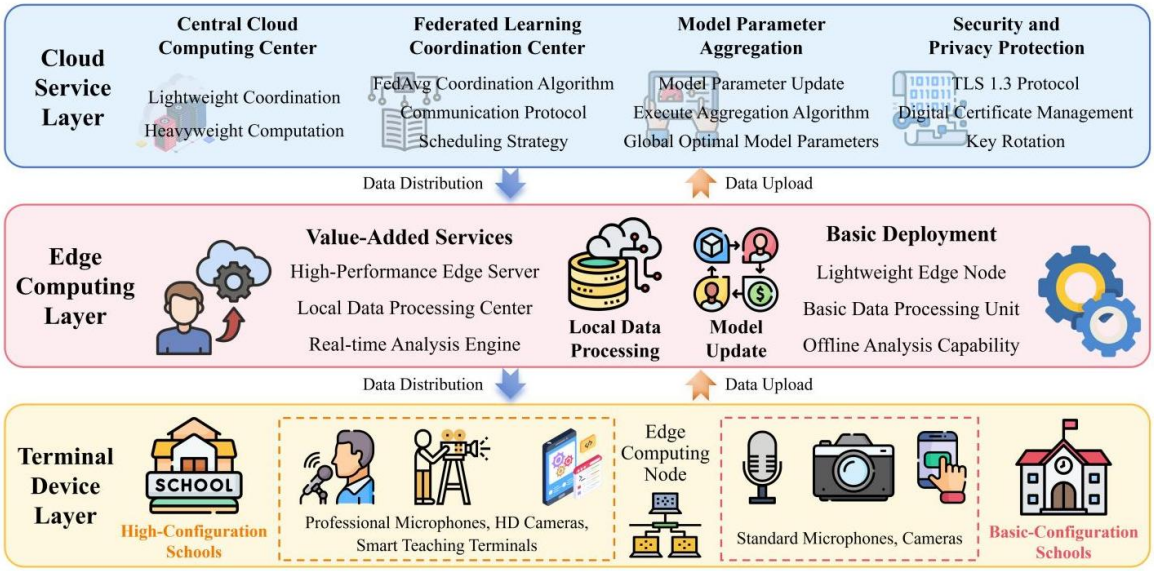

Figure 4. Flexible Deployment Architecture for Educational Equity and Privacy Protection

Acknowledgements

This research was supported by the Research Planning Project on Higher Education Science of the China Higher Education Society ("Research on the Coupling System of Large-Scale Precision Education and Intelligent Teaching Evaluation under the View of Multimodal Large Models," No. 24PG0103) and the Research Project on Higher Education Science of the Chongqing Higher Education Society ("Research on a Multidimensional Teaching Quality Evaluation System from the Perspective of Artificial Intelligence," No. cqgj23002B).

References

[1] The CPC Central Committee and the State Council. General Plan for Deepening the Reform of Education Evaluation in the New Era. Information on: https://www.gov.cn/zhengce/2020-10/13/content_5551032.htm.

[2] Yubo Hou, Qiangqiang Li, Hao Li. The Construction of Critical Thinking Structure and Scale Development in China. Journal of Peking University (Natural Science Edition). 2022, Vol. 58 (No. 02), p. 383-390.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0021

[3] Ismail S M, Rahul D R, Patra I, et al. Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. Language Testing in Asia. 2022, Vol. 12 (No. 01), p. 40.

[4] Xingnan Lu, Xuewei Gao. Artificial Intelligence Empowering Educational Evaluation Reform: Development Trends, Risk Assessment and Mitigation Strategies. China Journal of Education. 2023, Vol. (No. 02), p. 48-54.

[5] Zhinan Huang, Gen Li, Yafeng Zheng. Empowering the High-Quality Development of Science Education with Multimodal Large Models: Potential, Challenges, and Application Exploration. China Electro-education. 2025, Vol. (No. 06), p. 60-69.

[6] Jeon H, Jun Y, Laine T H, et al. Immersive virtual reality game for cognitive-empathy education: Implementation and formative evaluation. Education and Information Technologies. 2024, Vol. 29 (No. 02), p. 1559-1590.

[7] Han Y, Ji F, Jiang Z. Two-stage polytomous attribute estimation for cognitive diagnostic models: overcoming computational challenges in large-scale assessments with many polytomous attributes. Humanities and Social Sciences Communications. 2025, Vol. 12 (No. 01), p. 1-14.

[8] Yue Wang, Shujuan Chang, Xiaoling Han, et al. Item bank construction and validity testing based on item response theory: A case study of the public course "Modern Educational Technology". Modern Educational Technology. 2019, Vol. 29 (No. 10), p. 41-47.

[9] Gardner H. Frames of Mind: the theory of multiple intelligences. Basic Books, 1983.

[10] Vandenhende S, Georgoulis S, Van Gansbeke W, et al. Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, Vol. 44 (No. 07), p. 3614-3633.

[11] Zhang Y, Yang Q. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering. 2021, Vol. 34 (No. 12), p. 5586-5609.

[12] Rogers P M, Marine J M, Ives S T, et al. Validity evidence for a formative writing engagement assessment in elementary grades. Assessment in Education: Principles, Policy & Practice. 2022, Vol. 29 (No. 02), p. 262-284.

[13] Xie Q, Jiang S, Jiang L, et al. Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey. IEEE Internet of Things Journal. 2024, Vol. 11 (No. 14), p. 24569-24580.

[14] Tingting Feng, Dejian Liu, Lulu Huang, et al. Digital Education: Application, Sharing, Innovation-Summary of the 2024 World Digital Education Conference. China Electro-education. 2024, Vol. (No. 03), p. 20-36.

[15] Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions. IEEE Signal ProcesYunong Yang, Ao Xu, Chunjiong Zhang, et al. Privacy Protection in Smart Classrooms Based on Federated Multi-task Learning. Modern Educational Technology. 2024, Vol. 34 (No. 09), p. 123-132.sing Magazine. 2020, Vol. 37 (No. 03), p. 50-60.

[16] Yunong Yang, Ao Xu, Chunjiong Zhang, et al. Privacy Protection in Smart Classrooms Based on Federated Multi-task Learning. Modern Educational Technology. 2024, Vol. 34 (No. 09), p. 123-132.

[17] Leilei Zhao, Li Zhang, Jing Wang. Ethical risks of educational data in the intelligent era: typical representations and governance paths. China Distance Education. 2022, Vol. (No. 03), p. 17-25+77.

[18] Huibin Zhang, Lei Xu. Ethical Risks and Governance Approaches of Generative AI in Education: A Case Study of Russell Group. Modern Educational Technology. 2024, Vol. 34 (No. 06), p. 25-34.