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.

A Generative Feedback System to Support Failure At-risk Online Learners

David Bañeres

dbaneres@uoc.edu

Open University of Catalonia

Anna Espasa

Open University of Catalonia

Montserrat Martínez Melo

Open University of Catalonia

Pau Cortadas

Open University of Catalonia

Research Article

Keywords: feedback, failure risk, generative tools, early warning system, artificial intelligence, online learning, higher education

Posted Date: May 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9538769/v1

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

Additional Declarations: No competing interests reported.

Abstract

Educational environments are evolving rapidly as predictive analytics and generative artificial intelligence (AI) become more deeply integrated into the learning processes. One example is the provision of personalized feedback in large-scale learning environments. This study proposes a new semiautomated framework that delivers timely, high-quality, personalized feedback with corrective and suggestive information while reducing the teacher's time commitment. Generative tools are used to combine AI-generated content, considering the learner's risk of failure provided by predictive analytics, while maintaining teacher assessment and oversight. The approach has been tested in a fully online first-year course, with 566 participants from 918 enrolled learners across different degrees in the Faculty of Economics and Business at a fully online university. The results provide insights into the effectiveness of feedback on learning processes, enhancing engagement, reducing dropout, and improving academic performance. Furthermore, learners reported positive perceptions of the use of AI-generated feedback in their learning process, giving them a clearer awareness of the objectives needed for success.

Introduction

Feedback is a fundamental support through which learners acquire knowledge and skills and remain engaged in a course (Evans, 2013; Hattie & Timperley, 2007; Winstone et al., 2017). The main objective is to provide specific information to fill the gap between the desired and the actual understanding (Sadler, 1989). However, providing feedback requires considerable effort from teachers, who must balance personalization with time constraints to produce and deliver it (Dhananjaya et al., 2024). Personalized feedback motivates learners to a higher level, promotes better self-regulation, contributes to better knowledge acquisition, and enhances engagement (T. W. Kim, 2023; Wang & Lehman, 2021). Moreover, feedback may enable a dialogue between actors to foster the learner's knowledge construction (Espasa et al., 2018; N. Winstone & Carless, 2019; Yang & Carless, 2013).

Thus, automating feedback generation has been sought as a promising solution. Although different terminology has been used in the literature, i.e., automated feedback system, intelligent tutoring system, online feedback tool, teaching platform, among others; the common aim of automatization is to increase teachers' efficiency (Xie & Li, 2018). Different techniques have been used in the past with positive results, i.e., natural language processing for essay tasks (Akçapinar, 2015), solution comparison in programming (Keuning et al., 2018), or data-driven dashboards for collaboration tasks (Bodily et al., 2018). Nevertheless, artificial intelligence (AI) techniques offer an innovative opportunity in educational settings to promote automated feedback. At first, pre-trained language models, like BERT (Bidirectional Encoder Representations from Transformers), enabled semantic understanding capabilities for activities assessment (Jia et al., 2021). Nowadays, generative tools (GenAI) have added analysis, interpretation, and text generation (Dai et al., 2024) that can produce meaningful descriptions.

Due to the potential of GenAI tools for feedback automation, different studies have analyzed their capabilities to produce feedback, detecting some flaws based on common GenAI problems, such as

hallucinations (Jia, Cui, Xi, et al., 2024) or conceptual limitations in terms of lack of quality feedback, i.e., only cognitive, mostly negative, and no self-regulation information (Dai et al., 2024). Thus, at this stage, teacher oversight is still highly recommended (Naz & Robertson, 2024). In order to produce additional types of feedback, such as self-regulation or self-level, some authors have used analytical tools to gather and process learners' data, and automated feedback tools to produce simple recommendations based on dashboards (Davis et al., 2017). However, simple visualizations do not always provide sufficient support for self-regulation (Matcha et al., 2020), increasing the efforts to produce meaningful information in terms of recommendations or reminders from data (Afzaal et al., 2021; Yan et al., 2024).

Related to the application of analytical tools in the learning process, authors of this current work previously explored how an early warning system (EWS) can improve support for learners by detecting at-risk failure in individual courses. An EWS was developed, denoted as <SYS_BLINDED>, that can raise different alerts represented in a semaphore metaphor to notify teachers and learners about the potential risk of failure (BLINDED). This alarm served as automated additional feedback to learners after each activity was graded, promoting awareness of their current state. Additionally, the alarm was complemented with information tailored to the at-risk level, providing guidelines and recommendations to support learners' self-regulation process. However, at-risk information was not aligned with teachers' feedback, leading some learners to experience inconsistencies in the received recommendations.

This work proposes an integrated approach, denoted as <APP_BLINDED>, that combines cognitive and self-regulation feedback (i.e., teachers' activity feedback and at-risk recommendations from the EWS) with a generative tool. This approach aims to help teachers provide feedback efficiently at scale while enriching feedback quality and learners' experience. The main contributions of this paper are 1) low impact on teachers' work, although oversight is present, 2) high-quality personalized feedback on general-purpose activities, and 3) better learners' performance and satisfaction. As far as we know, there is no previous work with such contributions that provides feedback for general-purpose activities by adding predictive analytics capabilities and ensuring teacher oversight.

Theoretical framework and background

Feedback as a cornerstone in online education

Feedback is an essential support in the teaching-learning process for identifying gaps between the objective and the actual understanding. Following Hattie & Timperley (2007) model, effective feedback must include information about feed-up (i.e., goal setting), feedback (i.e., learner's progression), and feedforward (i.e., recommendations towards knowledge mastery). Feedback influences learner outcomes at four levels: task, process, self, and self-regulation. While task-level aims to identify errors and areas for improvement, and process-level focuses on the method used to solve the task (Giamos et al., 2024), self-level promotes awareness of the current state, and self-regulation enables the achievement of learning outcomes supported by effective guidance (Barnard et al., 2009; Cho & Shen, 2013; Wang et al., 2021).

Since feedback should be adapted to the learner's needs, personalized feedback has been proven as an effective methodology to communicate learners' strengths and weaknesses in their learning, improve self-regulatory skills, enhance engagement, and group belonging (Giamos et al., 2024; Kaufmann & Vallade, 2022; Wurster et al., 2021). Different authors have evidenced their impact on learners' self-regulation and reflection about their progress, which leads to a deeper cognitive level (Wang & Lehman, 2021). However, producing such feedback requires some effort from teachers who, in the end, need to apply an intermediate solution between quality and time constraints (Xie & Li, 2018). Therefore, automated feedback has been used as a promising solution to reduce workload and deliver personalized high-quality feedback.

Automated feedback as a supporting tool

Different literature reviews explore the advantages, drawbacks, and potential future works (Cavalcanti et al., 2021) focused on effects, technologies, and methods for automated feedback generation, giving insights that feedback promotes learners' performance, but there is a lack of evidence about teachers' workload reduction or enriched feedback effects. Similarly, Deeva et al. (2021) reviewed automated feedback systems focusing on the technology used, the type of feedback, educational contexts, and evaluation. The authors claimed that automated feedback should be more personalized to learners' needs by using data produced during the learning process. Following this recommendation, personalized feedback was reviewed on Maier & Klotz (2022) by exploring specific tools in different contexts. The authors concluded that higher personalization is obtained when the system is specifically designed for a specific domain. For this reason, many works focused on designing specific solutions, mostly in computer science, technology, engineering, and mathematics (STEM) domains (Keuning et al., 2018). In such domains, producing feedback that compares the learner's submission with a golden solution helps achieve the objective, task, or competence through a trial-and-error approach.

Automated feedback was also explored in other domains using different commercial tools, such as Ontask (Pardo et al., 2018), E2Coach (Huberth et al., 2015), or Sonic Divider (Kickmeier-Rust et al., 2014) based on a rule-based approach. Feedback is generated based on conditions related to grades, Learning Management System (LMS) events, or acquired competences. Although the impact on performance and expectations is positive, rule-based systems require some effort and expertise to configure.

Nowadays, GenAI tools are increasingly being used in education with little understanding of how decisions are made. They are used to produce feedback, but also to automatically assess activities. Some evidence suggests that automated assessment benefits some domains, such as language learning (Escalante et al., 2023), writing skills (Banihashem et al., 2024) or multimedia designs (Almasre, 2024). Although automated assessment seems to be one of the future advantages of GenAI in education, it is currently not recommended by some institutional or even governmental policies (Commission, 2021), since GenAI tools are error-prone systems due to hallucinations (Jia, Cui, Xi, et al., 2024). However, automated feedback generated by GenAI does not have such restrictions. For this reason, some works focused on evaluating accuracy (Kinder et al., 2025), acceptance (K. Qu & Wu, 2024;

Ruwe & Mayweg-Paus, 2024) and perceptions of the different actors (Bewersdorff et al., 2025; Lee & Song, 2024; Nazaretsky et al., 2024) with diverse conclusions. In terms of accuracy, GenAI tools can be capable of producing high-quality feedback (Wambsganss et al., 2022), but teacher-in-the-loop is highly suggested (Dai et al., 2024; H. Kim et al., 2024; Lin & Crosthwaite, 2024). In terms of perceptions, other studies show that feedback generated by AI promotes greater engagement (K. Qu & Wu, 2024) and is preferred for certain tasks (Lee & Song, 2024). However, human touch remains irreplaceable (Jia, Cui, Du, et al., 2024).

Although GenAI tools can provide high-quality feedback, content analysis reveals some flaws. GenAI tools tend to produce extensive, complex, and polarized (i.e., mostly negative) feedback, which weakens their effectiveness. Furthermore, GenAI is currently unable to produce feedback about self-regulation. Thus, learners know about how it is going but not where they are going. As Pardo et al. (2019) and Yan et al. (2024) suggested, feedback should consider learners' activities, as well as data about their learning process, to provide effective recommendations.

Early Warning Systems in the feedback loop

When citing works on EWS, researchers tend to initially describe the impact of the Course Signals at Purdue University (Arnold & Pistilli, 2012). The tool provided support for teachers, learners, and tutors focusing on enhancing performance, retention, and satisfaction. The warning-level information was provided to the different stakeholders through dashboards, and the university designed strategies to help learners in case of detecting failure or dropout risk, such as sending recommendations, making phone calls, or scheduling face-to-face meetings.

Following a similar approach, other tools were developed depending on the educational setting: online courses (Moreno-Marcos et al., 2019; Nagrecha et al., 2017; Srilekshmi et al., 2017; Xing et al., 2016) or face-to-face learning (Knowles, 2014; Márquez-Vera et al., 2016; Niyogisubizo et al., 2022; Soumya & Krishnamoorthy, 2022). Additionally, they can detect different risks such as institutional dropout (Vega et al., 2022), course dropout (Mubarak et al., 2020; Whitehill et al., 2017) or failure within a course (Casey & Azcona, 2017; Chen & Cui, 2020; Rafique et al., 2021; You, 2016).

Such risk conditions can be identified by applying learning analytics strategies (Kew & Tasir, 2022; Li et al., 2022). Analytical tools increase self-awareness and can be enhanced with predictive models. In recent years, full-fledged developments have emerged that have provided dashboards to teachers (Najdi & Er-Raha, 2016; Wolff et al., 2014), learners (Hu et al., 2014; Ortigosa et al., 2019), and tutors (Llauró et al., 2021; Qu et al., 2022) and empowered each stakeholder with crucial information to help learners sidestep the identified risk. However, EWS dashboards have a low influence on reversing the at-risk situations (Matcha et al., 2020).

Providing feedback or recommendations is the answer to the problem (Clow, 2012; Xavier & Meneses, 2022). Some systems use only analytical data on learners' grades, access to information, or activity

completion. Such data-oriented feedback can contribute to self-leveling, which some studies have reported as positively affecting learners (Mousavi et al., 2021). However, EWS with predictive capabilities could enhance self-regulation to guide future improvement. Some EWS include feedback capability in terms of when a risk is detected (Burgos et al., 2018; Márquez-Vera et al., 2016; Ortigosa et al., 2019; Vasquez et al., 2015), providing behavioral feedback combined with goal setting (Latham & Locke, 2007; Locke & Latham, 2002) or self-goal setting (Elliot & Fryer, 2008; Zimmerman, 1990).

Based on the aforementioned insights, the authors of this work previously proposed the <SYS_BLINDED> system, which provides analytical information and feedback capabilities by informing learners and teachers about failure (BLINDED) and dropout risk (BLINDED). The system is fed with LMS available data, and risk levels are computed using predictive models. Feedback on risk recommendations, additional resources, and guidance is provided to learners based on the detected risk level to prevent subsequent risk situations (BLINDED). Currently, such feedback is proposed by the teacher, and the system manages the delivery based on learners' risk levels. However, since <SYS_BLINDED> does not consider teachers' feedback at the task or process level, the cognitive potential of feedback is weakened (Wang & Lehman, 2021). In this work, the baseline <SYS_BLINDED> system has been enhanced with feedback from those levels. Such a new approach, denoted as <APP_BLINDED>, has automated feedback generation using GenAI tools to produce high-quality feedback while minimizing teachers' workload, even though the text is reviewed by them.

This contribution has been evaluated by analyzing the following research questions by testing the system on a specific first-year higher education course:

1. Is the failure risk-level identification accurate for self-level and self-regulation feedback?
2. Have learners' submissions increased during the continuous assessment when using the <APP_BLINDED> approach?
3. Has learners' performance increased when using the <APP_BLINDED> approach?
4. Do learners consider that the <APP_BLINDED> approach is significantly different in terms of content, utility, and personalization from the <SYS_BLINDED> baseline?

Methodology

A fully online university

The <BLINDED_UNIV > is a fully online university with a learner-centered educational model, with a focus on competence acquisition. The assessment process is based on a continuous assessment (CA) model that includes a set of activities denoted as Continuous Assessment Activity (CAA). Additionally, the CA is combined with a summative assessment at the end of the semester in the majority of courses, which is a final test (FT).

The final mark (FM) for each course is calculated using a formula that assigns different weights to CAA and FT based on the content's significance within the course. The grading system combines a qualitative

grade during the CA with a quantitative one at the end of the course. Each CAA is graded by the scale: A (very high), B (high), C+ (sufficient), C- (low), and D (very low), where a grade of C- and D means failing the CAA. There is also an additional grade (N, non-submitted) to mark a not-submitted CAA. The FM is a grade between 0 and 10 that combines the qualitative grades of each CAA and the quantitative grade of the FT. A value below 5 indicates failure in the course.

The learning process takes place on the Canvas LMS platform (Instructure, 2025) that includes learning materials, custom learning tools, CAA, and communication channels (i.e., announcements and forums). It stores all digital traces of learners, including submitted CAA, performance, and interactions with the LMS (i.e., navigational data, communication, accessed resources, and tools).

<BLINDED_UNIV> learner profiles differ significantly from those at face-to-face universities. They are mostly full-time employed and have family commitments, which condition their learning process (Sánchez-Gelabert et al., 2020; Xavier et al., 2026). They pursue new studies to improve their professional career or expand their knowledge in a specific domain.

Teachers offer support to promote participation and guide learners. Communication is based on asynchronous and mainly written texts. Feedback is mainly provided once the CAA is assessed, using different strategies, e.g., exemplars, general or individual comments highlighting common mistakes and suggesting ways for improvement, or through a rubric designed by the teaching staff (Guasch et al., 2019). The number of learners and the activity type constrain feedback personalization.

Research design and participants

A mixed research design combining a design-and-creation approach (Kuechler & Vaishnavi, 2012) and an action research methodology (Oates, 2006) has been implemented. The former is used to detect the problem (i.e., to enhance the support to learners); to suggest a solution (i.e., to improve the feedback on each CAA by combining the generated GenAI feedback with the at-risk suggestions); and, finally, to develop, evaluate, and test the proposed solution. The latter is used in this final step to iteratively develop the artifact through plan-act-reflect cycles. Each cycle comprises solution development, a test in real settings with learners and teachers, and an evaluation of the product outcomes using different measures. This evaluation aims to gain knowledge, assess expectations, and either start a new cycle or finish product creation.

This paper focuses on the first iteration of the <APP_BLINDED> approach which focuses on combining the teacher's feedback after the CAA assessment with the system’s at-risk recommendation. This cycle has been evaluated in an online course of 6 ECTS called *Markets and behavior* included in multiple degrees within the Faculty of Economics and Business. The course focuses on the microeconomy specialty by introducing learners to the characteristics of the modern economy, from supply (companies and their costs) to demand (consumers and their preferences). The assessment model follows the university policy by combining five CAA with an FT. The FM is computed as FM = 60% CA + 40% FT, subject to several conditions: the CA and the FT must be passed with grades greater than five and four

out of ten, respectively. Additionally, the CA grade is computed from the best four CAA grades out of the five to mitigate failure and dropout issues. Learners receive different feedback types after CAA is assessed: an exemplar, whole-class feedback with common mistakes, and individualized feedback upon request. This course has been selected for three reasons: 1) it is a first-year course with a large number of learners, 2) there is a high failure rate where at-risk detection could impact, and 3) providing personalized feedback at scale could effectively reduce failure and dropout rates and increase teachers' efficacy.

Figure 1 illustrates the research design. <APP_BLINDED > has been tested during the 2024 Autumn semester (i.e., from September 2024 to January 2025). The study employed a quasi-experimental design with two pre-existing learner groups equally distributed in online classrooms. The analysis focused exclusively on learners from both groups who voluntarily agreed to participate by signing a consent form, as the university's Research Ethics Committee requires explicit consent for any study, in line with the General Data Protection Regulation (GDPR, https://gdpr-info.eu/). After acceptance, one group will receive recommendations from the failure identification mechanism, denoted as the *Only Risk* approach (BLINDED). The other group will receive <*APP_BLINDED*> feedback, which combines at-risk recommendations with the teacher's feedback after each CAA has been assessed. Figure 1 shows an overall participation rate of 61.66% (i.e., 566 of 918), with unequal group sizes due to differences in acceptance rates across groups.

The pilot evaluation was conducted after the course concluded across three aspects: at-risk identification accuracy, course performance, and learners' expectations. Accuracy was evaluated for non-signed learners, as they are not affected by any feedback approach, while performance and expectations were evaluated across acceptance groups.

Study procedure and instruments

As previously stated, this iteration focuses on integrating CAA feedback with at-risk recommendations within the <SYS_BLINDED> system, whose capabilities have been previously described in multiple contributions. Only the at-risk failure detection has been used to avoid multiple treatment interference, which uses the <MODEL_BLINDED> predictive model (BLINDED).

The failure risk level is informed in dashboards for both learners and teachers and is represented on a semaphore metaphor with different possible colors depending on the prediction provided by the <MODEL_BLINDED> model; the model accuracy (i.e., the percentage of correct failure and pass cases identified in the training phase) regarding a confidence threshold to consider a reliable model (i.e., the accuracy percentage to consider an accurate model); the submission event and the grade of the current CAA; and the number of consecutive CAA that have not been submitted for a learner. Figure 2 shows the color distribution under these conditions, along with descriptions of each color for teachers and learners, and the internal code to reference each risk in the following sections. In this work, the threshold value was set to 75% based on an institutional accuracy study performed for the failure model (BLINDED).

The failure risk information is given as additional support to learners in a two-step process on each CAA. Once a CAA has been started, each learner receives the potential risk distribution based on the grades they could obtain on the CAA assessment (See Fig. 3). It is represented as a colored bar distribution, computed by simulating the risk associated with the range of possible grades using the <MODEL_BLINDED> model. This preliminary information provides the first alert about potential risks and feed-up information (i.e., goal setting). Learners know from the beginning of the CAA the minimum grade required to stay out of risk, and they must endeavor to acquire the CAA knowledge, submit the activity, and pass it with the minimum grade stated in the risk bar distribution. The second support occurs after the CAA is assessed and the real failure risk is revealed. Learners know the grade and, therefore, the risk associated with such a grade. Simultaneously, learners receive at-risk recommendations depending on the risk level. As shown in Fig. 2, although learners see only green, amber, and red, teachers have a broader range of risk colors with distinct meanings. Thus, teachers can tailor recommendations for each meaning to provide additional support and personalization. These recommendations are included in the system before the assessment phase (i.e., one recommendation per risk level and CAA) and are automatically delivered when the risk levels are revealed to the learners. Additionally, at this point, a new bar distribution is presented to each learner for the next CAA. It is worth noting that the <MODEL_BLINDED> predictive model considers profile information about the learner (i.e., newbie at the university, simultaneous enrolments in the current semester, repeated enrolments in the current course, and the grade point average) and all previous CAA grades to date. Thus, predictions may be different for two learners with the same grades but different profiles.

This paper enhances this information provided to learners when they receive their grade. The risk recommendations are combined with feedback on the assessed CAA using a GenAI tool. <APP_BLINDED> feedback generation is divided into a validation and a creation phase as depicted in Fig. 4. The validation phase is conducted before the CAA assessment, during which the assessment criteria and the competences to be addressed for each CAA exercise are defined. A Llama Large Language Model (LLM) (Grattafiori et al., 2024) is used to generate different feedback fragments for each competence. The prompt has been engineered to provide short, meaningful, localized (i.e., Spanish and Catalan), and teacher-gender perspective feedback for each addressed competence. After different experiments, we observed that generating five text fragments per competence is enough to obtain diverse feedback. Then, teachers review those fragments and decide whether to improve, discard, or add more fragments, focusing on selecting the most suitable feedback fragments for each competence to better support learners. Although one feedback fragment per competence might be enough, we consider that offering different possibilities and presenting them randomly to learners enriches the learning process, as learners sometimes discuss the received feedback among themselves.

During the creation phase, when a CAA is assessed, the teacher assesses each exercise and records the competence acquisition in a custom-format document. Finally, feedback is produced for each learner. The feedback includes both corrective and positive feedback, along with suggestions, and the risk-level recommendations stored within the EWS. The feedback fragments are randomly selected from the validated set for each competence, based on the teacher's gender and the learner's language

preferences, creating a personalized feedback message for each learner. It is worth noting that the validation step before the assessment process allows skipping the revision of the complete feedback messages, since all text fragments have been previously reviewed. This decision greatly increases the teacher's efficiency and minimizes their effort.

Finally, a questionnaire was designed and validated through expert consultation, ensuring that the questions were clearly understood and appropriately formulated. The questionnaire covered general perceptions of the feedback received (see Table 3 for a detailed description of the questions), grouped into the dimensions content, utility, personalization, satisfaction, and open-ended comments. The questionnaire was administered online to learners who participated in the study at the end of the course using the institutional Qualtrics tool, and responses were collected anonymously.

Data analysis

Data from different data sources have been gathered to answer the research questions. The accuracy of the <MODEL_BLINDED> model and at-risk failure identification (RQ1) have been evaluated by using data from the <SYS_BLINDED> system. Different accuracy metrics are automatically computed after the training process and at the end of the piloted semester. The <MODEL_BLINDED> model has been trained with data from six semesters (i.e., from Autumn 2018 to Spring 2021, with 4089 learners) and validated with one semester (i.e., Autumn 2022, with 1090 learners). It is worth noting that the course has changed slightly since Autumn 2022. Currently, the assessment model gives more emphasis on the CA. While the previous assessment model allowed learners to pass the course only with the FT, the CA is currently mandatory for all learners, which could impact their performance.

The performance of the <MODEL_BLINDED> model is analyzed by the following metrics:

$$TNR = \frac{TN}{TN+FP} \quad ACC = \frac{TP+TN}{TP+FP+TN+FN}$$

$$TPR = \frac{TP}{TP+FN} \quad F_{1.5} = \frac{(1+1.5^2)TP}{(1+1.5^2)TP+1.5^2FN+FP}$$

where TP denotes the number of at-risk learners correctly identified, TN the number of non-at-risk learners correctly identified, FP the number of non-at-risk learners not correctly identified, and FN the number of at-risk learners not correctly identified. These four metrics are used for evaluating the global accuracy of the model (ACC), the accuracy when detecting at-risk learners (true positive rate—TPR), the accuracy when distinguishing non-at-risk learners (true negative rate—TNR), and a harmonic mean of the true positive value and the TPR that weights correct at-risk identification (F score - F1.5). The accuracy of the failure risk-level detection is computed by comparing the percentage of correct failure detections among learners who did not consent to participate, as their performance was not affected by any feedback approach (i.e., Only Risk and <APP_BLINDED>).

Additionally, CAA submission information has been collected from the Canvas LMS to analyze the impact of the new feedback during the CA (RQ2). This data is used to conduct a statistical analysis in the R tool to determine whether belonging to groups participating in the pilot affects the CAA submission. Boschloo's unconditional test (Boschloo, 1970) is used to test the significance of the correlation between submitting the CAA and belonging to the different groups, represented as binary variables.

Next, the impact on the learners' performance (RQ3) is evaluated by gathering the FT score and FM from the Student Information System (SIS) using Oracle Discoverer. The significance of pilot participation on performance is analyzed using the Mann-Whitney test due to the final mark's non-normal distribution (Mann & Whitney, 1947).

Finally, learners' expectations were collected using the anonymous post-questionnaire instrument administered to learners who consented (RQ4). Data analysis was carried out using two basic procedures: (1) univariate and bivariate descriptive statistics and (2) bivariate inferential statistics. Statistical analysis was performed using SPSS. For the first procedure, basic descriptors such as mean (M) and standard deviation (SD) were analyzed, along with comparisons between experimental groups. For the bivariate inferential analysis and the quantitative variables, before addressing the hypothesis test, the Kolmogorov-Smirnov (K-S) normality test (Massey, 1951) was performed, and other contrasts were performed, which were positive. Given the sample size and the results of the normality tests, the Mann-Whitney test was also used for the hypothesis test. The survey was answered by 187 out of the 566 learners who consented to participate. The sample is subdivided into 64 learners who received <APP_BLINDED> feedback and 123 learners who received Only Risk recommendations. This is a small, finite sample with a coverage rate of 33%. Under the principles of simple random sampling for finite universes, with nc = 95.5% and p = q=50, the overall margin of error is ± 5.99. The sample is composed of 55.6% women, and the overall mean age is 29.3 years (SD = 9.73), with no significant differences between the two groups. Finally, an inductive thematic analysis was conducted on the open-ended comments (Braun & Clarke, 2006).

Results

RQ1. Is the failure risk-level identification accurate for self-level and self-regulation feedback?

Before knowing the impact of the <APP_BLINDED> approach, it is crucial to evaluate the accuracy of the <MODEL_BLINDED> model and risk identification system. Table 1 illustrates the accuracy metrics for the trained model on the validation test and after the semester ended for learners who did not participate. Thus, the model should predict them according to their risk level due to no additional feedback intervention.

Table 1
Accuracy of <MODEL_BLINDED> model.

Activity	Accuracy metrics on the validation test[1]				Accuracy metrics on not signed group[2]			
	ACC(%)	TPR(%)	TNR(%)	F1.5(%)	ACC(%)	TPR(%)	TNR(%)	F1.5(%)
CAA1	81.10	84.52	76.13	76.44	69.42	71.56	66.45	72.52
CAA2	86.79	90.71	81.08	82.45	77.96	80.09	75.00	80.56
CAA3	89.36	96.44	79.05	83.08	83.20	81.04	86.18	83.35
CAA4	91.47	97.21	83.11	86.53	83.47	83.41	83.55	84.65
CAA5	94.50	97.99	89.41	91.57	85.40	81.04	91.45	84.36

Notes:

[1] Train set: 4089, validation set: 1090

[2] Test set: 352

Table 1 summarizes the high-accuracy metrics that enable correct at-risk identification. On the one hand, the training process shows good accuracy in metrics for detecting at-risk (i.e., TPR) and non-at-risk learners (i.e., TNR), with values above the threshold quality of 75%. Thus, predictive models are considered high-quality, enabling the use of all risks based on the at-risk color distribution shown in Fig. 2. Note that lower accuracy values (i.e., less than 75%) would activate only *YFailLowAcc* and *YPassLowAcc* risk levels under the threshold configuration, and at-risk messages would be limited to superficial recommendations. On the other hand, the model accuracy has significantly degraded failure detection (i.e., TPR) on the piloted semester around 15%. However, non-at-risk detection accuracy (i.e., TNR) has been less impacted ranging from a 10% decrement to a 2% increment.

Figure 5 focuses on the percentage of correct failure identification for each risk level in the not signed group across the different CAA with respect to the final outcome (i.e., fail the course). At-risk levels (i.e., code risks *Red, NS1*, and *NS2* in Fig. 2) and the lowest risk, *Green*, are highly accurate. The variability in *GreenLow* and *YPassActivity* risk levels is due to the possibility of submitting four out of the five CAA, and mostly affects the first two activities. For instance, many learners who fail the first activity are classified as *GreenLow*, since the predictive model takes into account the possibility of submitting and passing the remaining activities. Both levels' accuracy improves from the third activity, CAA3, where the risks are better aligned with learners' previous performance. Consequently, the <MODEL_BLINDED> model and risk identification correctly classify nearly all potential failure instances.

RQ2. Have learners’ submissions increased during the continuous assessment when using the <APP_BLINDED> approach?

This section analyses the impact of the new approach during the CA by checking the number of submissions. Note that an increment in the number of submissions may impact the passing rate and course performance at the end of the course. Figure 6 depicts the submission rate for each group and CAA for the three groups (i.e., <APP_BLINDED>, Only Risk, and Not Signed). Note that the first CA has not been included, since learners' submissions have not been affected by any previous feedback. The impact of <APP_BLINDED > is positive and progressive across activities, with submissions remaining near 90%. The Only Risk group also outperforms the Not Signed group from 5% to 10%. Thus, supporting learners with additional feedback (i.e., <APP_BLINDED> feedback or only risk-based recommendations) impacts submission rates positively.

Boschloo's unconditional test is used to statistically evaluate whether belonging to a group is associated with submitting the CAA. Three hypotheses have been evaluated depending on the compared groups (i.e., <APP_BLINDED > vs. Not Signed, <APP_BLINDED > vs Only Risk, and Only Risk vs. Not Signed). The null hypothesis states that the proportion of non-submission is greater than or equal among learners assigned to the first group in each comparison. Based on the observations in Table 2, the null hypothesis can be rejected for all CAA comparisons between the <APP_BLINDED > and Not Signed groups. Similar rejection results can be concluded in activities for the Only Risk and Not Signed groups, except for CAA2. However, it cannot be rejected when comparing <APP_BLINDED > and Only Risk groups. These results suggest that both feedback mechanisms significantly outperform the Not Signed group in activity submissions, but there is no statistical difference between the two mechanisms.

Table 2
Boschloo's unconditional test p-values for each activity

Activity	<APP_BLINDED > vs. Not Signed	<APP_BLINDED > vs. Only Risk	Only Risk vs Not Signed
CAA2	< .001	.110	.052
CAA3	.037	.070	.003
CAA4	.002	.110	.023
CAA5	< .001	.080	.001

RQ3. Has learners' performance increased when using the <APP_BLINDED> approach?

The FM of the *Markets and behavior* course takes into account a final test. Thus, it is relevant to see whether the learners have acquired the minimum knowledge and skills to pass the course. The impact of each approach on the failure rate and the final mark distribution is analyzed. To avoid dependence on learners dropping out, their final mark (i.e., 0) has been excluded. Thus, the analysis focuses only on learners who have finished the course. For significance testing, the Mann-Whitney test was used because the final marks were non-normally distributed. The null hypothesis test is that the mark is less than or equal to the first group of each comparison. The comparison is made across all groups, and the

results are summarized in the notched box-and-whisker plot in Fig. 7, with p-values for group comparisons (values above the figure), median (in the middle of the boxes), and mean values (indicated by a diamond).

Failure rates are 29.22%, 39.44%, and 47.47% for <APP_BLINDED>, Only Risk, and Not signed groups, respectively. When comparing median and average values across groups, we observe that the <APP_BLINDED> group performs best. When comparing p-values, we observe significance in learners who participated in any of the piloted groups (i.e., <APP_BLINDED > or Only Risk) compared to learners who did not participate. Additionally, there is also significance in the <APP_BLINDED> group compared to the Only risk group. Thus, we can reject the null hypothesis and answer the research question affirmatively.

RQ4. Do learners consider that the <APP_BLINDED> approach is significantly different in terms of content, utility, and personalization from the <SYS_BLINDED> baseline?

In relation to learners' perceived experience (see Table 3), the descriptive results indicate higher scores across all indicators among learners who have received <APP_BLINDED> feedback. All mean values in the <APP_BLINDED> group are mostly above 7.0, with the content, utility, and general satisfaction dimensions being the highest-rated.

When checking significance, only two specific indicators related to content and personalization are positive. Specifically, the learners in the <APP_BLINDED> group score significantly better (M = 7.5; SD = 2.64) on the indicator *It allows me to receive feedback with more nuances, depth*, *etc.* (z=-2.366; p = 0.018). They also consider it is more personalized, so they declare *Feedback corresponds to the mistakes I made in the activities* (M = 7.8; SD = 2.70), significantly more than the Only Risk group (z=-2.777; p = 0.005).

Table 3
Learners' perception about content feedback, utility, personalization, and satisfaction with feedback.

Dimension	Indicator[1]	<APP_BLINDED > [2]		Only Risk[3]		U de Mann-Whitney[4]	
		M	SD	M	SD	z	*p*
Content	It gives me a clear perspective of what is behind the meaning of the teacher's feedback	7.5	2.56	7.0	2.60	-1.360	*0.174*
	It provides me with comprehensive feedback	7.3	2.72	6.7	2.74	-1.498	*0.134*
	It provides me with clear feedback	7.6	2.61	7.0	2.68	-1.706	*0.088*
	It makes easy for me to receive a great volume of information	7.4	2.60	6.7	2.76	-1.684	*0.092*
	It allows me to receive feedback with more nuances, depth, etc.	7.5	2.64	6.6	2.72	-2.366	*0.018*
	The risk of failure indicators has been useful for the following assessment activities	7.1	2.76	7.0	2.63	-0.350	*0.726*
Utility	I found it easy to understand	8.0	2.31	7.5	2.78	-1.319	*0.187*
	The feedback helped me learn	7.0	2.75	6.6	2.91	-0.842	*0.400*
	It allows me to have a better understanding so that I can improve my work	7.6	2.49	7.2	2.62	-0.880	*0.379*
	It facilitates reflection on the learning carried out	7.7	2.56	7.3	2.68	-1.328	*0.184*
	The feedback is useful and facilitates	7.7	2.53	7.0	2.67	-1.941	*0.052*

[4] To facilitate the results interpretation, the mean (M) and standard deviation (SD) are also shown.

Dimension	Indicator[1]	<APP_BLINDED > [2]		Only Risk[3]		U de Mann-Whitney[4]	
		M	SD	M	SD	z	*p*
	its implementation (making changes and improvements in the activity)						
Personalization	Feedback is personalized	7.1	2.75	7.0	2.80	-0.719	*0.472*
	I feel the feedback is tailored to my needs	7.2	2.62	6.7	2.80	-1.244	*0.213*
	The feedback fits how I learn	7.3	2.58	6.7	2.81	-1.280	*0.201*
	The feedback fits how I feel	6.9	2.69	6.6	2.92	-0.583	*0.560*
	Feedback corresponds to the mistakes I made in the activities	7.8	2.70	6.7	2.91	-2.777	*0.005*
Satisfaction	General satisfaction with feedback	8.0	2.19	7.5	2.37	-1.443	*0.149*

Notes:

[1] Indicators with a 0–10 scale

[2] n = 64 [3] n = 123

[4] To facilitate the results interpretation, the mean (M) and standard deviation (SD) are also shown.

Open-ended comments were also analyzed. The most relevant complaint was that the text generated was too generic, offering only a description of the errors made and the competences to be improved and reviewed. Even some learners reported sharing their feedback, finding that parts of it were identical (i.e., feedback fragments validated for a specific competence). However, while some learners described this as a drawback, others claimed that such feedback, even with GenAI tools, is an advantage, providing textual descriptions of competences to improve and insights into why the obtained grade. Also, some learners noted difficulty understanding the risk level and the associated recommendations, as reflected in the questionnaire (i.e., *The risk of failure indicators has been useful for the following assessment activities*). Some learners did not understand why a grade higher than 5 in a CAA could lead to a potential failure (i.e., *YPassActivity* is considered a risk level as shown in Fig. 2).

Conclusions, limitations, and future research

Related to RQ1 (*Is the failure risk-level identification accurate for self-level and self-regulation feedback?*), results show that the <MODEL_BLINDED> model achieves high accuracy, correctly identifying at-risk learners. On the one hand, correct risk identification provides precise recommendations on the competences needed for the following CAA, alternative learning paths to pass the course in case of low grades, and information on the minimum grade required in the next CAA to get out of the at-risk levels. On the other hand, model accuracy decreased in the piloted semester. This implies that current learners are not performing like learners used to train the models. The change in the assessment model may have conditioned or affected performance and the correlation between CA and final grades. Currently, it is mandatory to pass the CA to access the FT, while previously, learners could even pass the course without passing the CA. However, accuracy in detecting at-risk (i.e., TPR) and non-at-risk learners (i.e., TNR) is greater than 80% and 75%, respectively, from CAA2 to CAA5, allowing the delivery of the correct risk recommendation message most of the time. This is also observed in risk identification on at-risk and non-at-risk levels. Intermediate levels refine over time, leading to correct identification from the middle of the semester. Basing at-risk identification on predictive models rather than solely on past analytical data increases awareness of potential issues (e.g., failure) and reduces reliance on current learner state (e.g., current grades) (Matcha et al., 2020).

Concerning RQ2 (*Have learners' submissions increased during the continuous assessment when using the <APP_BLINDED> approach?*), results show an increase in total submissions from learners who received the <APP_BLINDED> feedback. These results are consistent with the literature, which shows that feedback impacts engagement and performance (Giamos et al., 2024). Such learners receive, in addition to the at-risk recommendation, positive, corrective, and suggestive feedback impacting task- and process-levels. When compared with learners who did not participate, all learners received the official teacher's exemplar, along with whole-class feedback highlighting the most common errors and recommendations. However, individual feedback is available upon request. In online learning, and especially in first-year courses, learners often do not request feedback due to inexperience in the learning context or because they are uncomfortable contacting the teacher. This might lead to feeling alone during the learning process, resulting in dropping out when difficulties arise (Kaufmann & Vallade, 2022). Providing additional feedback from the CAA assessment or potential risks improves engagement, group belonging, and, in the end, staying on track within the course (Wurster et al., 2021). The impact of EWS on learners' performance has been reported by other authors (Hu et al., 2014; Mubarak et al., 2020; Ortigosa et al., 2019) who found that risk information combined with feedback recommendations reduces dropout issues during the course and increases success (Rafique et al., 2021).

Related to RQ3 (*Has the learner's performance increased when using the <APP_BLINDED> approach?*), results also show that both groups participating in the pilot experience fewer failures and better performance than learners who do not participate. Risk recommendations in both groups effectively indicate learners' risk status and state short-term goals in case of at-risk detection that effectively impact further task performance (Locke & Latham, 2002). Such awareness and recommendations form the basis for self-level and self-regulation, which different authors have related to course success (Cho & Shen, 2013). Furthermore, complete feedback at the task- and process-level in the <APP_BLINDED>

approach enhance competence acquisition (Winstone & Carless, 2019), which improved learners' performance toward passing the FT and the course in this study.

Finally, concerning RQ4 (*Do learners consider that the <APP_BLINDED> approach is significantly different in terms of content, utility, and personalization from the <SYS_BLINDED> baseline?*), learners are satisfied with the additional feedback provided by the <APP_BLINDED> approach. Learners are informed about the utilization of GenAI tools for feedback creation. Thus, results may be biased by preconceptions against AI feedback as found in Ruwe & Mayweg-Paus (2024). Learners mostly agree that the content feedback was clear and comprehensible. Thus, this feedback aided learning because it was easy to understand and facilitated reflection on the self-work done during the activities, consistent with the findings in Dai et al. (2024). Feedback personalization was the lowest-rated indicator. Although the feedback addressed the errors and ways to improve, learners did not feel it was personalized enough.

Regarding open-ended comments, they align with perceptions gathered in Lee & Song (2024) and Nazaretsky et al. (2024) related feedback generated with GenAI. Participants consider the feedback content useful, but noted that some parts seem repetitive, suggesting human oversight (Lin & Crosthwaite, 2024). Also, they complained about the difficulty in understanding the risk level and the associated recommendations. They did not understand that predictive models are based on learners' behavior in previous semesters, and that such predictions indicate that most learners with the same profile failed the course. However, this suggestion in some cases increased engagement and questioned the teacher, leading to a dialogue, one of the main advantages of formative feedback (Yang & Carless, 2013). This advocates that AI literacy must be a main pillar for current learners (Bewersdorff et al., 2025), and data-supported feedback for self- and self-regulation levels completes the full feedback loop (Clow, 2012; Yan et al., 2024).

This work has some limitations. On the one hand, self-selection bias arises from the constraints imposed by the Ethical Committee. Mostly, active, engaged, and motivated learners are the ones who consented to participate. Additionally, the design of both feedback groups is imbalanced due to the invitation process (i.e., an equal number of invitations but different numbers of learners accepted). Such limitation restricts generalizability and may affect external validity. However, this experiment piloted an integrated feedback approach in a large-enrollment first-year economics course, where providing feedback at scale had previously been difficult. The presented approach, which proposed reviewing feedback fragments by competence rather than complete message feedback, contributed to drastically reducing teachers' effort (Cavalcanti et al., 2021). Performing a full experiment with randomized equal group sizes would enrich the results obtained. On the other hand, results on the questionnaire designed for RQ4 may be impacted by mortality bias. Learners who have dropped out during the course stop attending classroom announcements and, therefore, do not provide their opinions at the end of the course.

This study demonstrated the feasibility of a feedback approach combined with risk identification. As future work, we are interested in exploring whether suggestive scaffolding feedback, supported by

GenAI, can enable the feedback loop, ensuring a dialogue with corrective comments until mastery is reached. Additionally, generalizability must be sought by exploring these approaches in other courses in different domains. Each domain has different activities (e.g., essays, programming, mathematics, among others) with distinct feedback needs. Also, teachers' expectations were not explored in this study due to the limited number of teachers, but their perceptions are crucial for assessing the approach's usefulness and effectiveness. Thus, obtaining opinions across different domains will provide insights into whether delivering high-quality feedback at scale is feasible in the era of GenAI tools.

Abbreviations

Generative Artificial Intelligence (GenAI).

Science, Technology, Engineering, and Mathematics (STEM).

<SYS_BLINDED>.

<BLINDED_UNIV>.

Continuous Assessment Activity (CAA).

Continuous Assessment (CA).

Final Test (FT).

Learning Management System (LMS).

Final Mark (FM).

General Data Protection Regulation (GDPR).

<MODEL_BLINDED>.

Large Language Model (LLM).

True Positive (TP).

True Negative (TN).

False Positive (FP).

False Negative (FN).

True Positive Rate (TPR).

True Negative Rate (TNR).

Accuracy (ACC)

F Score 1.5 (F1.5)

Mean (M)

Standard Deviation (SD)

Artificial Intelligence (AI).

Generative Artificial Intelligence (GenAI)

Early Warning System (EWS).

Students Information System (SIS)

Kolmogorov-Smirnov normality test (K-S)

Declarations

Availability of data and materials: Data from the <SYS_BLINDED> system and Canvas LMS include personal identifying information when merged with information from the signed consent form, so they cannot be disclosed to the general public.

Competing Interests: The authors declare that they have no competing interests.

Funding: <blinded>

Author's contribution: <blinded>

Acknowledgments: <blinded>

Ethics approval and consent to participate: The Research Ethics Committee requires learners to provide explicit consent to participate in any study. A consent form was required for all participants prior to starting, and Committee approved the research procedure.

Consent for publication: All authors consent to its publication. A consent form was signed by all participants involved in this research paper.

References

1. Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., Li, X., & Weegar, R. (2021). Explainable AI for Data-Driven Feedback and Intelligent Action Recommendations to Support Students Self-Regulation. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.723447

2. Akçapinar, G. (2015). How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers and Education*, *87*. https://doi.org/10.1016/j.compedu.2015.04.007
3. Almasre, M. (2024). Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course. *Education Sciences*, *14*(2). https://doi.org/10.3390/educsci14020148
4. Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. *ACM International Conference Proceeding Series*, 267–270. https://doi.org/10.1145/2330601.2330666
5. Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4
6. Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S. L. (2009). Measuring self-regulation in online and blended learning environments. *Internet and Higher Education*, *12*(1). https://doi.org/10.1016/j.iheduc.2008.10.005
7. Bewersdorff, A., Hornberger, M., Nerdel, C., & Schiff, D. S. (2025). AI advocates and cautious critics: How AI attitudes, AI interest, use of AI, and AI literacy build university students' AI self-efficacy. *Computers and Education: Artificial Intelligence*, *8*. https://doi.org/10.1016/j.caeai.2024.100340
8. Bodily, R., Ikahihifo, T. K., Mackley, B., & Graham, C. R. (2018). The design, development, and implementation of student-facing learning analytics dashboards. *Journal of Computing in Higher Education*, *30*(3). https://doi.org/10.1007/s12528-018-9186-0
9. Boschloo, R. D. (1970). Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, *24*(1). https://doi.org/10.1111/j.1467-9574.1970.tb00104.x
10. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2). https://doi.org/10.1191/1478088706qp063oa
11. Burgos, C., Campanario, M. L., Peña, D., de la, Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541–556. https://doi.org/10.1016/J.COMPELECENG.2017.03.005
12. Casey, K., & Azcona, D. (2017). Utilizing student activity patterns to predict performance. *International Journal of Educational Technology in Higher Education*, *14*(1). https://doi.org/10.1186/s41239-017-0044-3
13. Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. In *Computers and Education: Artificial Intelligence* (Vol. 2). https://doi.org/10.1016/j.caeai.2021.100027
14. Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, *7*(2). https://doi.org/10.18608/JLA.2020.72.1

15. Cho, M. H., & Shen, D. (2013). Self-regulation in online learning. *Distance Education*, *34*(3). https://doi.org/10.1080/01587919.2013.835770
16. Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2330601.2330636
17. Commission, E. (2021). Annex III: High-risk AI systems referred to in Article 6(2). In *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final)*. European Commission. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF
18. Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100299
19. Davis, D., Chen, G., Jivet, I., Hauff, C., Kizilcec, R. F., & Houben, G. J. (2017). Follow the successful crowd: Raising MOOC completion rates through social comparison at scale? *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3027385.3027411
20. Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers and Education*, *162*. https://doi.org/10.1016/j.compedu.2020.104094
21. Dhananjaya, G. M., Goudar, R. H., Kulkarni, A. A., Rathod, V. N., & Hukkeri, G. S. (2024). A Digital Recommendation System for Personalized Learning to Enhance Online Education: A Review. *IEEE Access*, *12*. https://doi.org/10.1109/ACCESS.2024.3369901
22. Elliot, A. J., & Fryer, J. W. (2008). The goal construct in psychology. *Handbook of Motivation Science*, *18*, 235–250.
23. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00425-2
24. Espasa, A., Guasch, T., Mayordomo, R. M., Martínez-Melo, M., & Carless, D. (2018). A Dialogic Feedback Index measuring key aspects of feedback processes in online learning environments. *Higher Education Research and Development*, *37*(3), 499–513. https://doi.org/10.1080/07294360.2018.1430125
25. Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. In *Review of Educational Research* (Vol. 83, Number 1). https://doi.org/10.3102/0034654312474350
26. Giamos, D., Doucet, O., & Léger, P. M. (2024). Continuous Performance Feedback: Investigating the Effects of Feedback Content and Feedback Sources on Performance, Motivation to Improve Performance and Task Engagement. *Journal of Organizational Behavior Management*, *44*(3). https://doi.org/10.1080/01608061.2023.2238029
27. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar,

A., Korenev, A., Hinsvark, A., & Ma, Z. (2024). *The Llama 3 Herd of Models*. https://arxiv.org/abs/2407.21783.

28. Guasch, T., Espasa, A., & Martinez-Melo, M. (2019). The art of questioning in online learning environments: the potentialities of feedback in writing. *Assessment and Evaluation in Higher Education*, *44*(1). https://doi.org/10.1080/02602938.2018.1479373
29. Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487
30. Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*. https://doi.org/10.1016/j.chb.2014.04.002
31. Huberth, M., Chen, P., Tritz, J., & McKay, T. A. (2015). Computer-tailored student support in introductory physics. *Plos One*, *10*(9). https://doi.org/10.1371/journal.pone.0137001
32. Instructure (2025). *Our Story. Canvas Learning Management System*. https://www.instructure.com/about
33. Jia, Q., Cui, J., Du, H., Rashid, P., Xi, R., Li, R., & Gehringer, E. (2024). LLM-generated Feedback in Real Classes and Beyond: Perspectives from Students and Instructors. *Proceedings of the International Conference on Educational Data Mining*. https://doi.org/10.5281/zenodo.12729974
34. Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024). On Assessing the Faithfulness of LLM-generated Feedback on Student Assignments. In B. PaaÃŸen & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 491–499). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729868
35. Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. (2021). ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. *Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021*.
36. Kaufmann, R., & Vallade, J. I. (2022). Exploring connections in the online learning environment: student perceptions of rapport, climate, and loneliness. *Interactive Learning Environments*, *30*(10). https://doi.org/10.1080/10494820.2020.1749670
37. Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education*, *19*(1). https://doi.org/10.1145/3231711
38. Kew, S. N., & Tasir, Z. (2022). Developing a Learning Analytics Intervention in E-learning to Enhance Students' Learning Performance: A Case Study. *Education and Information Technologies*, *27*(5). https://doi.org/10.1007/s10639-022-10904-0
39. Kickmeier-Rust, M. D., Hillemann, E. C., & Albert, D. (2014). Gamification and smart feedback: Experiences with a primary school level math app. *International Journal of Game-Based Learning*, *4*(3). https://doi.org/10.4018/ijgbl.2014070104
40. Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for Writing Evaluation: Examining the Accuracy and Reliability of AI-Generated Scores Compared to Human Raters. In *Exploring AI in Applied Linguistics*. https://doi.org/10.31274/isudp.2024.154.06

41. Kim, T. W. (2023). Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. In *Journal of Educational Evaluation for Health Professions* (Vol. 20). https://doi.org/10.3352/jeehp.2023.20.38
42. Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, *8*. https://doi.org/10.1016/j.caeai.2024.100349
43. Knowles, J. (2014). Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. *JEDM - Journal of Educational Data Mining*, *7*(3), 1–52. https://doi.org/10.5281/zenodo.3554725
44. Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: Multiple perspectives. *Journal of the Association for Information Systems*, *13*(6). https://doi.org/10.17705/1jais.00300
45. Latham, G. P., & Locke, E. A. (2007). New developments in and directions for goal-setting research. *European Psychologist*, *12*(4). https://doi.org/10.1027/1016-9040.12.4.290
46. Lee, S., & Song, K. S. (2024). Teachers' and students' perceptions of AI-generated concept explanations: Implications for integrating generative AI in computer science education. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100283
47. Li, C., Herbert, N., Yeom, S., & Montgomery, J. (2022). Retention Factors in STEM Education Identified Using Learning Analytics: A Systematic Review. In *Education Sciences* (Vol. 12, Number 11). https://doi.org/10.3390/educsci12110781
48. Lin, S., & Crosthwaite, P. (2024). The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback. *System*, *127*. https://doi.org/10.1016/j.system.2024.103529
49. Llauró, A., Fonseca, D., Villegas, E., Aláez, M., & Romero, S. (2021). Educational data mining application for improving the academic tutorial sessions, and the reduction of early dropout in undergraduate students. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3486011.3486449
50. Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9). https://doi.org/10.1037/0003-066X.57.9.705
51. Maier, U., & Klotz, C. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. In *Computers and Education: Artificial Intelligence* (Vol. 3). https://doi.org/10.1016/j.caeai.2022.100080
52. Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1). https://doi.org/10.1214/aoms/1177730491
53. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, M., H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*,

33(1), 107–124. https://doi.org/10.1111/exsy.12135

54. Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253). https://doi.org/10.1080/01621459.1951.10500769
55. Matcha, W., Uzir, N. A., Gasevic, D., & Pardo, A. (2020). A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. In *IEEE Transactions on Learning Technologies* (Vol. 13, Number 2). https://doi.org/10.1109/TLT.2019.2916802
56. Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Kloos, C. D. (2019). Prediction in MOOCs: A Review and Future Research Directions. *IEEE Transactions on Learning Technologies*, *12*(3), 384–401. https://doi.org/10.1109/TLT.2018.2856808
57. Mousavi, A., Schmidt, M., Squires, V., & Wilson, K. (2021). Assessing the Effectiveness of Student Advice Recommender Agent (SARA): the Case of Automated Personalized Feedback. *International Journal of Artificial Intelligence in Education*, *31*(3). https://doi.org/10.1007/s40593-020-00210-6
58. Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2020.1727529
59. Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017). MOOC dropout prediction: Lessons learned from making pipelines interpretable. *26th International World Wide Web Conference 2017, WWW 2017 Companion*. https://doi.org/10.1145/3041021.3054162
60. Najdi, L., & Er-Raha, B. (2016). A Novel Predictive Modeling System to Analyze Students at Risk of Academic Failure. *International Journal of Computer Applications*, *156*(6), 25–30. https://doi.org/10.5120/ijca2016912482
61. Naz, I., & Robertson, R. (2024). Exploring the Feasibility and Efficacy of ChatGPT3 for Personalized Feedback in Teaching. *Electronic Journal of E-Learning*, *22*(2). https://doi.org/10.34190/ejel.22.2.3345
62. Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or Human? Evaluating Student Feedback Perceptions in Higher Education. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *15159 LNCS*. https://doi.org/10.1007/978-3-031-72315-5_20
63. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*. https://doi.org/10.1016/j.caeai.2022.100066
64. Oates, B. J. (2006). Researching Information Systems and Computing. *Inorganic Chemistry* (Vol. 37). fcgi?artid=2836698&tool=pmcentrez&rendertype=abstract. http://www.pubmedcentral.nih.gov/articlerender. SAGE.
65. Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J., & Blanco, Ó. (2019). From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention

System. *IEEE Transactions on Learning Technologies*, *12*(2), 264–277. https://doi.org/10.1109/TLT.2019.2911608

66. Pardo, A., Bartimote, K., Buckingham Shum, S., Dawson, S., Gao, J., Gašević, D., Leichtweis, S., Liu, D., Martínez-Maldonado, R., Mirriahi, N., Moskal, A. C. M., Schulte, J., Siemens, G., & Vigentini, L. (2018). OnTask: Delivering Data-Informed, Personalized Learning Support Actions. *Journal of Learning Analytics*, *5*(3). https://doi.org/10.18608/jla.2018.53.15
67. Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of persona<sys_blinded > ed feedback</sys_blinded >. *British Journal of Educational Technology*, *50*(1). https://doi.org/10.1111/bjet.12592
68. Qu, K., & Wu, X. (2024). ChatGPT as a CALL tool in language education: A study of hedonic motivation adoption models in Eng<sys_blinded > h learning environments</sys_blinded >. *Education and Information Technologies*, *29*(15). https://doi.org/10.1007/s10639-024-12598-y
69. Qu, Y., Li, F., Li, L., Dou, X., & Wang, H. (2022). Can We Predict Student Performance Based on Tabular and Textual Data? *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2022.3198682
70. Rafique, A., Khan, M. S., Jamal, M. H., Tasadduq, M., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2021). Integrating Learning Analytics and Collaborative Learning for Improving Student's Academic Performance. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3135309
71. Ruwe, T., & Mayweg-Paus, E. (2024). Embracing LLM Feedback: the role of feedback providers and provider information for feedback effectiveness. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1461362
72. Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2). https://doi.org/10.1007/BF00117714
73. Sánchez-Gelabert, A., Valente, R., & Duart, J. M. (2020). Profiles of online students and the impact of their university experience. *International Review of Research in Open and Distributed Learning*, *21*(3). https://doi.org/10.19173/irrodl.v21i3.4784
74. Soumya, M., & Krishnamoorthy, S. (2022). Student performance prediction, risk analysis, and feedback based on context-bound cognitive skill scores. *Education and Information Technologies*, *27*(3). https://doi.org/10.1007/s10639-021-10738-2
75. Srilekshmi, M., Sindhumol, S., Chatterjee, S., & Bijlani, K. (2017). Learning Analytics to Identify Students At-risk in MOOCs. *Proceedings - IEEE 8th International Conference on Technology for Education, T4E 2016*, 194–199. https://doi.org/10.1109/T4E.2016.048
76. Vasquez, H., Fuentes, A. A., Kypuros, J. A., & Azarbayejani, M. (2015). Early identification of at-risk students in a lower-level engineering gatekeeper course. In *2015 IEEE Frontiers in Education Conference (FIE)* (Vol. 2015, pp. 1–9). IEEE. https://doi.org/10.1109/FIE.2015.7344361
77. Vega, H., Sanez, E., De La Cruz, P., Moquillaza, S., & Pretell, J. (2022). Intelligent System to Predict University Students Dropout. *International Journal of Online and Biomedical Engineering*, *18*(7). https://doi.org/10.3991/ijoe.v18i07.30195

78. Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers and Education*, *191*. https://doi.org/10.1016/j.compedu.2022.104644

79. Wang, H., & Lehman, J. D. (2021). Using achievement goal-based personalized motivational feedback to enhance online learning. *Educational Technology Research and Development*, *69*(2). https://doi.org/10.1007/s11423-021-09940-3

80. Wang, H., Tlili, A., Lehman, J. D., Lu, H., & Huang, R. (2021). Investigating feedback implemented by instructors to support online competency-based learning (CBL): a multiple case study. *International Journal of Educational Technology in Higher Education*, *18*(1). https://doi.org/10.1186/s41239-021-00241-6

81. Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). MOOC dropout prediction: How to measure accuracy? *L@S 2017 - Proceedings of the 4th (2017) ACM Conference on Learning at Scale*, 161–164. https://doi.org/10.1145/3051457.3053974

82. Winstone, N., & Carless, D. (2019). Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach. In *Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach*. https://doi.org/10.4324/9781351115940

83. Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes. In *Educational Psychologist* (Vol. 52, Number 1). https://doi.org/10.1080/00461520.2016.1207538

84. Wolff, A., Zdrahal, Z., Herrmannova, D., & Knoth, P. (2014). Predicting Student Performance from Combined Data Sources. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (Vol. 524, pp. 175–202). Springer International Pub her. https://doi.org/10.1007/978-3-319-02738-8_7

85. Wurster, K. G., Kivlighan, D. M., & Foley-Nicpon, M. (2021). Does person-group fit matter? A further examination of hope and belongingness in academic enhancement groups. *Journal of Counseling Psychology*, *68*(1). https://doi.org/10.1037/cou0000437

86. Xavier, M., & Meneses, J. (2022). Persistence and time challenges in an open online university: a case study of the experiences of first-year learners. *International Journal of Educational Technology in Higher Education*, *19*(1), 31. https://doi.org/10.1186/s41239-022-00338-6

87. Xavier, M., Meneses, J., & Fiuza, P. J. (2026). Dropout, stopout, and time challenges in open online higher education: A qualitative study of the first-year student experience. *Open Learning*, *41*(1). https://doi.org/10.1080/02680513.2022.2160236

88. Xie, X., & Li, X. (2018). Research on Personalized Exercises and Teaching Feedback Based on Big Data. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3232116.3232143

89. Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, *58*. https://doi.org/10.1016/j.chb.2015.12.007

90. Yan, L., Martinez-Maldonado, R., & Gasevic, D. (2024). Generative Artificial Intelligence in Learning Analytics: Contextua ing Opportunities and Challenges through the Learning Analytics Cycle. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3636555.3636856
91. Yang, M., & Carless, D. (2013). The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education*, *18*(3). https://doi.org/10.1080/13562517.2012.719154
92. You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education*, *29*, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003
93. Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, *25*(1). https://doi.org/10.1207/s15326985ep2501_2

Figures

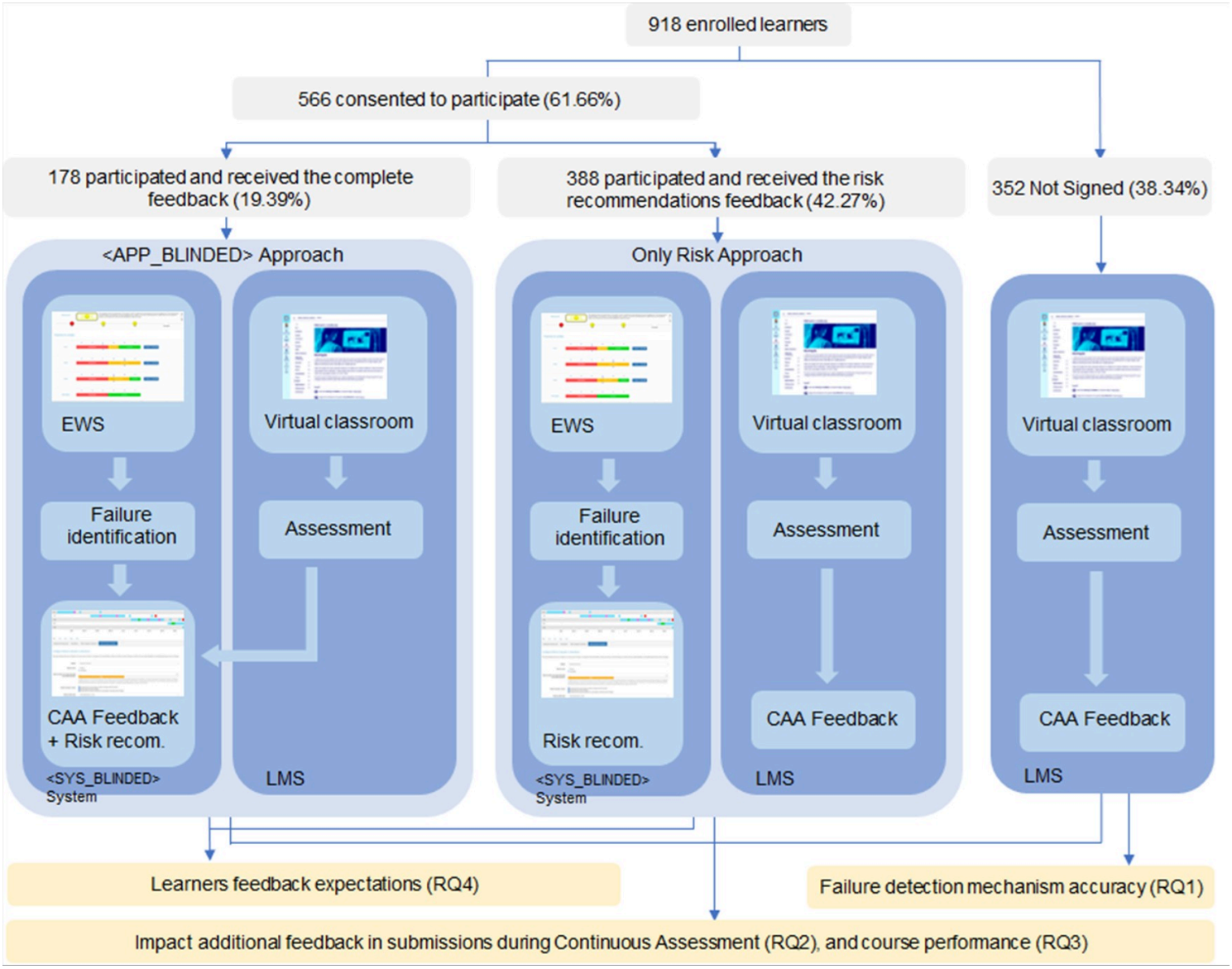


Figure 1

Research design and participants.

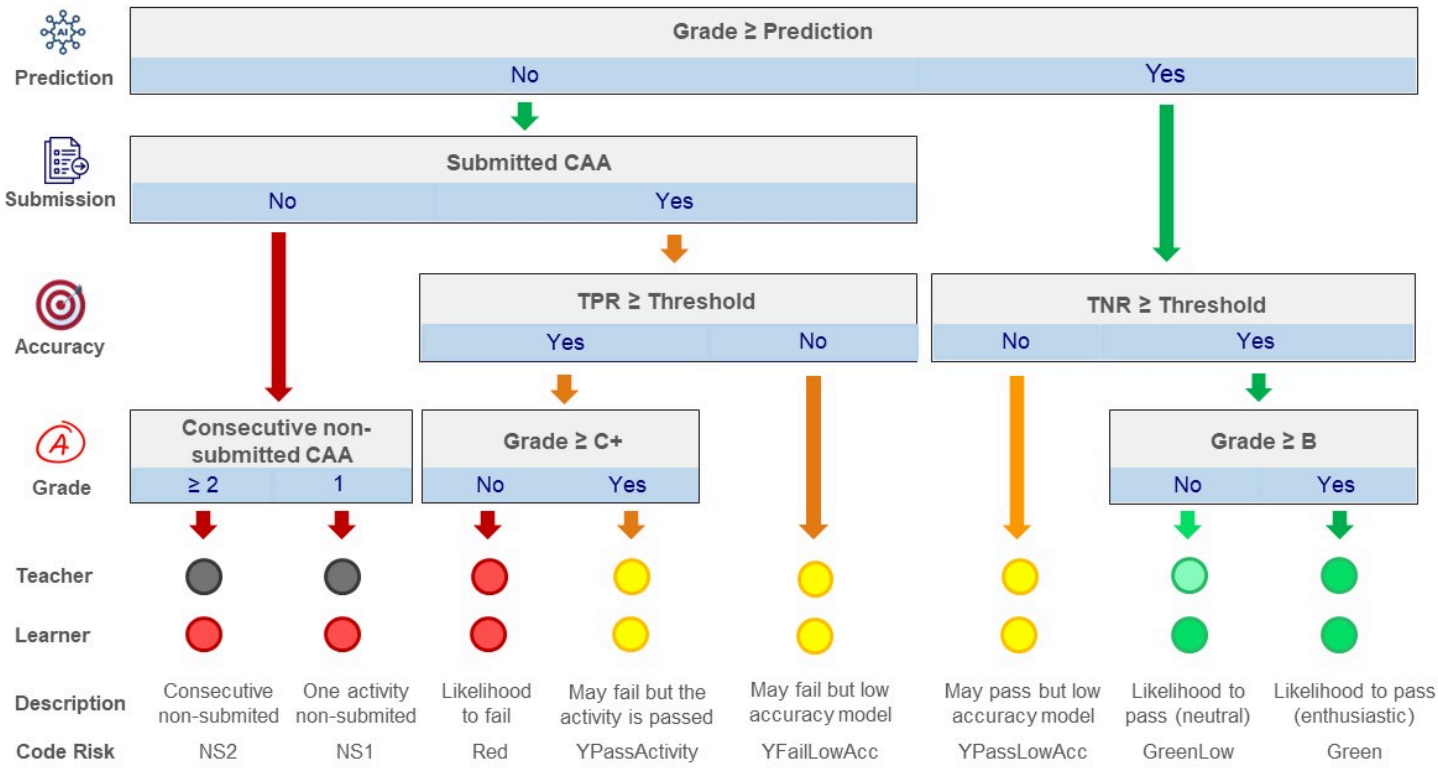


Figure 2

Failure at-risk color distribution for teachers and learners.

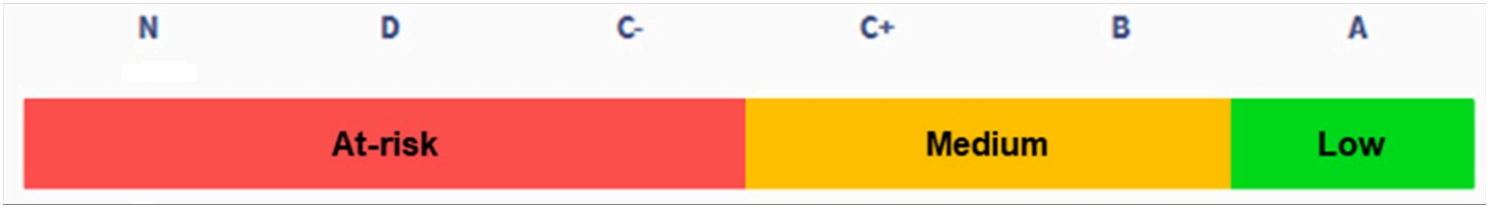


Figure 3

Example of failure at-risk classification distribution for a CAA.

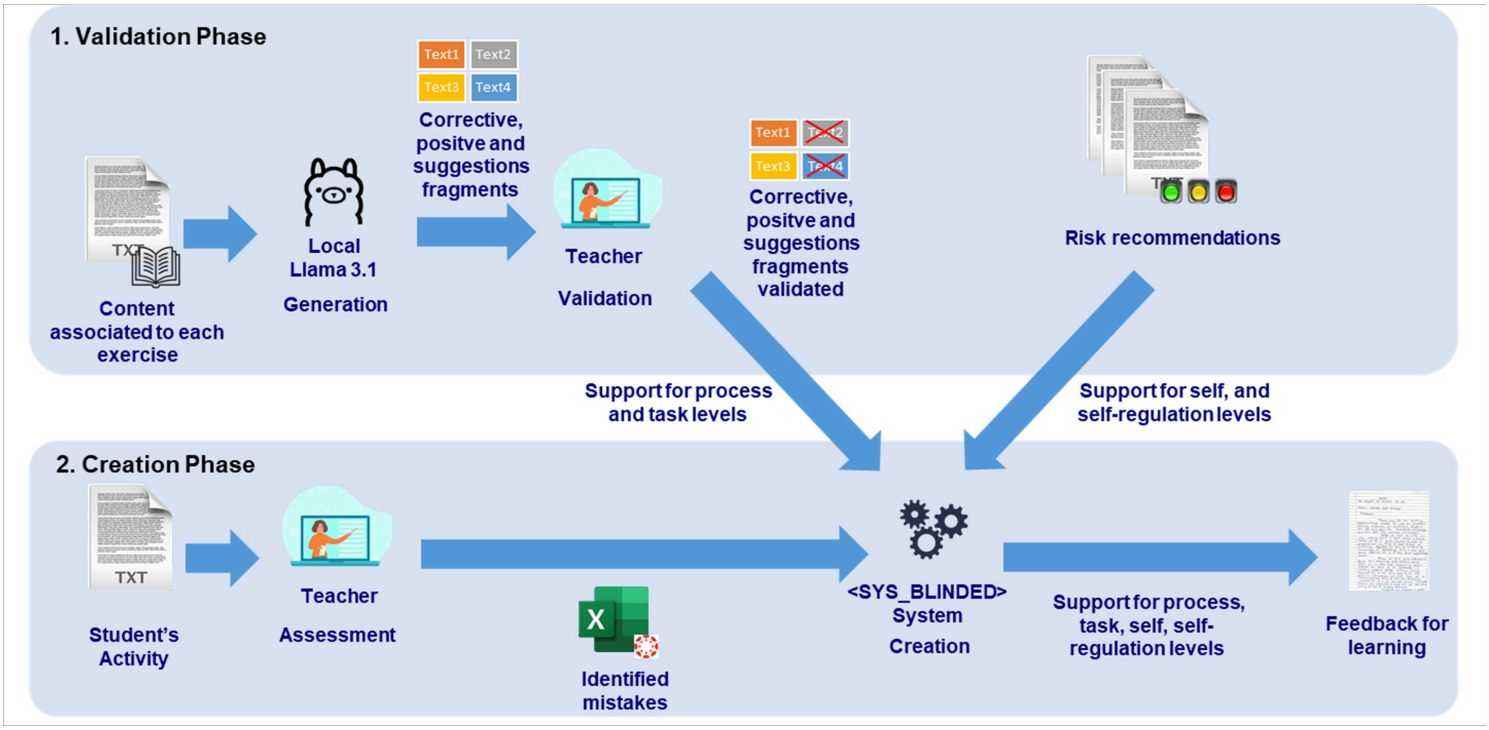


Figure 4

Approach to create <APP_BLINDED> feedback.

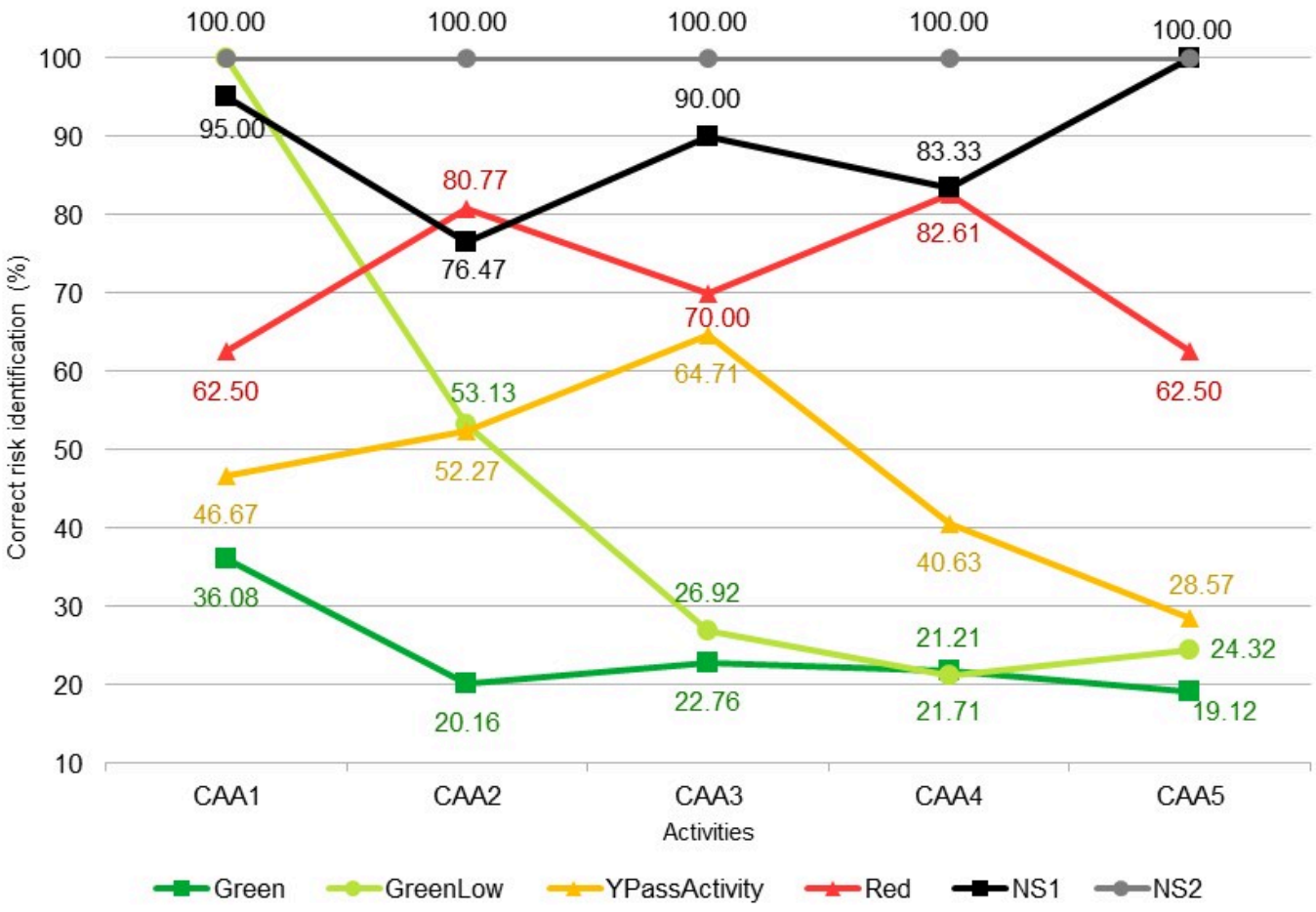
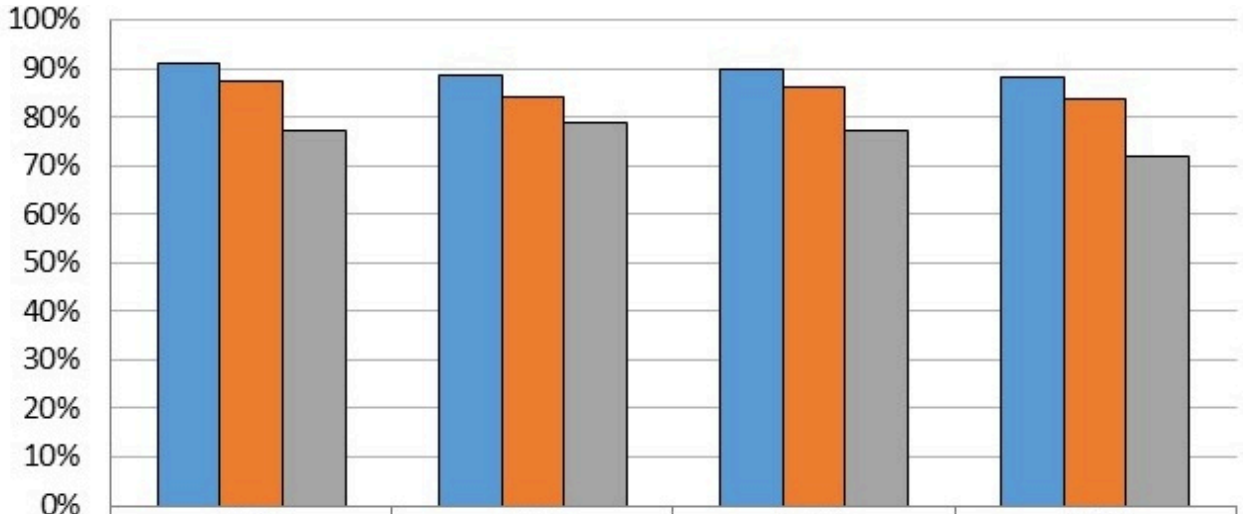


Figure 5

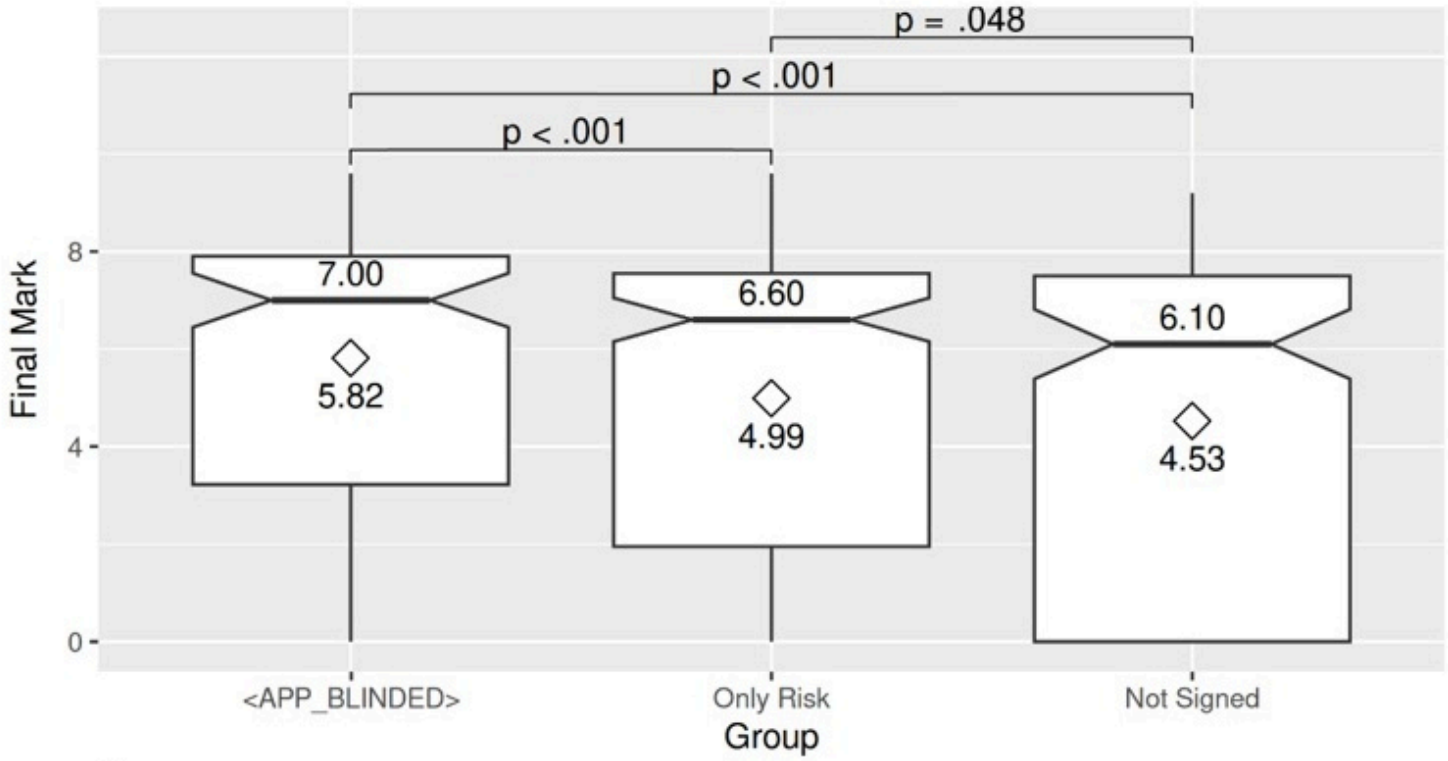
Success rate of at-risk failure identification distribution by risk level on not signed group.



	CAA2	CAA3	CAA4	CAA5
<APP_BLINDED>	91.01%	88.76%	89.89%	88.20%
Only Risk	87.47%	84.14%	86.19%	83.63%
Not Signed	77.17%	78.62%	77.17%	71.74%

Figure 6

Submission rate during Continuous Assessment.



Notes:

p-value of the Mann-Whitney test (above figure), median, and mean (diamond sign)

Figure 7

Notched box-and-whisker plot of the final mark distribution