Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Submission completeness and follow-up flags in AI-assisted review of undergraduate coursework

Xiangyong Duan

duanxy@gyhvu.edu.cn

Guiyang KangYang Vocational University https://orcid.org/0009-0007-1191-355X

Yuefang Sun

Guiyang KangYang Vocational University

Research Article

Keywords: AI literacy, coursework assessment, learning analytics, formative feedback, human oversight, higher education

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11871834/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: Xiangyong Duan developed the coursework platform and taught the course analysed in this study. The authors declare no commercial financial interest.

Abstract

In a retrospective observational study, we examined how often coursework with entries in all three required in-class text fields carried a follow-up flag. An artificial intelligence (AI) system screened work in an undergraduate AI literacy course at a health-professions institution in China. An instructor could accept, override or return its recommendations. We analysed 7,080 student–lesson records from 295 students in five classes, covering the same 24 lessons in each class. Of 6,775 records with entries for notes, practical work and classroom sharing, 1,588 carried a supplementation flag (23.44%, 95% confidence interval 21.79%–25.18% from a student-cluster bootstrap). These records accounted for 95.78% of all flags. The class-specific proportion flagged among records with all three entries ranged from 14.32% to 30.68%. Mean platform scores were lowest in professional application (69.89/100), where 12.47% of records were pending review. Restricting the analysis to non-pending records narrowed the gap between foundations and professional application from 17.6 to 10.6 points. A missing-work list would have omitted most flagged records. The snapshot could identify work for inspection but could not establish flag accuracy, teacher follow-up or learner revision.

Introduction

In a practical AI literacy lesson, a student can fill in every text field and still leave the instructor with a question. A note can name the tool, an account of practical work can describe a procedure, and a classroom-sharing entry can report a result. Whether those entries explain the decisions the student actually took is a separate judgement. A dashboard that counts submissions can tell a teacher when to chase missing work. It gives the teacher little basis for deciding which of the submitted pieces need a second look. This distinction matters in a compulsory course where a single instructor may receive several hundred sets of entries for one lesson.

Evaluating the output of AI tools and explaining how it can be used are central to accounts of AI literacy (Long & Magerko, 2020; Ng et al., 2021). In health-professions education, the demand for judgement extends to the limits of tools and the responsible handling of material entered into them (Masters, 2023). These goals give process evidence a place alongside the final product. For example, a plausible chart says little about whether its author chose an appropriate data range. An account of the steps taken can help the instructor ask a more precise question. A record that such an account exists cannot settle whether the explanation is adequate.

Learning analytics offers a way to study the distinction. Lockyer et al. (2013) relate the meaning of an indicator to the activity that the instructor designed. Gašević et al. (2016) showed how instructional conditions changed the interpretation of platform variables across undergraduate courses. Both observations matter when indicators are used to select work for review. A missing screenshot has a different meaning in a lesson that requires a screenshot and one that does not. Even a measure as simple as a submission rate needs a specified task, eligible population and unit of analysis.

The same care is needed when connecting review to feedback. Boud and Molloy (2013) locate feedback in a process in which information can change subsequent work. Carless and Boud (2018) describe the learner's role in interpreting that information and acting on it. Research on learning environments has accordingly examined how dashboard design supports regulation (Hu et al., 2024) and what students want from tools that help them use feedback (Jin et al., 2024). A flag in a review queue captures an earlier event. It can identify a piece of work for attention even when the subsequent conversation, revision and learning outcome remain unobserved.

We examined submission indicators and review statuses recorded at a single extraction date in five classes taking the same AI literacy course. Our primary question was how often a record with all three in-class text fields present also carried a flag for additional evidence, and how this frequency varied across classes. A secondary question concerned the use of the same records for curriculum review. How large were differences in routine scores across the three stages of the course, and how much did these differences change when analyses accounted for pending reviews and records without a learning trace?

Methods

Setting and participants

We conducted a retrospective observational study of routine records collected over one semester at an undergraduate health-professions institution in China. The same instructor taught the five first-year classes using the same digital platform. The cohort comprised all 295 students on the class registers: 60 in Nursing A, 39 in Nursing B, 60 in speech and hearing rehabilitation, 60 in rehabilitation assistive technology and 76 in rehabilitation therapy. Each student belonged to one class. The two nursing classes are distinguished throughout.

The compulsory course comprised 36 lessons of 45 minutes. The first phase covered foundations, the middle phase applied AI, and the final phase professional applications. Students came to the course with different levels of experience. Of 295 enrolled students, 289 completed an entry profile (98.0%). Among respondents, 117 (40.5%) reported no systematic prior AI instruction and 61 (21.1%) reported having no personal computer or relying on borrowing one. These aggregate responses describe the teaching context. They cannot identify which students had difficulty with particular tasks.

How submitted work entered review

For each lesson, the platform recorded preparation activities and provided three in-class text fields: class notes, an account of the practical process and a contribution for classroom sharing. Selected assignments also requested screenshots. A separate post-class entry asked students to reflect on the lesson. Our definition of text completeness required a recorded entry in each of the three in-class fields. It did not depend on the presence of an optional screenshot, completion of a post-class reflection or the text meeting a quality threshold.

The platform sent text and available images to an asynchronous AI review queue. The prompt directed the model to consider relevance to the lesson, specificity of the steps described, duplication across fields and consistency with any images supplied. Its output included a provisional judgement and advice. The instructor could accept or override the recommendation, request another review or return the work. Students could view the rationale and resubmit. The export retained a broad review status for each student and lesson. It did not preserve separately identifiable initial AI recommendations, subsequent teacher decisions and student revisions. Throughout this study, a follow-up flag means a recorded status that called for supplementation, rejection or return. It does not indicate who initiated that status or that a student received a request.

Figure 1 traces the route available to a submission, from the student's entry through provisional screening to possible teacher and learner action. Table 1 identifies which parts of that route the research extract could support.

Table 1
What each layer of the workflow represents

Layer	Available indicators	Interpretation	Unresolved question
Preparation and access	Electronic-textbook progress, video progress, check-in	Opportunity to engage with preparation materials	Attention, comprehension or learning gain
Evidence production	Notes, practical-process account, classroom sharing, task-specific screenshot	Presence of a requested artefact	Quality, authenticity or independent authorship
Review status	Review outcome and supplementation status	A status calling for follow-up on the work	Accuracy of the flag or reason for it
Teacher and learner action	Review coverage and completion status. The platform permits teacher verification.	Work accessible in the review workflow	Actual verification, feedback use or subsequent improvement
Curriculum review	Class- and stage-level score patterns	Lessons or stages for closer inspection	Cause of a score difference

Review and revision opportunities describe platform functions. Actual teacher decisions and learner responses require separate event records.

Selecting comparable lessons

The course migrated during the semester from paired sessions to individual lesson identifiers. A lesson visible in the platform was consequently an unreliable indication that all five classes had reached it. We used the existing course export rule to identify a lesson as taught in a class when more than half of

enrolled students had at least one trace for that lesson. Traces included resource progress, check-in, in-class evidence, reflection, a non-zero component score, completion or a review status.

At extraction (10 July 2026, 20:31 China Standard Time), between 24 and 36 lessons met this criterion in each class. The intersection comprised 24 of the 36 course lessons. Eleven belonged to foundations, eight to applied AI and five to professional application. We retained every enrolled student for these identifiers, including students with no trace for a retained lesson. The resulting panel contained 295 × 24 = 7,080 unique student–lesson records. The rule holds lesson identity constant across classes while restricting interpretation to lessons with substantial observable participation. The curriculum description in Online Resource 1 lists all retained identifiers.

Measures and analysis

We cross-tabulated text completeness and review status within each record. The extraction script classified blank or waiting statuses as pending and grouped labels for supplementation, rejection or return as flagged. The “passed” category also included operational labels for teacher review, acceptance or completion. Remaining non-pending labels were grouped as other. Thus an entry coded “passed” can indicate that processing was completed without specifying an explicit pass judgement. Non-pending coverage combined the first three categories. The principal estimate divided the number of records with all three fields present and a supplementation flag by the number with all three fields present. Pending records remained in that denominator. We separately calculated the share flagged among all records, the share among non-pending records and the share of all flagged records with all three fields present. These quantities answer different questions even though some of their percentages are numerically similar.

We calculated a percentile 95% confidence interval for the principal estimate using 10,000 bootstrap samples (NumPy PCG64 seed 25112026). Within each of the five classes we drew the original number of students with replacement, retained each sampled student’s entire set of lesson records, and recomputed the pooled ratio of counts. The interval describes variation across students conditional on the observed classes and lessons. Separately, we report class-specific counts and conditional percentages. Class membership overlapped with programme, so the observed spread is descriptive.

The course rubric allocated 30 points to preparation, 20 to in-class completeness, 40 to evidence quality and 10 to post-class work. The export contained an overall score and three phase subtotals (pre-class, in-class and post-class). We used the recorded overall score on its 0–100 scale. We also summed the three phase subtotals to check agreement with that field. The course score was an operational measure and had no independent validation as an AI-literacy outcome. For each student we first averaged lesson scores within each curriculum stage and then averaged the student means. This gave every student equal weight within each stage. We used the Friedman test for the three repeated stage measures and report Kendall’s W. Between-class comparisons used each student’s overall mean and a one-way analysis of variance, with η^2 as the effect size.

Two sensitivity analyses repeated the within-student stage calculation. The first retained records with at least one learning trace. The second retained records with a non-pending review status. In each analysis we required a student to have at least one eligible record in every stage. These restrictions address different features of an operational snapshot. They also change which lessons contribute to a student's mean. The export script encoded absent numeric values as zero and absent completion indicators as false. A stored zero may therefore reflect missing source information. We report the two restrictions separately and retain the original coding in the full panel.

For the course-wide implementation indicators we averaged each indicator across the 24 common lessons within student. We calculated two-sided 95% t intervals over the 295 student means. For adjusted class contrasts we fitted an ordinary least-squares model to the 7,080 scores with class and lesson identifier as fixed effects and student-clustered standard errors. Rehabilitation therapy was the reference class. The model estimates mean score differences on the original 100-point scale.

Use of generative AI

OpenAI Codex assisted with drafting and revising manuscript text, writing analysis code, and preparing tables and figures. The statistical checks and revision of Fig. 6 reported here used OpenAI Codex desktop version 26.1120.5083.0. The desktop version identifies the tool used for these checks and does not identify the model used for classroom review. The authors retain responsibility for the analysis, interpretation and submitted text.

Results

Entry profile and course participation

Among the 289 entry-profile respondents, 181 (62.6%) reported non-systematic or no prior information-technology instruction. Nearly one third (89/289, 30.8%) lacked confidence in applying AI within their major. Figure 2 sets these responses alongside prior AI instruction, career awareness and computer access. Students could report more than one support need.

Across the common lessons, notes were present in 96.78% of records, practical-process entries in 96.74% and classroom-sharing entries in 95.79%. These are separate field-level rates. Table 2 gives the student-level intervals and Fig. 3 places submission alongside resource use, check-in and review coverage.

Table 2
Implementation indicators across the 24 common lessons

Indicator	Mean (%)	95% CI
Electronic-textbook progress	88.85	[88.08, 89.62]
Video progress	97.30	[96.79, 97.81]
Pre-class check-in	60.38	[57.47, 63.30]
Notes submitted	96.78	[96.28, 97.28]
Practical-process evidence submitted	96.74	[96.24, 97.23]
Classroom sharing submitted	95.79	[95.23, 96.35]
Screenshot evidence present	24.53	[23.19, 25.88]
Post-class summary submitted	95.88	[95.28, 96.47]
Non-pending AI-assisted or teacher review status	96.99	[96.52, 97.47]
Overall completion status	90.20	[89.30, 91.09]
Record flagged for supplementation	23.42	[21.58, 25.25]

Resource indicators measure mean progress. The other indicators measure the presence of a trace or status. Intervals were calculated across 295 student means.

Submitted work and the review decision

Of 7,080 records, 6,775 (95.69%) contained all three in-class text entries. The platform recorded a non-pending status for 6,867 records (96.99%). Across the entire panel, 1,658 records carried a supplementation flag (23.42%). Among the 6,867 non-pending records the corresponding share was 24.14%. There were 264 other non-pending records whose broad status could not be assigned to the passed or supplementation categories (Table 3).

Table 3
In-class text entries and review status at the snapshot

Text entries	Passed	Flagged	Other	Pending	Total
All three present	4,934	1,588	246	7	6,775
At least one missing	11	70	18	206	305
Total	**4,945**	**1,658**	**264**	**213**	**7,080**

Counts are student–lesson records. Other includes non-pending labels outside the passed and supplementation categories.

Among the records with all three entries present, 1,588 of 6,775 were flagged (23.44%, 95% CI 21.79–25.18). Viewed from the instructor's queue, the same cross-tabulation yields a different result: 1,588 of all 1,658 flags (95.78%) occurred on work with all three text fields present. A filter that displayed only records with a missing text field would therefore have displayed 70 of the 1,658 flagged records at this snapshot. For context, 95.69% of all records had all three entries. The large share of flags on text-complete records partly reflects how common such records were. The conditional flag rate of 23.44% describes the size of the follow-up list within this group.

Review status was pending for 206 of the 305 records missing at least one text entry and for seven of the 6,775 records with all three entries. Among non-pending records, 70 of 99 with missing text (70.71%) and 1,588 of 6,768 with all three entries (23.46%) were flagged. These descriptive rates apply only to records that had reached a non-pending state by extraction. Differences in review coverage preclude a direct comparison of evidence quality across the two groups.

Figure 4 Review status across all 7,080 student–lesson records. "Other reviewed" denotes a non-pending label outside the passed and supplementation categories. "Passed" includes accepted or completed statuses under the export coding rule. AI-assisted and teacher statuses are combined

Variation across classes

Across all records in each class, the share flagged ranged from 14.72% to 30.56% (Table 4 and Fig. 5). This comparison includes records with missing text fields. Table 5 addresses the primary question by restricting the denominator to records with all three in-class text entries.

Table 4
Routine score, participation and supplementation across all records by class

Class	n	Mean score (SD)	Check-in (%)	Notes (%)	Practice (%)	Flagged (%)
Nursing A	60	83.80 (3.90)	42.85	95.76	95.69	14.72
Nursing B	39	83.46 (4.14)	77.56	97.33	97.33	30.13
Speech and hearing rehabilitation	60	82.03 (3.52)	67.50	96.74	96.60	29.17
Rehabilitation assistive technology	60	82.79 (4.34)	72.36	97.78	97.71	30.56
Rehabilitation therapy	76	84.14 (3.77)	50.33	96.55	96.60	16.67

n denotes students. SD is the standard deviation of student mean scores. Percentages use all records in the corresponding class across the 24 common lessons.

All five classes contributed examples of work that had entries in all three fields and was flagged (Table 5). The conditional share ranged from 14.32% in Nursing A to 30.68% in Nursing B. The two

nursing classes sat at opposite ends of the observed range despite sharing a programme and instructor. Their contrast illustrates the limits of interpreting a class difference as a disciplinary characteristic. Mean course scores showed less separation, ranging from 82.03 to 84.14 points ($F(4, 290) = 2.959$, $p = .020$, $\eta^2 = .039$).

Table 5
Follow-up flags among records with all three in-class text entries, by class

Class	n	All three present	Flagged	Flagged (%)
Nursing A	60	1,369	196	14.32
Nursing B	39	906	278	30.68
Speech and hearing rehabilitation	60	1,376	409	29.72
Rehabilitation assistive technology	60	1,389	423	30.45
Rehabilitation therapy	76	1,735	282	16.25
Total	**295**	**6,775**	**1,588**	**23.44**

n denotes students. The percentage denominator is the number of records with all three entries present within that class.

In the lesson-adjusted analysis, speech and hearing rehabilitation averaged 2.11 points below rehabilitation therapy (95% CI − 3.33 to − 0.88). The intervals for the other three class contrasts included zero (Table 6). These estimates describe differences between the observed classes under a common set of lesson identifiers.

Table 6
Lesson-adjusted differences in routine scores relative to rehabilitation therapy

Class	Difference	SE	95% CI	p
Rehabilitation assistive technology	-1.35	0.70	[-2.73, 0.03]	.055
Nursing A	-0.34	0.66	[-1.63, 0.96]	.609
Nursing B	-0.68	0.79	[-2.22, 0.86]	.388
Speech and hearing rehabilitation	-2.11	0.62	[-3.33, -0.88]	< .001

Estimates are score points. Standard errors are clustered by student. Confidence intervals and p values are for each contrast without multiplicity adjustment.

Scores and review progress across the curriculum

Mean scores in the full panel were 87.50 for foundations, 85.84 for applied AI and 69.89 for professional application. A within-student comparison indicated that the three stage distributions differed (Friedman $\chi^2(2) = 386.49$, $p < .001$, Kendall's $W = .655$). The later stage also had the largest share of complete-text

records flagged: 621 of 1,239 (50.12%), compared with 588 of 3,226 (18.23%) in foundations and 379 of 2,310 (16.41%) in applied AI. Flags and routine scores partly reflect the same review process.

Figure 6 Mean student score across the three curriculum stages using the full panel. The annotated change is from applied AI (85.84) to professional application (69.89). Each point equally weights 295 students. The chart was drawn from the de-identified CSV using Python with code assistance from OpenAI Codex desktop 26.1120.5083.0

Review progress varied across the three stages (Table 7). Professional application accounted for 184 of the 213 pending records, although it supplied only five of the 24 lessons. All 295 students contributed an eligible record in each stage under both sensitivity restrictions. When only records with a learning trace were retained, the professional-application mean rose to 73.14. With only non-pending records it rose to 77.17. The gap between foundations and professional application narrowed from 17.6 points in the full panel to 10.6 points on the non-pending restriction (Fig. 7).

Table 7
Records available for comparing curriculum stages

Stage	Lessons	Records	No trace	Pending	Pending (%)
Foundations	11	3,245	6	11	0.34
Applied AI	8	2,360	7	18	0.76
Professional application	5	1,475	63	184	12.47
Total	**24**	**7,080**	**76**	**213**	**3.01**

Percentages use all student–lesson records in each row. All 76 no-trace records were pending and had an exported score of zero. The remaining 137 pending records had at least one trace.

Even within professional application, low scores and pending reviews were concentrated in particular lessons. All five classes contributed records to lessons 27 and 28. Lesson 27, on word processing and AI-assisted writing, had 76 pending records out of 295 and a mean of 59.94. Lesson 28, on document formatting and AI writing practice, had no pending records and a mean of 86.68. Lesson 30, on data processing and chart creation, had 86 pending records and a mean of 51.57. These are observations at a common extraction date. Time from the submission deadline to extraction was unavailable for each lesson.

The exported overall score differed from the sum of the three exported phase subtotals in 4,536 of 7,080 records (64.07%). Across the full panel, stage means calculated from the subtotal sum were 85.50 for foundations, 83.04 for applied AI and 67.28 for professional application. This calculation preserved the ordering of the stage means. The available extract does not document how differences between the total and subtotals arose.

Discussion

What an instructor can learn from two statuses

The practical finding is a count of work an instructor could overlook. At the snapshot, a missing-text filter would have hidden 1,588 of the 1,658 flagged records. The size of that overlap is invisible in separate class-level submission and supplementation percentages. Pairing the indicators at the level of the same student and lesson reveals which work a teacher would find under each view. A useful display would allow the instructor to open a populated submission from the follow-up queue and inspect the material that prompted the flag.

This result has a narrower interpretation than a measure of assessment quality. The three text indicators record the presence of entries. A flag records a judgement within this particular workflow. The research extract contains neither the underlying artefacts nor an independent judgement of their quality. The 1,588 records could include entries with generic descriptions, a missing task-specific image, disagreements about the expected level of detail or erroneous flags. The prompt explicitly allowed several grounds for follow-up, but their frequency cannot be recovered from the broad status. Treating all 1,588 as poor-quality submissions would assign a meaning that the data do not support.

For feedback design, what happens after a flag matters as much as the flag itself. Jin et al. (2024) used students' accounts of feedback to inform tool design. Here the corresponding student perspective is missing. A well-founded request may reach a learner too late to be useful, while an incorrect request may be contested or resolved immediately. Both can leave the same coarse status in an extract. Linking the original work to the reason shown, the teacher's response and the next version would make it possible to study those different episodes. It would also give instructors a more defensible basis for deciding which requests should recur in future lessons.

Reading scores alongside review progress

The professional-application stage provides a second illustration of the value of keeping statuses visible. An instructor seeing only a stage mean of 69.89 might reasonably prioritise the entire stage for redesign. Seeing that two neighbouring writing lessons scored 59.94 and 86.68, with different proportions of work still pending, suggests a more specific investigation. The teacher could check the submission windows, the task instructions, the evidence requested and the time available for review in those lessons. These records identify where to start that inspection. They do not distinguish a difficult assignment from a slow review queue, an absent student or a change in scoring practice.

Restricting analysis to non-pending records reduced the stage gap by about seven points. This restriction cannot estimate the score that would have been observed had every review finished. Work that reached a non-pending state by the snapshot may differ systematically from work still in the queue. Yet the professional-application mean remained lower in all three analyses. A future study could compare lessons at the same interval after their submission deadlines and follow the same records

through final grading. That design would separate changes within a record from changes in the mix of records being averaged.

The class results likewise favour inspection before intervention. Although the same person taught all five classes, the two nursing classes differed considerably in conditional flag frequency. The data supply no independent measure of a class's evidence quality and no estimate of teacher time saved. Research with vocational teachers has examined digital-skills training (Duan & Ismail, 2025) and factors associated with teachers' digital competence (Duan et al., 2026). The present study held instructor identity constant across classes. In this cohort, a small overall class score difference offers little basis for assigning remediation to an entire programme. A comparison of sampled assignments against a common rubric would be more informative.

Scope and limitations

The study followed one instructor, one institution and one semester. Selecting identical lesson identifiers reduced the risk of comparing different curricula across classes. The participation threshold also excluded lessons with little activity in any class. In particular, the five retained professional-application lessons do not represent the entire final block of the 36-lesson course. Students and lessons were observed repeatedly under the same local workflow. The bootstrap interval addresses clustering within students for these particular classes and tasks. It provides no estimate of between-institution or between-instructor variability.

Review categories were derived from operational labels. The export combined preliminary AI and teacher decisions and grouped some unclassified non-pending labels together. It lacked versioned model and prompt histories, an independent set of rubric ratings, documented override sequences and stable timestamps for feedback viewing and resubmission. It therefore supports estimates of flag frequency but leaves accuracy, consistency, turnaround time and learner uptake unresolved. A comparison of AI recommendations with blinded ratings from two instructors, stratified by lesson and submission type, could estimate false flags and missed cases. A subsequent process study could examine who reads, contests or acts on those recommendations.

Finally, review contributed to the nominal course rubric, and the source of discrepancies between the overall score and phase subtotals is unresolved. This limits interpretation of the absolute score differences. Task requirements, curriculum stage and position in the semester also varied together. The stage contrasts should be used to select lessons for closer examination. Claims about learning gains would require a separate assessment with comparable tasks over time. A privacy-conscious extension could retain event dates, decision roles and version links while limiting access to the work itself (Pardo & Siemens, 2014).

Conclusion

In five undergraduate AI literacy classes, nearly one in four student–lesson records with entries in all three in-class text fields still carried a follow-up flag. For a teacher choosing what to inspect, linking submission and review status would expose a substantial set of records that a missing-work list would overlook. The lower scores in professional application warrant lesson-level follow-up, particularly where many reviews remained pending. Determining whether the workflow made judgements more accurate or helped students revise requires independent ratings and a record of what happened after a flag.

Declarations

Ethics approval and consent to participate

This retrospective analysis used de-identified records generated during routine teaching and introduced no research-specific activity. Student identifiers were replaced with salted hashes. Names, student numbers, IP addresses, free text and screenshots were excluded from the research extract. No identifiable student work is reproduced. On 13 July 2026, the ethics review committee for medicine and laboratory animals at Guiyang KangYang Vocational University approved this retrospective study following regular review (approval no. GKD-LS-Y-2026-025) and waived the requirement for individual informed consent. The title recorded in the ethics application was “Human-in-the-loop AI for formative assessment in an undergraduate AI literacy course: a multi-class learning analytics study”.

Competing interests and funding

Xiangyong Duan developed and taught with the platform evaluated in this study. The authors declare no commercial financial interest. This work received no specific grant from a public, commercial or not-for-profit funding agency.

Author contributions

Xiangyong Duan contributed to conceptualisation, curriculum design, software, methodology, investigation, data curation, formal analysis, visualisation, project administration, drafting and revision. Yuefang Sun contributed to review and editing.

Acknowledgements

The authors thank the students and teaching colleagues who contributed to the course.

Data availability

Online Resource 1 contains the de-identified student–lesson dataset, aggregate entry-profile data, the curriculum and workflow description, and the Python analysis script. Original student submissions and

screenshots are withheld to protect privacy.

References

1. Boud D, Molloy E (2013) Rethinking models of feedback for learning: The challenge of design. Assess Evaluation High Educ 38(6):698–712. https://doi.org/10.1080/02602938.2012.691462
2. Carless D, Boud D (2018) The development of student feedback literacy: Enabling uptake of feedback. Assess Evaluation High Educ 43(8):1315–1325. https://doi.org/10.1080/02602938.2018.1463354
3. Duan X, Ismail MN (2025) Digital ability enhancement strategy and effect evaluation of vocational education teachers based on improved BP and KKT model. J Artif Intell Technol 5:563–573. https://doi.org/10.37965/jait.2025.0837
4. Duan X, Ismail MN, Abdul Wahid K (2026) Determinants of digital competencies among vocational college teachers in Southwest China: A quantitative analysis. Int J Inf Visualization 10(3):951–958. https://doi.org/10.62527/joiv.10.3.5430
5. Gašević D, Dawson S, Rogers T, Gasevic D (2016) Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. Internet High Educ 28:68–84. https://doi.org/10.1016/j.iheduc.2015.10.002
6. Hu Y-H, Liao B-K, Hsieh C-L (2024) Effectiveness of a gamified learning analytics dashboard with coregulation mechanism for self-regulated learning in college ethics courses. Interact Learn Environ 32(10):6628–6644. https://doi.org/10.1080/10494820.2023.2277741
7. Jin F, Maheshi B, Martinez-Maldonado R, Gašević D, Tsai Y-S (2024) Scaffolding feedback literacy: Designing a feedback analytics tool with students. J Learn Analytics 11(2):123–137. https://doi.org/10.18608/jla.2024.8339
8. Lockyer L, Heathcote E, Dawson S (2013) Informing pedagogical action: Aligning learning analytics with learning design. Am Behav Sci 57(10):1439–1459. https://doi.org/10.1177/0002764213479367
9. Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*(pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727
10. Masters K (2023) Ethical use of artificial intelligence in health professions education: AMEE Guide 158. Med Teach 45(6):574–584. https://doi.org/10.1080/0142159X.2023.2186203
11. Ng DTK, Leung JKL, Chu SKW, Qiao MS (2021) Conceptualizing AI literacy: An exploratory review. Computers Education: Artif Intell 2:100041. https://doi.org/10.1016/j.caeai.2021.100041
12. Pardo A, Siemens G (2014) Ethical and privacy principles for learning analytics. Br J Edu Technol 45(3):438–450. https://doi.org/10.1111/bjet.12152

Figures

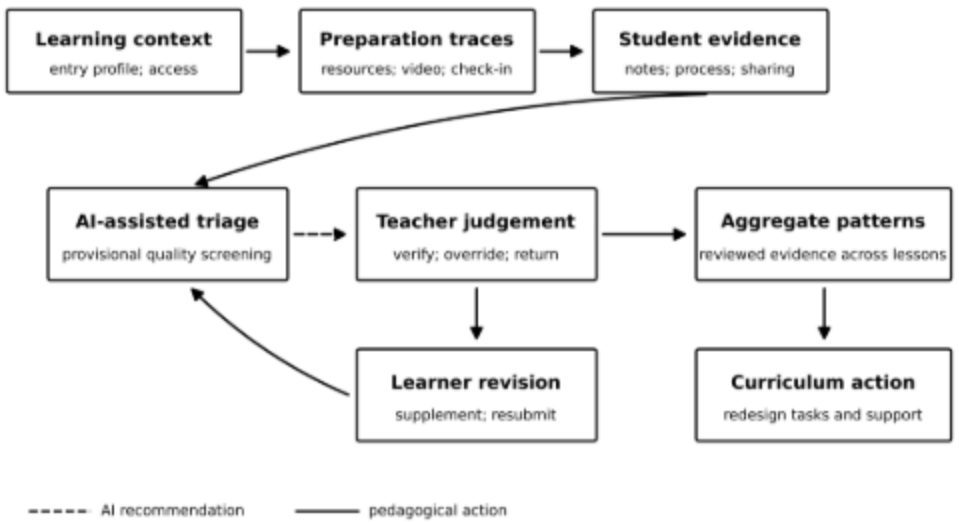

Figure 1

Designed workflow for AI-assisted review. The dashed arrow conveys a provisional recommendation. The teacher could return work and the learner could revise it. The extract did not track those actions separately

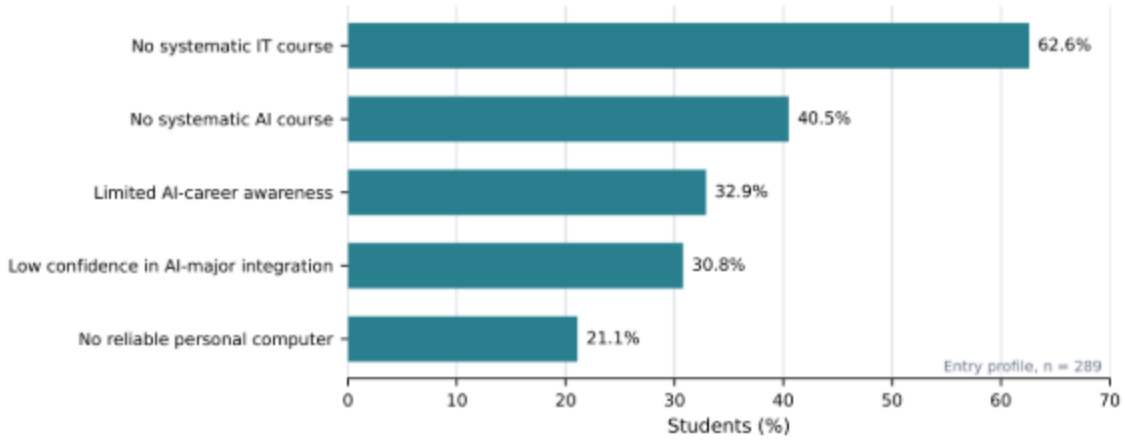

Figure 2

Self-reported support needs at course entry. Each percentage uses the 289 respondents as its denominator. The IT category includes respondents who reported little systematic instruction or none. Computer access refers to having no personal computer or relying on borrowing one

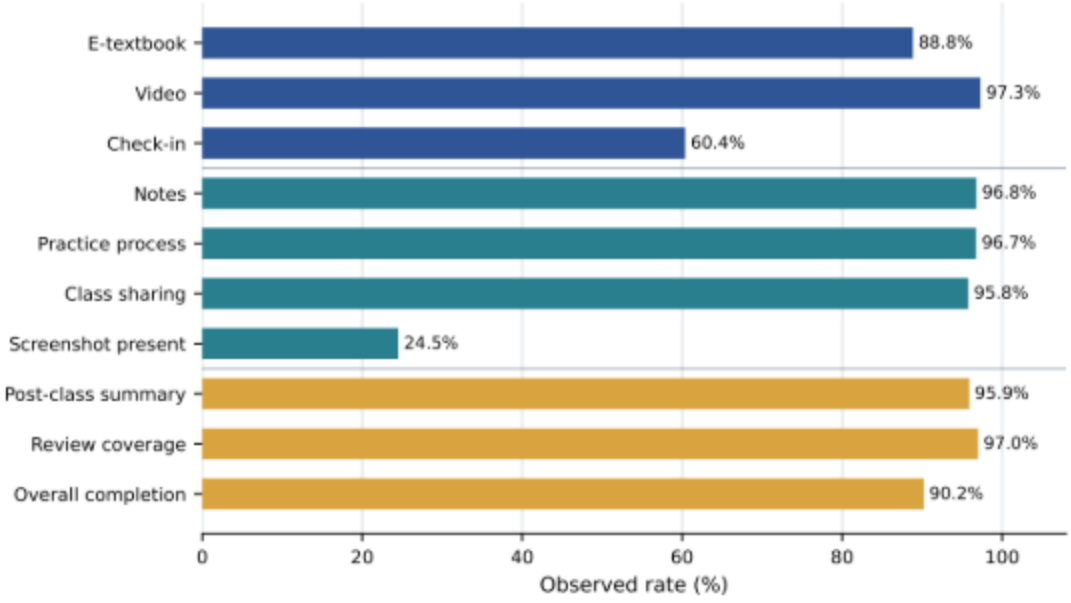

Figure 3

Preparation, submission and review indicators in the common-lesson panel. Resource bars show mean percentage progress. Other bars show the percentage of records with that indicator. Only selected assignments requested screenshots

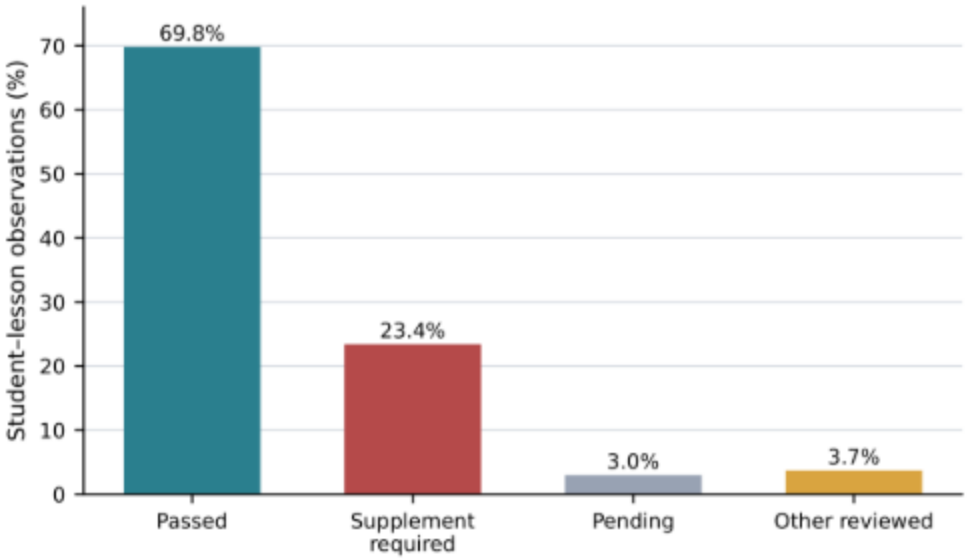

Figure 4

Review status across all 7,080 student–lesson records. “Other reviewed” denotes a non-pending label outside the passed and supplementation categories. “Passed” includes accepted or completed statuses under the export coding rule. AI-assisted and teacher statuses are combined

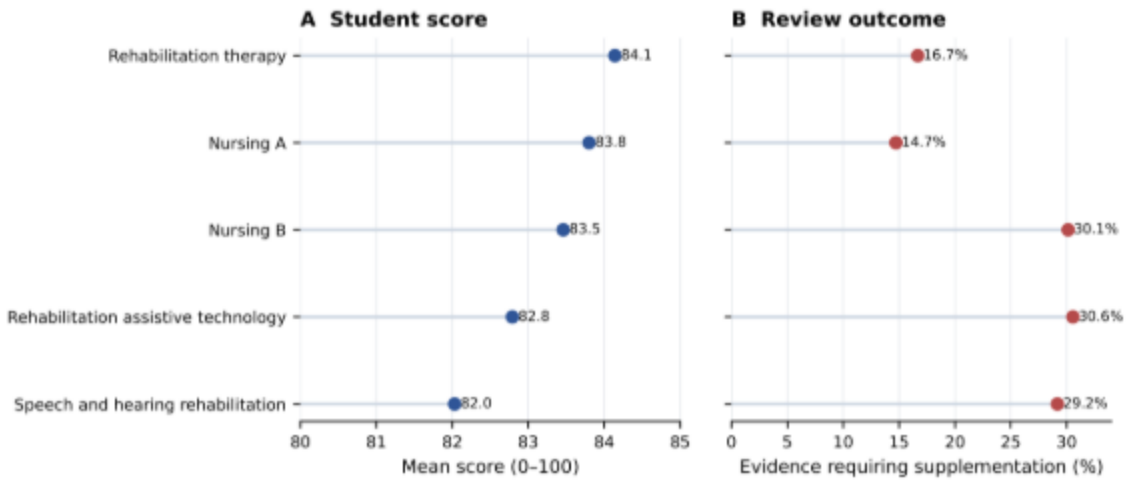

Figure 5

Mean routine score (panel A) and percentage of all class records flagged for supplementation (panel B). The left axis is expanded to show differences on the 100-point scoring scale

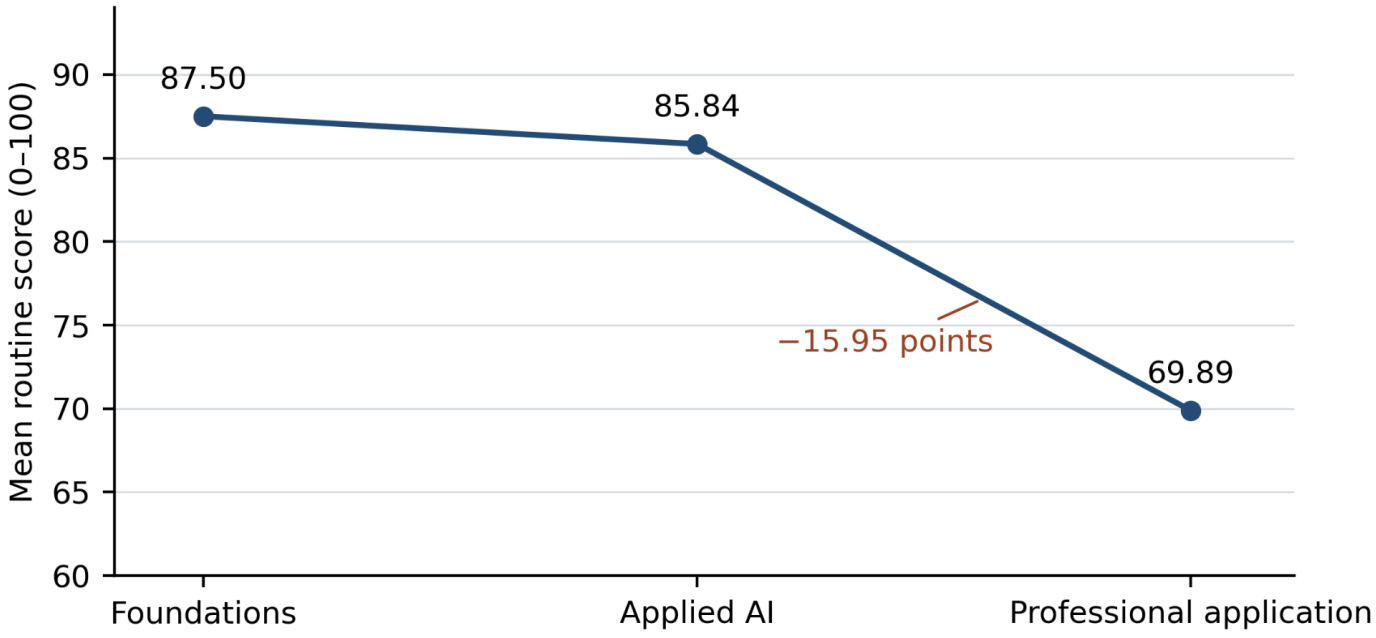

Figure 6

Mean student score across the three curriculum stages using the full panel. The annotated change is from applied AI (85.84) to professional application (69.89). Each point equally weights 295 students. The chart was drawn from the de-identified CSV using Python with code assistance from OpenAI Codex desktop 26.1120.5083.0

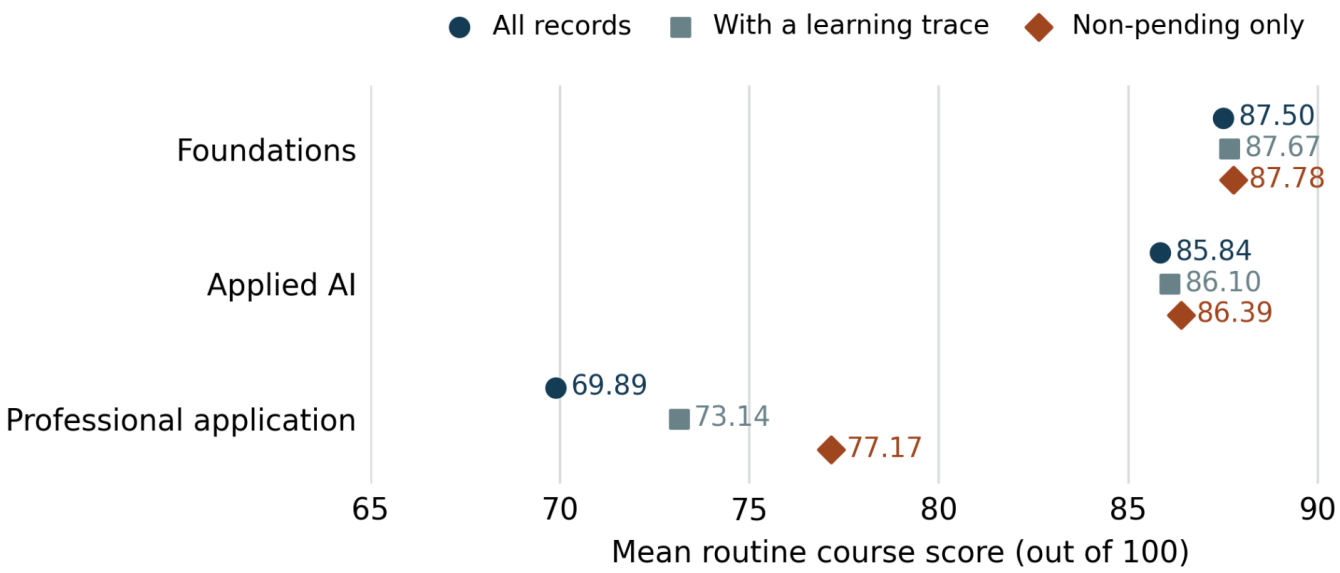

Figure 7

Mean routine score by curriculum stage under three record-selection rules. Each estimate equally weights the 295 students after averaging their eligible lessons within stage. The horizontal axis spans 65–93 points on the 100-point scale

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip