A Lightweight AI Agent for Classroom Teaching Diagnosis in Higher Vocational Education

Zheng Jiangwei, Qi Xinying, Wu Cuijiao
School of Artificial Intelligence, Caofeidian Vocational and Technical College, Tangshan, Hebei, China

Abstract: Against the background of the digital transformation of vocational education, routine classroom evaluation in higher vocational colleges remains constrained by subjective judgement, fragmented criteria and weak alignment between lesson design and occupational competency requirements. Existing commercial AI systems are further limited by generic evaluation logics, dependence on video-capture infrastructure and insufficient support for integrated assessment across posts, courses, competitions and certificates. To address these limitations, this study develops XiaoQ, a lightweight school-based intelligent agent for classroom teaching diagnosis built on prompt engineering, modular role constraints and a domain-specific knowledge base. The agent supports two modes: standalone lesson-plan diagnosis and integrated diagnosis combining instructional documents with classroom observation records. We further propose a dual-track framework that separates static evaluation of instructional design from dynamic quantification of classroom behaviour, thereby distinguishing objective behavioural statistics from vocationally oriented interpretation. Application testing indicates that the agent can be deployed without dedicated hardware, has a low threshold for routine use, and produces diagnostic feedback aligned with school-based talent-training objectives. The proposed approach offers a replicable and low-cost model for data-informed teaching research and standardised classroom evaluation in higher vocational education.

Keywords: Artificial Intelligence; Classroom Teaching Diagnosis; Higher Vocational Education; Prompt Engineering; School-Based Knowledge Base

1. Introduction

Higher vocational education is shifting from expansion in scale to improvement in educational quality, which has intensified the need for the digital and standardised reconstruction of the full teaching process. Within an outcome-based education (OBE) framework, talent cultivation follows a hierarchical logic: professional-cluster objectives inform programme competency standards, programme standards shape course-level goals, and course goals are further translated into operational lesson objectives [1]. At the same time, rapid advances in large language models have created new opportunities for lightweight teaching-research tools that can support lesson observation, lesson-plan revision and post-class reflection [2]. Bibliometric research likewise suggests that generative AI is becoming an important driver of classroom-evaluation reform in vocational colleges [3].

Despite these developments, routine classroom evaluation in higher vocational colleges still faces several practical constraints. Conventional peer review continues to rely heavily on handwritten observation notes, subjective commentary and simple student feedback, with limited use of standardised quantitative indicators. As a result, such evaluations rarely examine whether teaching activities align with occupational requirements, skills-competition standards or vocational certification criteria, and therefore provide only limited guidance for improving practice-oriented teaching.

Although some institutions have introduced intelligent teaching-analysis platforms, three obstacles continue to limit their routine adoption. First, video-based commercial systems usually require dedicated smart classrooms, sustained cloud-service subscriptions and continuing technical maintenance, which together impose substantial financial pressure on regional vocational colleges [4]. Second, generic AI tools seldom embody the logic of vocational education. Most focus on superficial behaviour counts rather than on whether teaching design supports the integration of posts, courses, competitions and certificates, or on whether it cultivates the professionalism and craftsmanship expected in technical programmes [5]. Third, visual-recognition approaches remain vulnerable in practical training settings characterised by equipment occlusion, group operations and complex interactions, which can distort data and reduce the credibility of diagnostic results [6]. Some scholars therefore argue that objective behaviour statistics should be analytically separated from profession-specific interpretation [7].

To address these challenges, this study develops a zero-hardware intelligent agent for classroom teaching diagnosis in response to the routine teaching-research needs of the School of Artificial Intelligence at a local higher vocational college. The agent is implemented on Coze, a low-code large language model development platform that supports customised knowledge bases and logical constraints [8]. It provides two diagnostic modes, one for pre-class lesson-plan review and the other for full-session diagnosis based on both teaching documents and classroom observation data. In addition, the study proposes a dual-track evaluation framework that separates static instructional design assessment from dynamic classroom-behaviour quantification, thereby avoiding the methodological confusion that arises when subjective judgement and objective measurement are merged into a single score. The remainder of the paper presents the system design, evaluation framework, implementation route and application results.

2. System Design, Dual-Track Evaluation Framework and Technical Implementation

The proposed intelligent classroom diagnosis agent is built on Coze, a low-code large language model platform that supports skill configuration, private knowledge-base management and third-party tool integration. Supported by school-specific teaching resources, teacher-provided instructional materials and a dedicated AI diagnostic engine, the system forms a closed-loop digital teaching-research framework. This section introduces the overall architecture, the dual-track evaluation framework and the lightweight implementation route.

2.1 Overall Architecture and Collaborative Operation

The system comprises four interrelated functional layers. The role and risk-control layer defines the agent as a specialist in vocational education evaluation and imposes strict grounding constraints to prevent fabricated evidence and generic comments. All diagnostic outputs are therefore derived exclusively from uploaded lesson plans and classroom observation records, ensuring traceability and authenticity. The standardised evaluation layer constitutes the analytical core of the system and integrates eight static instructional-design dimensions with a set of dynamic classroom-behaviour indicators. The visual reporting layer automatically generates radar charts and proportional graphs to present diagnostic results in an intuitive manner and reduce teachers' interpretive burden. Finally, the school-based knowledge-base layer incorporates the talent cultivation plans of the School of Artificial Intelligence and extracts structured benchmarks, including professional-group training objectives, typical job tasks, 1+X certificate standards and skills-competition requirements.

Within this architecture, two diagnostic modes are provided to address different teaching-research needs. Standalone lesson-plan diagnosis supports pre-class refinement by examining the internal consistency of four-level training objectives, from professional-cluster goals to individual lesson objectives, while activating only the static evaluation dimensions. Comprehensive classroom diagnosis combines lesson plans with structured classroom observation records to compare intended instructional design with actual classroom implementation, and therefore activates both static and dynamic indicators.

The collaborative workflow linking the college, teachers and the intelligent agent forms a closed-loop process of input, diagnosis and output. Teachers provide instructional materials, the college supplies institutional standards through the school-based knowledge base, and the agent performs evidence verification, indicator quantification and vocationally oriented interpretation. The system ultimately outputs core evaluation indicators and an assessment of the alignment between teaching content and occupational competency requirements.

2.2 Dual-Track Evaluation Framework

To overcome the methodological weaknesses of commercial AI evaluation tools, this study separates static grading standards from dynamic classroom quantification indicators and constructs a two-layer analytical logic: objective behavioural statistics and vocationally oriented interpretation. This dual-track design reduces the distortion that may arise when subjective judgement and objective measurement are collapsed into a single weighted score, while preserving the explanatory value required in vocational education.

2.2.1 Static Evaluation Dimensions for Instructional Design

Guided by the outcome-based education (OBE) philosophy, eight standardised static evaluation dimensions were established to assess lesson-plan quality. The corresponding judgement criteria are shown in Table 1.

Table 1. Static Instructional-Design Evaluation Dimensions and Judgement Criteria

Evaluation dimension	Core assessment standard
Teaching objective design	Examines the completeness of objectives, their alignment with learner characteristics, and the consistency of the four-level objective system, while covering occupational tasks, 1+X certificates and competition-related competencies.
Key and difficult points design	Focuses on high-frequency occupational skills, common certificate-related errors and competition-critical difficulties, rather than on isolated theoretical content.
Question design quality	Uses the 4MAT framework to design questions at multiple cognitive levels, with priority given to vocationally relevant higher-order questioning.
Teaching activity design	Adopts project-based tasks, reconstructs enterprise workflows, integrates competition and certificate preparation, and cultivates craftsmanship.
Teaching process logic	Follows a complete enterprise workflow, including information acquisition, planning, implementation, inspection and evaluation.
Teaching resource adaptation	Prioritises enterprise cases, occupational operation manuals and loose-leaf industry teaching materials.
Professional literacy integration	Assesses AI-related professional literacy, including data security, coding norms, engineering ethics and a rigorous practical spirit.
Overall implementability	Evaluates whether the design matches campus training facilities, students' prior preparation and occupational training needs.

2.2.2 Dynamic Quantification of Classroom Behaviour

For dynamic in-class measurement, this study draws on the coding framework of the Flanders Interaction Analysis System and adapts it to higher vocational post- and competition-oriented training scenarios. The resulting classroom observation scale was refined through three rounds of iterative revision. In the formal reliability test, two trained teachers simultaneously recorded the same classroom sessions, and the intraclass correlation coefficient (ICC) was used to evaluate inter-rater consistency. The resulting ICC of greater than 0.85 met the usual reliability threshold for educational measurement.

Because classroom observation records are neutral, objective and non-interventive behavioural data, the quantification module performs raw frequency statistics only and does not assign additional vocational weights. The indicator system covers five dimensions. Effective teacher questioning is classified into routine management, memory, reasoning, creative and critical questions, together with respondent-selection modes, response forms and response quality. Teacher feedback is divided into verbal and non-verbal forms and coded across five attitudinal levels: positive feedback, negative feedback, zero feedback, interruption or replacement of student answers, and restatement plus interpretation of student replies. The 4MAT framework is used to calculate the proportion of four cognitive question types (What, Why, How and If) [9]. Classroom dialogue is coded into five levels of depth, from shallow exchanges to advanced reasoning. Finally, the quality of group-based cooperative learning is scored on a five-point scale covering participation, task allocation, peer interaction, role performance and conflict management.

2.3 Technical Implementation of the Lightweight Agent

The lightweight diagnostic assistant was implemented entirely on the Coze low-code platform and rests on three technical pillars: hierarchical prompt engineering, a school-specific professional knowledge base and a direct-output interaction mode requiring no dedicated hardware.

First, a high-stability prompt system was designed around seven functional modules: role definition, scenario boundary control, unified rating rules, visualisation standards, vocational interpretation logic, risk-control constraints and a standardised output format. Three principles were treated as highest-priority rules throughout the diagnostic process: data objectivity, occupation-oriented interpretation and cross-verification of hierarchical objectives.

Second, the school-based knowledge base was bounded by the School of Artificial Intelligence at Caofeidian Vocational and Technical College. It integrates talent cultivation plans across all relevant programmes and extracts structured evaluation benchmarks, including professional-group objectives, occupational orientations, core competency standards, 1+X certificate requirements, skills-competition priorities and norms of professional conduct.

Third, the tool adopts a lightweight deployment model that requires no specialised hardware, no independent server construction and no paid cloud authorisation. Teachers need only upload instructional materials to obtain automatic scoring, chart generation, quantitative analysis and vocationally oriented interpretation within seconds, which makes the system suitable for routine teaching-research activities in resource-constrained higher vocational colleges.

3. Application Case Analysis

3.1 Diagnostic Procedure

To examine the practical applicability of the proposed tool, this study used "HTML5 New Common Tags and General Attributes" from the course "Front-End Application Technology" in the Mobile Application Development programme as a representative lesson. The complete lesson plan and a structured classroom observation record were uploaded simultaneously to the XiaoQ intelligent assistant to initiate comprehensive classroom diagnosis.

The diagnostic process followed a two-stage standard of "dual-track isolation and post-coupling". In the first stage, the static module evaluated the lesson across the eight instructional-design dimensions and assigned an overall grade of B. In parallel, the dynamic observation module recorded only the original frequencies and proportions of teacher-student interaction behaviours, without adding interpretive weight to the raw data.

3.2 Main Problems Identified

The dual-track analysis identified two major weaknesses in this lesson. At the instructional-design level, classroom questioning was dominated by low-level "What" questions centred on memorisation and comprehension, whereas higher-order reasoning and creative questions rooted in authentic vocational scenarios were insufficient. Some practical-training activities were also insufficiently aligned with the assessment requirements of the Web Front-End Development 1+X certificate. In addition, the ideological and political education component remained overly general and did not translate into discipline-specific professional literacy for artificial intelligence students.

At the classroom-implementation level, Figure 1 shows that teachers mainly used managerial and memory-oriented questions, which coincided with low student willingness to participate voluntarily. Most teacher-student dialogue remained at Levels 1 and 2, with only a small proportion of higher-order reasoning and creative exchange. Group training also lacked explicit role allocation, leading to uneven participation among students.

3.3 Targeted Improvement Strategies

The intelligent agent generated targeted strategies for each of the identified deficiencies. To improve instructional design, scenario-based questions derived from authentic front-end development tasks were added, and the relationship between practical-training activities and the assessment criteria of the Web Front-End Development 1+X certificate was restructured into a clearer matching list. Generic ideological and political requirements were further refined into major-specific literacy objectives, including standardised coding awareness. To optimise classroom implementation, laddered questioning chains were designed to guide students from factual recall to higher-order reasoning and innovation. Enterprise code cases were introduced to enrich advanced classroom questioning, and a standardised group-work mechanism was proposed to clarify roles such as team leader, coder and tester.

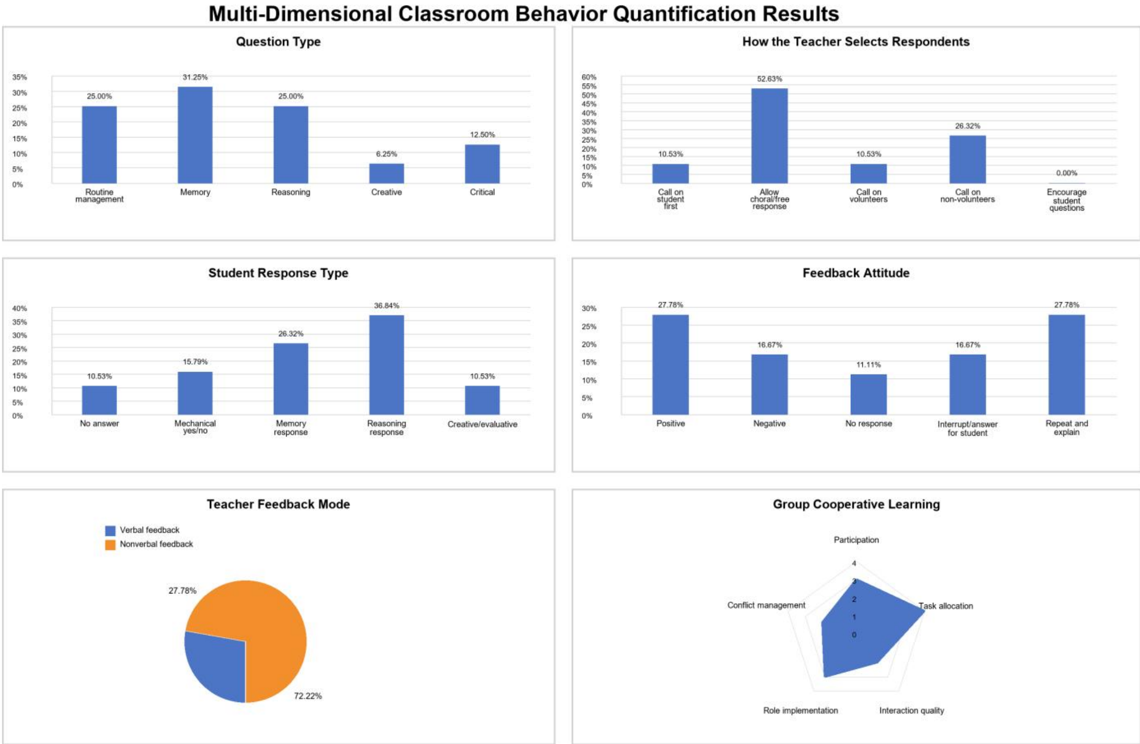

Figure 1. Quantitative Results of the Multi-Dimensional Classroom-Behaviour Analysis.

4. Conclusion

This study developed a lightweight intelligent assistant for classroom teaching diagnosis in higher vocational colleges on the basis of prompt engineering and a school-specific knowledge base. The proposed system addresses the subjectivity, high cost and vocational misalignment of conventional classroom evaluation by separating static lesson-design assessment from dynamic classroom-behaviour quantification and by preserving a clear distinction between objective statistics and vocational interpretation.

Three main contributions can be highlighted: first, a dual-track evaluation method that separately analyses lesson plans and classroom performance before integrating the two; second, a classroom observation scale for artificial intelligence majors with acceptable inter-rater reliability (ICC > 0.85); and third, a hardware-free collaborative model linking the college, teachers and the intelligent agent for routine teaching research. The tool has entered pilot use, but the current evidence remains limited to a school-based application scenario. Future work should expand the knowledge base, extend empirical validation across disciplines and institutions, and explore selective automation of data collection.

References

[1] Alon-Barkat T, et al. AI-Powered Framework for Assessing Teacher Performance in Classroom Interactions[J]. Frontiers in AI, 2025, 8: 1553051.

[2] Yu H T, Guo Y Y. Generative AI Empowers Educational Reform: Current Status, Issues, and Prospects[J]. Frontiers in Education, 2023, 8: 1-15.

[3] Abuhassna H. The Information Age for Education via AI and ML: A Bibliometric Analysis[J]. International Journal of Information and Education Technology, 2024, 14(5): 678-695.

[4] LIU Y, ZHANG H, WANG L. A Camera-Based Visual Sensor Pipeline for Fine-Grained Human Activity Recognition in Classroom Scenes[J]. Sensors, 2026, 26(12): 3666.

[5] HOU R K, FUTTERER T, BUHLER B, et al. Automated Assessment of Encouragement and Warmth in Classrooms Leveraging Multimodal Emotional Features and ChatGPT[EB/OL]. arXiv preprint arXiv:2404.15310, 2024.

[6] Wang Z Q, Yang B Q, Wang S H. Enhancing Online Teaching Effectiveness Through Computer Vision Analysis[J]. Signal Image and Video Processing, 2024, 41(3): 1193-1204.

[7] Wen J C, Lin Y, Si N. Behavioral Analysis

of Classroom Interactions Supported by Generative AI[C]//Proceedings of the 2024 7th ICETM. ACM, 2024: 45-52.
[8] Coze. Coze LLM Agent Development Platform[EB/OL]. https://www.coze.cn, 2024.
[9] McCarthy B. The 4MAT System: Teaching to Learning Styles with Right/Left Mode Techniques[M]. Barrington: Excel, Inc., 1980.