An Empirical Study of Assessing and Mitigating Trustworthiness of Educational AI

Evangelia Anagnostopoulou[1,*], Nikolaos Antonios Grammatikos[1,2], Dimitris Apostolou[1,2] and Gregoris Mentzas[1]

[1]*Information Management Unit, Institute of Communication and Computer Systems, National Technical University of Athens, Athens, Greece*

[2]*Department of Informatics, University of Piraeus, Piraeus, Greece*

Abstract

This paper presents the trustworthiness assessment of an AI-driven platform for inquiry-based STEM education, deployed in real secondary school classrooms. The assessment was conducted using a structured trustworthiness risk assessment methodology, covering cartography, threat analysis, consequence assessment, vulnerability analysis, risk analysis, and risk management, and spanning all AI lifecycle phases—design, development, and deployment. Following this methodology, a set of mitigation controls was identified and implemented across the platform's three core AI models: the Learning Companion, the Student Performance Classifier, and the Learning Recommendations system. After the implementation of these controls, the models achieved the following results: the Learning Companion achieved 100% intent classification accuracy across all intents; the Student Performance Classifier reached 96.7% accuracy on unseen data; and the Learning Recommendations system attained an 80% teacher acceptance rate in its initial deployment. The assessment will be repeated in subsequent pilot phases. These results illustrate the application of the trustworthiness assessment methodology in a real-world educational context, providing a replicable methodology for EU AI Act-compliant AI deployment in high-risk educational settings.

Keywords

Trustworthy AI, AI Trustworthiness Assessment, STEM Education, EU AI Act, Inquiry-Based Learning, Risk Assessment

1. Introduction

The rapid integration of Artificial Intelligence (AI) into educational settings has created both significant opportunities and critical responsibilities. AI-powered platforms can enable personalized learning on a large scale, automate formative assessment, and produce suggestions that can customize instruction for students. However, these features may create significant challenges related to trustworthiness. For example, wrong classifications of student performance can lead to wrong learning paths, recommendation engines that carry bias can put certain student groups on a disadvantageous track and conversation agents that are not able to reliably infer student intent can reduce pedagogical coherence. In educational settings, where AI's output shapes the educational journey of students, trustworthiness becomes a fundamental prerequisite for the use of AI responsibly.

While a variety of AI-supported educational tools have emerged, there is a lack of empirical research regarding the relationship between AI risk management and the resulting measurable outcomes of trustworthiness in the real world. While many previous research efforts focus on a specific dimension of trustworthiness, such as robustness [1], bias [2] and explainability [3], few studies report on the entire process of identifying a threat, implementing a mitigation strategy, and quantifying the validity of the entire process after deployment, using multiple AI components of a single system. This gap is particularly acute for multi-component educational AI platforms, where different models present distinct risk profiles and require differentiated mitigation strategies.

TRUST-AI 2026: Second European Workshop on Trustworthy AI, co-located with IJCAI 2026, August 15–17, 2026, Bremen, Germany

*Corresponding author.

eanagn@mail.ntua.gr (E. Anagnostopoulou); ngrammatikos@unipi.gr (N. A. Grammatikos); dapost@unipi.gr (D. Apostolou); gmentzas@mail.ntua.gr (G. Mentzas)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This paper addresses this gap by reporting on the trustworthiness assessment of an AI-driven platform for inquiry-based STEM education, deployed in real secondary school classrooms. The assessment spans all AI lifecycle phases—design, development, and deployment. This paper presents the trustworthiness assessment results, and the assessment will be repeated in subsequent pilot phases to track trustworthiness evolution. Building on earlier work examining the trustworthiness roadmap for AI in education [4], the research objectives are: (i) to apply a comprehensive AI trustworthiness assessment methodology across three AI models—the Learning Companion, the Student performance classifier, and the Learning Recommendations engine—operating within a single educational platform; (ii) to document the mitigation measures identified, implemented, and deferred across all seven trustworthiness dimensions; and (iii) to report illustrative quantitative indicators alongside the implemented mitigation measures for each model. The paper is organized as follows. Section 2 reviews the literature on trustworthy AI in education and the regulatory landscape shaping its deployment. Section 3 describes the AI platform and its three constituent models. Section 4 presents the trustworthiness assessment methodology applied. Section 5 reports the risk assessment findings across all assets. Section 6 details the implemented mitigations per model along with the quantitative validity results. Section 7 discusses the implications of the findings, and Section 8 concludes with directions for future work.

2. Background and Related Work

The use of AI in education has received increasing interest due to the increasing number of AI-based adaptive learning systems, intelligent tutoring systems, and automatic assessment systems [5]. AI systems involved in decisions regarding students' performance, learning paths, and teaching recommendations must meet high standards of trustworthiness, including transparency, fairness, accountability, and validity. These requirements are reported in EU frameworks including the EU High-Level Expert Group (HLEG) guidelines on trustworthy AI [6] and the EU AI Act [7] which defines AI systems with educational consequences as high-risk, including requirements for data governance, human oversight, accuracy, robustness and transparency. However, making these normative standards operational at the level of deployed AI systems is still a challenge.

Current tools and methods focus on individual dimensions of trustworthiness, such as IBM's AI Fairness 360 for bias and fairness, and Microsoft's Responsible AI toolkit for explainability and accountability [8]. Despite this, the integration of frameworks covering the full lifecycle of an AI system across multiple trustworthiness dimensions—validity, explainability, fairness, safety, security, privacy, and accountability—is still limited, especially in educational AI platforms [9]. In multi-component platforms for education, such as AI-driven learning platforms that integrate conversational AI, classifiers, and generative recommendation engines, the risk profiles vary, making it difficult to mitigate them with a one-size-fits-all approach. Studies on trustworthy AI roadmaps for education have started to outline ethical and practical considerations for embedded assessments [4] but empirical end-to-end deployments of educational platforms with quantitative evidence have yet to emerge.

3. AI Platform for Inquiry-Based STEM Education

The AI platform under investigation [10, 11] supports inquiry-based STEM learning by combining three AI components that operate in concert. The platform utilises three AI models, Learning Companion, Student Performance Classifier, and Learning Recommendations and two datasets as depicted in Figure 1. Each addresses a distinct pedagogical need and presents distinct trustworthiness challenges.

The Learning Companion is a conversational agent built on Rasa 3.6.6, an open-source NLU and dialogue management framework. It interprets student inputs during inquiry-based scenarios, classifies them into one of eleven educational intents (e.g., "ask_me", "provide_answer", "request_hint", "confirm_read"), and routes the interaction accordingly. A custom action ("action_handle_answer") tracks cumulative student scores in real time and assigns students to performance tiers (High, Moderate, Low),

adapting the learning path dynamically. Teachers author the inquiry-based scenarios, which form the basis for all training data, ensuring expert control over content quality.

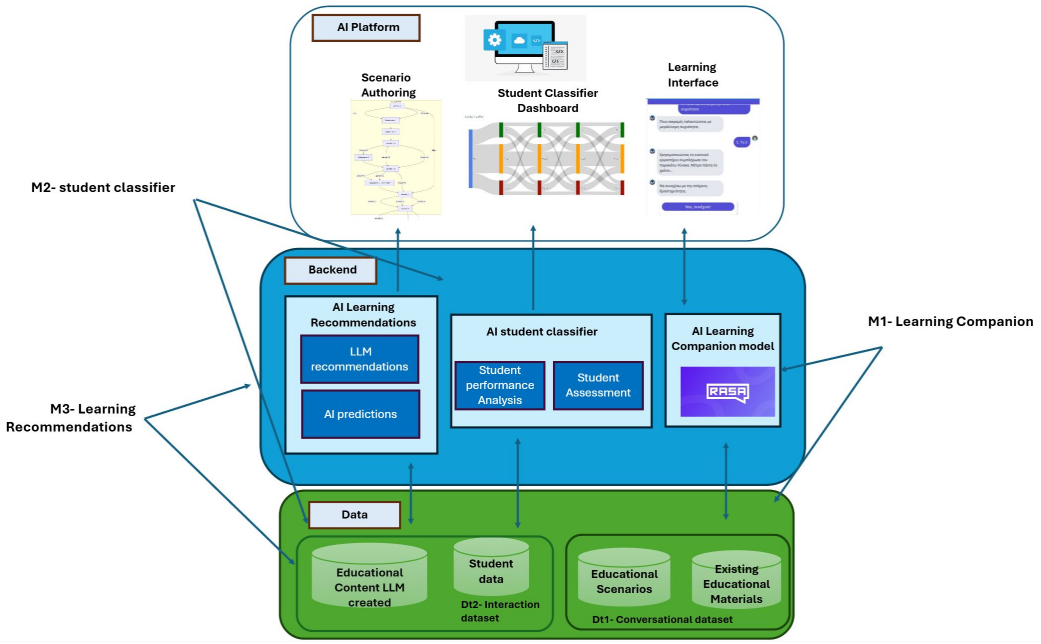

Figure 1: AI platform for inquiry-based STEM education

Student Performance Classifier has been trained on data from 759 students who completed the Pendulum inquiry-based activity. For each student, 29 features were extracted across four activity phases, including normalized scores, penalized scores, response times, and engagement indicators. The labeling scheme extends the PISA achievement framework (Low, Moderate, High) with engagement-based penalties: students with two or more disengaged phases are downgraded by one category, and those with three or more by two categories.

Learning Recommendations help teachers improve their inquiry-based scenarios by identifying activities where students show struggle, disengagement, or misconception. The system identifies activities that need improvement based on patterns in student answer and timing data. Teachers review each recommendation and can accept, modify, or reject it.

4. Trustworthiness Assessment Methodology

The trustworthiness assessment of the platform builds on earlier work that established a trustworthiness roadmap for AI in education [4] and on the platform's prior development and evaluation [10, 11]. The trustworthiness assessment of the AI platform was conducted using the AI-TAF framework [12], which provides a structured assessment process applicable to any AI system by combining structured risk identification, consequence modeling, and a mitigation workflow that guides risk assessors through the full assessment cycle and generates auditable mitigation registers. The framework was applied into all three AI assets of the educational AI platform to assess its trustworthiness across all AI lifecycle phases: design, development, and deployment. The framework comprises six sequential phases, each producing structured outputs that feed into the next phase, as depicted in Figure 2. Our work contributes to the first complete, empirical trustworthiness assessment of a multi-component educational AI system, covering all AI lifecycle phases and reporting quantitative validity evidence alongside the implemented mitigation measures.

This study employed a co-design risk assessment methodology to evaluate the structural and operational vulnerability of the AI platform across multiple stakeholder domains. By combining structured questionnaire-based elicitation from active classroom teachers with architectural asset characterization

from system developers, researchers cross-examined human-centric consequence severity against technical defensive controls. Teachers provided structured input through a questionnaire-based elicitation process, contributing their domain knowledge and pedagogical experience to the consequence assessment phase—enabling the risk assessors to evaluate the real-world impact of potential threats from the perspective of those directly affected by the system's outputs in the classroom. The AI developers, responsible for the design and development of the platform, provided the technical input necessary to characterize each AI asset, identify applicable threats, and assess existing controls during the vulnerability analysis phase. This collaborative, multi-stakeholder approach ensured that the assessment integrated both the end-user perspective on consequence severity and the technical perspective on system architecture and defensive controls, producing a more complete and contextually grounded risk picture.

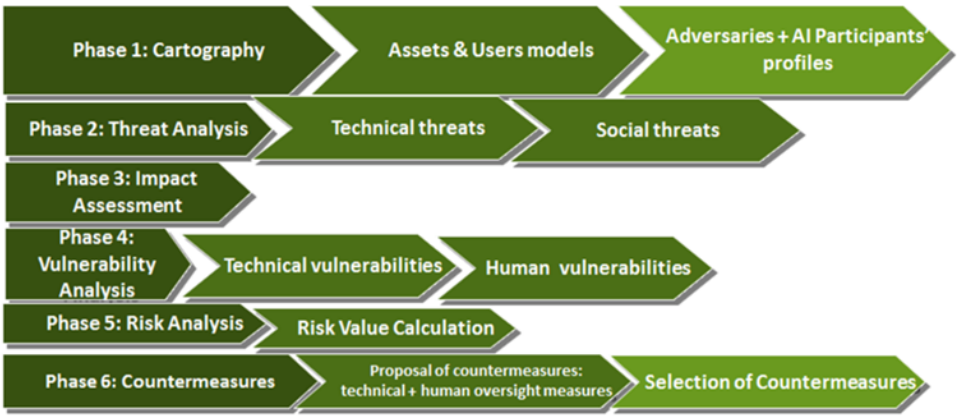

Figure 2: AI-TAF framework

During the cartography phase, the criticality level of the AI system was calculated as established by Article 6 of the EU AI Act, enabling the risk assessors to determine whether each asset falls within the scope of high-risk AI system provisions and to calibrate the rigor of subsequent assessment phases accordingly. The criticality level of the Educational AI platform was assessed as high risk. Additionally, the AI trustworthiness maturity level of all participant groups involved with the platform was assessed. The maturity level of the teachers, who are the primary end-users of the platform, was assessed as very high by the risk assessors.

During the threat analysis, the risk assessors identify applicable threats from the framework's threat taxonomy for each asset, scoring their likelihood of occurrence. The identified threats include Failure or Malfunction of ML Application, Model Drift Leading to Inaccurate Assessments, Human Error, Evasion, Bias in Personalized Learning Recommendations, and Propagation of Misinformation.

Consequence Assessment evaluates the severity of impact should each identified threat materialize, across the seven trustworthiness dimensions of the AI-TAF framework: Valid & Reliable, Explainable & Interpretable, Safe, Secure & Resilient, Fair, Privacy Enhanced, and Accountable & Transparent [12]. The assessment was conducted through a structured questionnaire completed by the team of risk assessors with the support of the teachers who are the primary end-users of the platform. For each trustworthiness dimension, teachers responded to a targeted question regarding the business consequence level, rated on a five-point scale from Negligible to Very High, should that dimension be compromised. Validity and reliability were rated highly, given the serious consequences of student misclassification and its potential impact on learning outcomes and student welfare. Additionally, explainability was rated high, as teachers rely on interpretable AI outputs to make informed pedagogical decisions. An opaque system would undermine their ability to critically evaluate and act on recommendations.

Vulnerability Assessment is used to assess existing controls and their effectiveness to generate an adjusted risk level that reflects the current state of the system's defenses. Risk Analysis combines likelihood, consequence, and vulnerability scores to produce an overall risk score per threat-asset pair.

During the risk assessment phase, the integrated risk engine calculated composite risk scores for each AI asset of the Educational AI platform, synthesizing the outputs of the preceding four phases into an overall risk profile. Finally, Risk Management produces the mitigation register, populated with control recommendations drawn from the framework's control library, which the risk assessor then reviews, assigning each control a status: Treated, Postponed, or Ignored. The output of this stage is a list of selected controls, a list of implemented controls and controls that will be implemented, and preliminary testing and evaluation reports of controls that are implemented. The full assessment cycle is designed to be repeated in subsequent pilot phases, enabling continuous monitoring of trustworthiness evolution as the system matures and the student dataset grows. The results of the trustworthiness assessment of the educational platform are presented in the following sections.

5. Risk Assessment Results

The trustworthiness assessment of the AI platform was conducted across all three of the AI assets (Learning Companion, Student performance classifier and Learning Recommendations) for all the AI lifecycle phases (Design, development, deployment). The assessment was intentionally carried out across the complete lifecycle to ensure comprehensive coverage of trustworthiness risks and vulnerabilities from the earliest stages of system conception through to live operational use. The risk assessment identified a set of Very High and High-risk findings across all three models, concentrated primarily in the Valid & Reliable, Safe, and Explainable & Interpretable dimensions. Table 1 summarizes the highest risk findings.

Table 1
Summary of highest-risk findings from the trustworthiness assessment across all three AI assets.

Threat	Dimension	Asset	Risk Level
Failure / Malfunction	Valid & Reliable	Learning Companion, Student performance classifier and Learning Recommendations	Very High
Model Drift Leading to Inaccurate Assessments	Valid & Reliable	Learning Companion, Student performance classifier and Learning Recommendations	Very High
Human Error	Explainable, Accountable, Safe	Learning Companion, Student performance classifier and Learning Recommendations	Very High
Evasion	Secure & Resilient, Safe	Student performance classifier, Learning Recommendations	Very High
Bias in Personalized Learning Recommendations	Fair	Learning Recommendations	Very High
Propagation of Misinformation	Explainable & Interpretable	Learning Recommendations	High

The Validity dimension presents high risk of Failure or Malfunction and Model Drift. The Explainable and Interpretable dimension has an extremely high Human Error risk and a high Propagation of Misinformation risk which is an indication of the lack of uncertainty quantification mechanisms at present. The Evasion threat is evaluated as very high in the Safe and Secure and Resilient dimensions and with Human Error and Bias in Personalized Learning Recommendations which are both rated as very high. The risk of privacy is also equally low in all sub-assets, since there are no stored sensitive personal data and all the model outputs are regulated by restricted API access and internal hosting. Accountable and Transparent dimension documents have a very high risk of Human Error, which supports the main idea of teacher review quality as an important dependency on system control.

6. Risk Mitigations

Following the risk assessment, a structured mitigation register was compiled for each asset. The following subsections summarize the mitigations implemented for each model.

6.1. Learning Companion

All core mitigations for the Learning Companion have been treated and presented in Table 2.

Table 2
Implemented mitigations for the Learning Companion

Dimension	Threat	Implemented Mitigation
Valid & Reliable	Model Drift; Failure / Malfunction	Continuous monitoring of accuracy, precision, recall, and F1-score; model evaluated on held-out test set of 85 examples
Safe	Failure / Malfunction	Formal risk analysis covering model failure scenarios; performance indicators defined and monitored
Secure & Resilient	Poisoning; Data Quality	All training data created by teachers—expert-controlled, closed-environment source resistant to poisoning
Explainable & Interpretable	Human Error; Misinformation	Teacher authorship of all inquiry scenarios ensures content reviewed for accuracy and pedagogical appropriateness
Accountable & Transparent	Human Error; Lack of Governance	ML components registered with defined ownership, versioning, and access control; risk analysis covering data disclosure scenarios
Privacy Enhanced	Data Disclosure; GDPR	GDPR compliance confirmed; differential privacy principles validated; model outputs restricted to prevent attacks
Fairness	Failure / Malfunction	All scenarios reviewed by teachers for fairness and pedagogical appropriateness before deployment

Continuous monitoring of accuracy, precision, recall, and F1-score was performed, ensuring on-going evidence of model performance and mitigating the threat of model drift and model malfunction for the Valid & Reliable dimension. To implement mitigation control related to continuously monitoring model performance, the Learning Companion was evaluated on pilot data. The Rasa NLU component was evaluated on a held-out test set of 85 examples covering all 11 domain intents. The model achieved 100% precision, recall, and F1-score across all intents, with no inter-class confusion observed. In the context of educational AI, this level of accuracy is particularly significant: a misclassified intent would result in an incorrect pedagogical response—for example, providing a hint when the student intended to submit an answer. The perfect classification results are consistent with the continuous performance monitoring. Pedagogically consistent training examples produce a model that operates with the reliability required for live classroom deployment.

6.2. Student Performance Classifier

The student performance classifier model received a comprehensive set of treated mitigations aligned with its very high risks in the Valid & Reliable dimension. Table 3 provides an overview of the mitigations

Table 3
Implemented mitigations for the student performance classifier

Dimension	Threat	Implemented Mitigation
Valid & Reliable	Model Drift; Failure / Malfunction	Continuous monitoring of accuracy, precision, recall, and F1-score
Safe	Failure / Malfunction	Formal risk analysis of ML application conducted
Secure & Resilient	Poisoning; Data Quality	Noisy and incomplete interaction logs filtered at ingestion; reliable data sources verified; awareness strategy updated
Accountable & Transparent	Human Error; Lack of Governance	Defined ownership, access control, and versioning in place; performance indicators documented and reported
Fairness	Bias in Recommendations; Evasion	Fairness mitigations will be implemented during the Replication Phase pending larger dataset

integrated into the student performance classifier.

Continuous monitoring of accuracy, precision, recall, and F1-score was performed, ensuring on-going evidence of model performance and mitigating the threat of model drift and model malfunction for the Valid & Reliable dimension. Concerning fairness, an initial assessment confirmed the absence of systematic bias in the model's outputs across student performance categories. This assessment was conducted by inspecting the distribution of predicted performance categories (Low, Moderate, High) across the student cohort and verifying the absence of systematic over-representation of any group. Formal fairness metrics were not computed at this stage, as the dataset does not include demographic attributes required for group-level analysis. Three fairness-related mitigations—integration of fairness criteria into CI/CD pipelines, automated fairness testing, and systematic bias regression testing during system updates—were classified as Postponed, to be implemented in the replication phase once a sufficiently large dataset is available to support statistically robust bias analysis. These postponed controls represent the primary residual risk for this model and are to be treated as priority actions in subsequent assessment cycles. The student performance classifier was trained in 80% of the engagement-adjusted dataset (607 students) and evaluated on a stratified hold-out test set of 152 students. Five-fold stratified cross-validation was applied on the training set to ensure robust estimation.

The model correctly categorized 96.7% of previously unseen students, with consistently high F1 scores across all three performance categories. Notably, the High category achieved perfect precision (1.00), meaning no student was incorrectly classified as a high performer. The slight recall deficit for High (0.93) reflects a conservative bias toward the Moderate category for borderline cases, which is preferable in an educational setting where over-classification of struggling students as high achievers carries greater risk than the reverse. The most informative features for classification were mean_penalized_score and total_score, confirming that the engagement-based penalty mechanism contributes meaningfully to the model's discriminative ability beyond raw correctness measures.

6.3. Learning Recommendations

For the Learning Recommendations, the treated mitigations are presented in Table 4.

Table 4
Implemented mitigations for the AI Learning Recommendations

Dimension	Threat	Implemented Mitigation
Valid & Reliable	Model Drift; Failure / Malfunction	Performance monitored via Teacher Acceptance Rate
Safe	Failure / Malfunction	Risk analysis conducted across all pipeline components; teacher review mandatory before any scenario changes is applied to students
Secure & Resilient	Poisoning; Data Quality	Learning Material Dataset created and reviewed by teachers; Student Interaction Dataset sourced from the controlled educational setting; all stakeholders informed of ML-specific risks
Explainable & Interpretable	Human Error; Misinformation	Mandatory teacher review of all AI-generated recommendations; RAG retrieval grounded in teacher-authored PDF materials to constrain LLM outputs
Accountable & Transparent	Human Error; Lack of Governance	ML components integrated into asset management with defined ownership, access control, and versioning; data provenance tracked for Student Interaction Dataset
Privacy Enhanced	Data Disclosure; GDPR	Local model hosting; no student data transmitted to external services; GDPR compliance confirmed; no personal data stored in the recommendation pipeline; privacy-by-design principles applied to system architecture
Fairness	Bias in Recommendations	Fairness evaluated by tracking acceptance rates; teachers perform ongoing reviews to detect and correct potential biases

Mitigation measures focused on anchoring RAG outputs exclusively within expert-validated, teacher-authored content deployed on secure, locally hosted infrastructure. Methodological integrity and data

provenance are maintained through localized, GDPR-compliant architectures that eliminate external data transmission risks. The model formalizes human-in-the-loop oversight by mandating expert pre-screening of all machine learning outputs, leveraging teacher acceptance rates as a core metric to systematically identify, audit, and mitigate algorithmic bias. Despite these controls, five fairness-related mitigations, including automated bias detection and mitigation, regular fairness audits, uncertainty quantification and drift detection, automated fairness testing within the development lifecycle, and the embedding of fairness criteria into CI/CD pipelines, are planned to be addressed during the replication of trustworthiness assessment. Recommendation quality was assessed using the Teacher Acceptance Rate, defined as the proportion of AI-generated proposals accepted by teachers without requiring substantial revision. In the initial pilot deployment, the system achieved an acceptance rate of 80%, representing an early but encouraging baseline.

7. Discussion

The paper presents the trustworthiness assessment of an AI-driven platform using the AI-TAF methodology. The six-phase assessment process identified specific risks, and the implemented mitigations directly addressed these risks across all three AI models. For the Learning Companion, the proposed mitigation actions were implemented, with the model reaching 100% intent classification accuracy across all intents. For the Student Assessment model, the model achieved 96.7% accuracy on unseen data following the implementation of the proposed mitigations. For the Learning Recommendations engine, a human-in-the-loop teacher review mechanism was implemented, reflected in an 80% teacher acceptance rate in its initial deployment.

The AI-TAF methodology also highlights the importance of distinguishing between mitigations that can be implemented immediately and those that require additional data or tooling. Fairness-related mitigations (automated bias testing, fairness CI/CD integration) were consistently postponed across all three models, due to the current absence of demographic diversity data. In the replication phase, data will be collected from students across multiple countries, enabling proper evaluation and implementation of these mitigations. This is a transparent acknowledgment of residual risk, consistent with the EU AI Act's iterative risk management philosophy.

Sustainability in resource-constrained school environments is also a consideration. While the platform's design assumes active teacher participation in authoring inquiry-based scenarios, the platform already maintains a library of scenarios that have been authored and reviewed by participating teachers. Schools with limited staff time or lower digital expertise can adopt these scenarios directly, substantially lowering the resource barrier to deployment. Future work will assess the implications of trustworthiness of this reuse pathway.

Another limitation concerns the scope of the risk assessment: the three AI models were assessed independently, each treated as a self-contained asset. While this approach enables focused, dimension-specific risk analysis for each model, it does not capture the cascading threat dynamics that may emerge from their sequential operation as an integrated pipeline. In the deployed platform, the output of one model constitutes the input context for the next: student interaction data processed by the Learning Companion informs the student performance classifier, whose performance categorizations in turn condition the flagging logic and recommendation policy of the Learning Recommendations engine. A failure or drift in an upstream model can therefore propagate and amplify through downstream components in ways that an asset-isolated assessment cannot detect. For example, systematic misclassification by the student performance classifier model could cause the recommendations engine to consistently generate inappropriate activities for misclassified students, compounding the educational harm beyond what the model's individual risk score would suggest. This inter-model dependency, and the cascading threats it introduces, will be explicitly addressed in the replication assessment phase, where a system-level trustworthiness assessment will treat the three models as an integrated socio-technical pipeline and evaluate cross-component threat propagation pathways accordingly.

Moreover, additional limitations concern the generalizability and interpretability of the quantitative

validity results reported in Section 6. The Student Performance Classifier was trained and evaluated exclusively on data from the Pendulum inquiry-based scenario, involving 759 students. While the model demonstrates strong performance on held-out data from the same activity (96.7% accuracy), it remains an open question whether the learned feature importance rankings and classification boundaries will transfer to inquiry scenarios with different pedagogical structures, content domains, or student interaction patterns. Cross-activity validation should be treated as a priority in the replication phase, once data from additional inquiry scenarios become available. The Pendulum scenario was selected for its largest number of completed implementations. While the validity results are specific to this scenario, the AI-TAF methodology will be applied to additional inquiry-based scenarios in the replication phase for cross-activity validation.

8. Conclusions

This paper presented the trustworthiness assessment of an AI platform encompassing three AI models, a Learning Companion, a student performance classifier, and a Learning Recommendations engine. The application of the AI-TAF framework identified significant risks in various trustworthiness dimensions, and the implemented mitigations are illustrated.

The key contributions of this work are: (1) a complete, real-world application of the AI-TAF structured trustworthiness assessment methodology to educational AI; (2) illustration of how the framework's mitigation register connects to performance indicators; and (3) a replicable assessment process for EU AI Act compliance in high-risk educational deployments.

Future work will focus on the replication phase of the pilot, in which the AI-TAF assessment cycle will be repeated to track trustworthiness evolution across all lifecycle phases. The transition to a socio-technical system assessment will incorporate broader organizational and societal considerations. The postponed fairness-related mitigations will be addressed.

Acknowledgments

The work presented here is funded by the European Union's Horizon FAITH project (Fostering Artificial Intelligence Trust for Humans towards the optimization of trustworthiness through large-scale pilots in critical domains), Grant agreement No 101135932. The work presented here reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Declaration on Generative AI

During the preparation of this work, the author(s) used a Generative AI language assistant to check grammar and improve language. After using this tool, the author(s) reviewed the content and take(s) full responsibility for the publication's content.

References

[1] E. Ferrara, Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, Sci 6 (2024) 3. doi:10.3390/sci6010003.

[2] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, H. V. Poor, Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey, IEEE Communications Surveys & Tutorials (2023). doi:10.1109/COMST.2023.3317573.

[3] M. Mersha, K. Lam, J. Wood, A. K. AlShami, J. Kalita, Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction, Neurocomputing 599 (2024) 128111.

[4] E. Anagnostopoulou, N. A. Grammatikos, D. Apostolou, G. Mentzas, Trustworthy ai in education: A roadmap for ethical and effective implementation, in: Proceedings of the 13th Hellenic Conference on Artificial Intelligence (SETN '24), ACM, 2024. doi:10.1145/3688671.
[5] W. Holmes, M. Bialik, C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning, Center for Curriculum Redesign, 2019.
[6] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, 2019.
[7] European Parliament, Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (ai act), Official Journal of the European Union, 2024.
[8] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: A review, ACM Computing Surveys 55 (2022) 1–38.
[9] O. Zawacki-Richter, V. I. Marín, M. Bond, F. Gouverneur, Systematic review of research on artificial intelligence applications in higher education—where are the educators?, International Journal of Educational Technology in Higher Education 16 (2019) 39.
[10] N. A. Grammatikos, E. Anagnostopoulou, D. Apostolou, G. Mentzas, A conversational digital assistant for stem education, in: 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA), IEEE, 2023, pp. 1–7.
[11] N. A. Grammatikos, E. Anagnostopoulou, D. Apostolou, G. Mentzas, An ai-powered learning companion for adaptive and personalized stem education, in: International Symposium on Chatbots and Human-Centered AI, Springer Nature Switzerland, Cham, 2024, pp. 87–95.
[12] E. Seralidou, K. Kioskli, T. Fotis, N. Polemi, Ai_taf: A human-centric trustworthiness risk assessment framework for ai systems, Computers 14 (2025) 243.