Ref. Ares(2026)4470152 - 30/04/2026

D6.2

Report on the initial pilot - education

Related Work Package	**WP6 – Implementation of education LSP**
Related Task	Task 6.2. – Initial pilot
Lead Beneficiary	EA
Contributing Beneficiaries	ICCS
Document version	v.1.0
Deliverable Type	R- Document, Report
Distribution level	SEN – Sensitive
Contractual Date of Delivery	30/04/2026
Actual Date of Delivery	30/04/2026

Authors	**D.Koulentianos (EA)** **S.Sotiriou (EA)** **E. Anagnostopoulou (ICCS)** **N.Grammatikos (ICCS)** **D.Apostolou (ICCS)** **G.Mentzas (ICCS)**
Contributors	-
Reviewers	Nic Fair (UoS), Mark Tanner (VPF), Daniel Diosdado (VPF)



Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.

Version history

Version	Description	Date completed
0.1	First Draft	23/03/2026
0.2	Appendixes added	24/03/2026
0.3	Internal review version	27/03/2026
1.0	Final version	30/04/2026

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer

This document contains material, which is the copyright of one or more FAITH consortium parties, and may not be reproduced or copied without permission.

All FAITH consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the FAITH consortium as a whole, nor individual FAITH consortium parties, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, accepting no liability for loss or damage suffered by any person using this information.

Executive summary

This deliverable reports on the initial pilot phase of the Educational Large-Scale Pilot (LSP) based on the work done in Task 6.2. The main goal of this task is to conduct the in-isolation assessment of the Educational LSP AI Solution using the FAITH TAF V1 Framework across all the AI lifecycle phases.

The Educational LSP AI Solution was deployed, and the platform is available at https://trust-ai-lab.eu/, where both teachers and students can access and use it with credentials assigned to them by the teachers. During the initial pilot phase, the inquiry-based scenarios for different STEM subjects were implemented in the real classroom environments. During the initial pilot, in addition to student implementations, several workshops and training activities for teachers were conducted concerning the trustworthy use of AI in education, where the partners from LSP3 have participated and the Trust AI Lab platform has been presented.

The in-isolation assessment has been successfully conducted, covering the full six-phase trustworthiness evaluation of all AI assets using the FAITH AI TAF V1 methodology, across all lifecycle stages (design, development, and deployment). Furthermore, all models and datasets have been registered and fully described in the FAITH Model Hub, the Trustworthy AI maturity of all relevant participant groups has been measured using TrustSense with the resulting scores incorporated into risk calculations, and the in-isolation trustworthiness assessment conducted using TrustGuard.

Additionally, both formal and informal feedback was submitted for all three FAITH tools (FAITH Model Hub, TrustGuard, and TrustSense) throughout the reporting period. In parallel, formal ethical approval was obtained from the Ethics Committee of EA following the submission of comprehensive system and data documentation. Finally, the regulatory and ethical assessment has been completed, including compliance mapping with EU AI Act high-risk requirements, and assessment of GDPR compliance.

Table of Contents

1 Introduction 10
1.1 Scope and objectives of the 1st pilot in the Education LSP 10
1.2 Structure of the document 10
2 Progress on the AI system deployment and execution of the Pilots 12
2.1 Status of the 1st Pilot execution 12
2.1.1 Deployment context & Domain specifics 12
2.1.2 Training 13
2.1.3 Users 14
2.1.4 Evaluation process 15
2.2 Status of the Replication Pilot preparation/initialization 17
2.3 Data collection throughout the 1st Pilot execution 18
2.4 Integration into LSP workflows: Lessons learned 21
3 Application of FAITH Tools 23
3.1 Application of the FAITH AI TAF V1 methodology 23
3.1.1 Cartography 24
3.1.2 Threat Assessment 25
3.1.3 Consequence Assessment 26
3.1.4 Vulnerability assessment 33
3.1.5 Risk assessment 34
3.1.6 Risk management 34
3.2 Tool Validation Planning & Training Processes 35
3.2.1 Dry-run 35
3.2.2 Plan development 37
3.2.3 Tool Training 39
3.3 FAITH Tool Validation Activities 39
3.3.1 TrustSense 39
3.3.2 Data & AI model Passport in the FAITH AI Model Hub 42
3.3.3 TrustGuard 43
3.4. Threat Analysis 44

3.4.1 Learning Companion 44
3.4.2 Student Assessment 47
3.4.3 AI Learning Recommendations 50
3.5. Initial Risk Assessment results 54
3.5.1 Learning Companion 54
3.5.2 Student Assessment 56
3.5.3 AI Learning Recommendations 57
3.6. Countermeasures 59
3.6.1 AI Learning Companion 59
3.6.2 Student Assessment 61
3.6.3 AI Learning Recommendations 63
3.7 Formal Tool Feedback 69
3.7.1 FAITH Model Hub 69
3.7.2 TrustSense 70
3.7.3 TrustGuard 70
4 AI System Evaluation 71
4.1 Quantitative metrics on pilot data 71
4.1.1 Learning Companion 71
4.1.2 Student Assessment 72
4.1.3 AI Learning Recommendations 74
4.2 Trustworthiness Assessment 78
4.2.1 Validation & Reliability 79
4.2.2 Explainability & Interpretability 79
4.2.3 Security, Transparency, Safety, Fairness and Privacy 80
4.3 Preliminary Impact Evaluation 81
5 Regulatory and Ethical Assessment 82
5.1. Alignment with the EU AI Act (High-Risk AI requirements) 82
5.2. GDPR Compliance Assessment 84
5.3 Ethics Committee Approval 84
6 Conclusions and Roadmap for the Final Pilot Phase 87
6.1 Summary of achievements in the second period 87

6.2 Plan for Replication Phase 87

List of Abbreviations

Abbreviation	Explanation
LSP	Large-Scale Pilot
STEM	Science, Technology, Engineering, and Mathematics
GDPR	General Data Protection Regulation
AI	Artificial Intelligence
WP	Work Package
AWS	Amazon Web Services
LAN	Local Area Network
SUT	System Under Test
LLM	Large Language Model
RAG	Retrieval-Augmented Generation
TWAs	Trustworthiness Attributes
HLEG	High-Level Expert Group on Artificial Intelligence
NLU	Natural Language Understanding

List of Tables

Table 1: Events that have taken place concerning the use of AI in education, where the partners from LSP 3 have participated and the Trust AI Lab platform has been presented. 14
Table 2: A detailed time plan of the implementations that have taken place within LSP 3 until M24 of the FAITH project. During the first pilot most of the implementations aimed at training the model, nevertheless AI-assisted implementations took place for the simple pendulum scenario. 20
Table 3: Overview of the training workshops attended by the LSP3 team 39
Table 4: List of the identified threats, along with their frequency and a corresponding explanation for the Learning companion. 47
Table 5: List of identified threats, along with their frequency and a corresponding explanation for the student assessment. 50
Table 6: List of the identified threats, along with their frequency and a corresponding explanation for the AI Learning Recommendations. 54
Table 7: Overall risk levels identified by TrustGuard for the Learning Companion 56
Table 8: Overall risk levels identified by TrustGuard for the Student Assessment model 57
Table 9: Overall risk levels identified by TrustGuard for the AI Learning Recommendations 58

Table 10:Mitigation measures suggested by TrustGuard for the Learning Companion 61
Table 11: Mitigation controls related to the Student Assessment 63
Table 12: Mitigation measures suggested by TrustGuard regarding the AI Learning Recommendations 66
Table 13: Rasa Intent and its equivalent F1 score and support. 71
Table 14: XGBoost metrics for classifying students in the three categories 74
Table 15: Per category results on the hold-out test set 74
Table 16: Initial Q-Table from the first set of teachers. 76
Table 17: Mapping of EU AI Act High-Risk AI Requirements in the Educational LSP AI Solution. 84
Table 18: Educational LSP AI Solution — Alignment with HLEG Trustworthy AI Principles 86

List of Figures

Figure 1: The Trust-AI-Lab platform allows the teacher to create anonymized credentials and assign specific inquiry-based scenario(s) to them, certifying student anonymity in this way. 13
Figure 2: Each user that gets a teacher's account should create an organization for the school he/she belongs to or be added to an existing organization. 13
Figure 3: Activities at risk. 16
Figure 4: AI recommendations 16
Figure 5: The iterative evaluation processes. The left panel illustrates how the inquiry-based scenario created by the teacher is the basis for the creation of the AI-enhanced scenario, through a process that is constantly monitored by the teacher. In the right panel, a new iterative cycle is shown, with the AI-enhanced scenario being the basis for the acquisition of data, new AI recommendations, and the creation of further AI-enhanced scenarios. 17
Figure 6: Initial scenario created by teachers s (left), AI-enriched scenario (right). 19
Figure 7: Sankey plot for simple pendulum scenario, showing how the students transition between the Low (red), Moderate (orange) and High (green) performance levels, as they move through the different inquiry-phases. 21
Figure 8: Sankey plot summarizing the results of the photoelectric effect implementations. 21
Figure 9: FAITH AI TAF V1 methodology 23
Figure 10: Preliminary Threat, Vulnerability, and Control Measure Register 24
Figure 11: Cartography of the Educational LSP AI Solution 24
Figure 12: Cartography of the Educational LSP AI solution. 35
Figure 13: Criticality assessment of the Educational LSP AI solution 36
Figure 14: Business consequences of the lack of trustworthiness in the Educational LSP AI Solution 36
Figure 15:Control questions for the deployment phase of the Educational LSP AI solution. 37
Figure 16:Detailed Implementation Plan of the Educational LSP 38
Figure 17: Trustworthiness level result of the technical team. 40
Figure 18: Trustworthiness level result of the teachers 41
Figure 19: Datasets of the Educational LSP AI Solution created using the FAITH AI Model Hub . 42

Figure 20: AI Models of the Educational LSP AI Solution created using the FAITH AI Model Hub43
Figure 21: List of in-isolation trustworthiness assessment using the TrustGuard 43
Figure 22: Threat Assessment for the Learning Companion.....44
Figure 23:Consequence assessment for the Learning Companion.....45
Figure 24: Vulnerability assessment for the Learning Companion 46
Figure 25: Threat Assessment for the Student Assessment 48
Figure 26: Threat Assessment for the AI Learning Recommendations.....51
Figure 27: Mitigation controls selected by the risk assessor for threats related to the validity of the AI Learning Companion.....59
Figure 28: Mitigation controls selected by the risk assessor for threats related to the validity of the Student Assessment.....61
Figure 29: Mitigation controls selected by the risk assessor for threats related to the validity of the AI Learning Recommendations 63
Figure 30: Feature Importances of XGBoost Classifier. 73
Figure 31: Confusion Matrix between XGBoost and PISA's actual categorization in the platform. 74
Figure 32: Validity (left) and Explainability (right) of the AI Learning Recommendations as evaluated by teachers.....77
Figure 33:Transparency (left) and Fairness (right) of the AI Learning Recommendations as evaluated by teachers.....77
Figure 34:Privacy (left) and Security (right) of the AI Learning Recommendations as evaluated by teachers 78
Figure 35: Safety (left) and Trust (right) of the AI Learning Recommendations as evaluated by teachers 78

1 Introduction

1.1 Scope and objectives of the 1st pilot in the Education LSP

The Educational Large-Scale Pilot (LSP) is one of the LSPs where the trustworthiness assessment of the FAITH project is being implemented, tested, and validated under realistic conditions of operation. The pilot was implemented in accordance with Task 6.2, which is run by EA and ICCS. ICCS operationalizes it in cooperation with EA where the Educational LSP AI Solution was deployed, and the main source of pilot participants were students and teachers whose interaction with the AI system formed the empirical core of the pilot.

The partners involved in the context of Task 6.2 arrange a set of implementation practices whereby the AI technologies of deeper learning in the field of Science, Technology, Engineering, and Mathematics (STEM) are introduced into practice in the real classroom environments. The application piloting is performed in cycles of repeated user-extended trials based on inquiry-based learning. This iterative process that occurs in cycles was intended to make sure that the AI solution was improved with the evidence of the field and that the credibility of the solution was evaluated with the help of successive data collection, analysis, and reflection.

The Educational LSP AI Solution is an AI-driven platform designed to support personalized learning in STEM, help teachers in assessing student problem-solving skills, and generate AI learning recommendations to improve educational outcomes. The initial pilot has the following specific objectives in the framework of the FAITH project:

- Implement the Educational LSP AI Solution into practice in a real-world school setting, involving both teachers and students in the school of the intervention in systematic and repeated contact with the AI system according to the principles of inquiry-based learning.
- The initial in-isolation trustworthiness assessment of all AI assets of the Educational LSP AI Solution in the FAITH AI TAF V1 methodology, including all six assessment phases Cartography, Threat Analysis, Consequence Assessment, Vulnerability Analysis, Risk Analysis, and Risk Management, in all AI lifecycle phases (design, development, deployment).
- Implement the FAITH tools TrustGuard, TrustSense and the FAITH Model Hub on the Educational LSP context and give formal feedback on the FAITH tools development teams in a structured way to inform tool improvement.
- Determine the regulatory and ethical compliance between the Educational LSP AI Solution and the high-risk system requirements of the EU AI Act as well as the General Data Protection Regulation (GDPR).
- Prepare lessons learned of the implementation of the AI solution in the LSP educational workflows and establish a roadmap of the Final Pilot Phase, encompassing the preparation of the Replication Pilot and transition to the socio-technical system trustworthiness assessment based on TAF V2.

1.2 Structure of the document

This deliverable is divided into six key areas, each dealing with a specific aspect of the Educational LSP pilot activities and trustworthiness evaluation that will be undertaken in the first pilot phase.

Section 1 - Introduction prepares the groundwork of the deliverable. It determines the scope and goals of the first pilot in the Educational LSP, giving the reader the needed context to comprehend the scope and purpose of the reported activities

Section 2 - Progress on the AI System Deployment and Execution of the Pilots reports concerning the practical implementation of the first pilot. It addresses the context of deployment and domain-related peculiarities of the Educational LSP, the training processes that take place with the participating educators and students, the user groups profile of the involved people, and the evaluation process. It also deals with the status of the Replication Pilot preparation, and the data collection operations performed during the pilot, the implementation time plan and the results thereof. The section ends with the lessons gained when incorporating the AI solution into the LSP workflows.

Section 3 - FAITH Tools Application explains the implementation of FAITH AI TAF V1 to perform an in-isolation trustworthiness assessment of all AI assets of Educational LSP AI Solution by using the six steps of the methodology one by one. It is also concerned with the tool validation planning and training procedures that were related to the FAITH tools during the pilot. Also, a formal tool feedback report on the systematic feedback process used on the three FAITH tools, including when informal and formal feedback was given, by whom, and providing a high-level evaluation of tool usefulness and usability.

Section 4 - AI System Evaluation, describes quantitative metrics based on pilot data and reports on how the FAITH Tools have provided insights into how the AI assets are performing against the seven Trustworthiness Attributes. Additionally, it reports on the impact the AI system is having in the Educational LSP.

Section 5 - Regulatory and Ethical Assessment evaluates the Educational LSP AI Solution against the regulatory and ethical frameworks that it is subjected to, discussing the high-risk requirements of the EU AI Act and compliance with GDPR standards.

Section 6 - Conclusions and Roadmap to the Final Pilot Phase summarizes the main findings of the deliverable, outlines the major conclusions of the pilot implementation and trustworthiness analysis.

2 Progress on the AI system deployment and execution of the Pilots

2.1 Status of the 1st Pilot execution

In this section we will discuss in detail how pilot implementation took place, the results obtained from the implementation, as well as how the results were used in order to get a number of recommendations from the AI system to improve the students' performance and create personalized learning pathways.

2.1.1 Deployment context & Domain specifics

The Educational LSP AI Solution enables the transformation of traditional classroom settings into dynamic, individualized learning environments. Each student in the deeper learning classroom follows a customized learning pathway suggested by the AI learning companion and tailored to his or her needs. Our approach evaluates students' problem-solving skills and tracks their development over time. It also supports students by providing them with tailored feedback.

The Educational LSP AI Solution consists of three main components: the AI Learning Companion, the authoring tool, and the analytics dashboard. Deliverable 6.1 provides additional details about the implementation of the Education LSP. Currently, our solution is TRL 6.

The Educational LSP AI Solution is hosted on a robust web infrastructure to ensure reliability, scalability, and efficiency. The platform operates as a web application that can be accessed at https://trust-ai-lab.eu/ . The Educational LSP AI Solution was created to be used in classroom laboratories, where the students have their own computers. The platform has been designed to work in a hybrid environment which is cloud-based infrastructure and on-site student and teacher devices.

The platform is implemented using the Django framework, runs on Amazon Web Services (AWS) and has PostgreSQL that runs on AWS RDS that offers high availability, automatic backup and secure access control. Students can access the platform using the computer in the labs during lessons, and teachers can use the platform using their laptops to create learning scenarios and track the progress of students. Every connection is made to the cloud-based application by the local area network (LAN) of the school and establishes a smooth learning process.

The Educational LSP Solution handles anonymized student interaction data, and it doesn't store any information that allows the identification of a student. Before implementation for a specific inquiry-based scenario takes place, the teacher is creating anonymized credentials for the students through the system, as shown in Figure 1. The specific credentials correspond to the educational scenario(s) assigned by the teacher and once students log in to the platform, they only have access to the specific scenario(s). This process certifies student anonymity. Moreover, each teacher can only know either how his/her students performed in the specific scenario or get a total overview of how all students have performed in the specific scenario, without knowing any information about student groups created by other teachers.

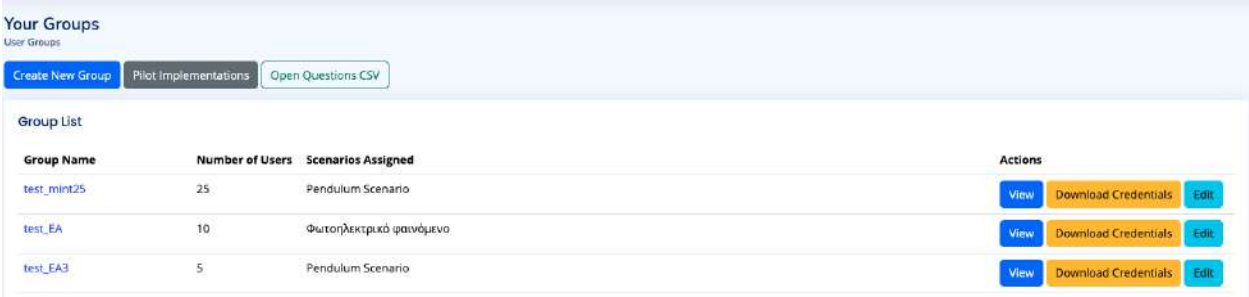


Figure 1: The Trust-AI-Lab platform allows the teacher to create anonymized credentials and assign specific inquiry-based scenario(s) to them, certifying student anonymity in this way.

In addition, any new member that signs up to the platform needs to provide a dedicated access code to get a teacher account and create an organization (see Figure 2) for his/her school or be added to an existing one. The above processes certify full student anonymity, even in the unlikely event that the platform has been compromised by a third party.

Finally, the platform has been thoroughly inspected by the ethics committee of EA, from which the collaborators of EA working on LSP 3 have been obviously excluded, and an approval letter has been sent to the coordinator of the FAITH project.

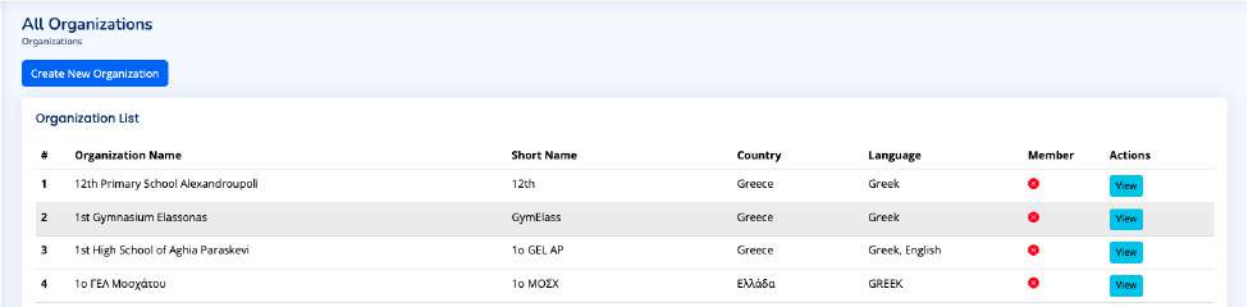


Figure 2: Each user that gets a teacher's account should create an organization for the school he/she belongs to or be added to an existing organization.

The development of the Educational LSP AI Solution has been based on a participatory design approach, ensuring that the system can accurately assess students' problem-solving competence and create personalized student pathways, in order to fully enhance the learning experience. This has been done in a way which is both technically robust and pedagogically sound. The main goal of this use case is to evaluate the trustworthiness of AI-based tools in the context of inquiry learning during laboratory activities.

2.1.2 Training

Parallel to the technical training of the system, the LSP 3 team prioritized extensive outreach and validation with the educational community. The platform was prominently featured during dedicated summer schools and workshops, where educators had the opportunity to engage with the tool and interact with the team of LSP 3. These sessions served as a critical testing ground for the platform's user interface and functional logic, providing the development team with

important insights into how the AI-driven pathways align with established teaching methodologies. Through workshops, teachers evaluated the trustworthiness of the AI system, highlighting which trustworthiness characteristics are most essential for establishing trust in classroom AI tools. These events are collated in Table 1.

To broaden the scope of feedback beyond physical events, a series of online meetings were conducted with educators across Greece. These digital consultations focused on demonstrating the platform's core mechanics and discussing its practical integration into the classroom. The reception from the teaching community has been overwhelmingly positive, characterized by a high level of professional interest and curiosity. This enthusiasm is quantitatively reflected in the platform's growing user base, which currently counts approximately 400 registered educators. This milestone indicates a strong market readiness and a genuine demand for AI-enhanced personalized learning tools within the national educational landscape.

Date	Name of the event	No of educators that participated
April 2024	AI in primary and secondary education	15
July 2024	ESIA summer school 2024	7
January 2025	MINT EC 2024	30
July 2025	ESIA summer school 2025	10
November 2025	EDEN Open Classroom Conference: Empower Educators and Learners for the Age of AI	60
November 2025	11th Panhellenic eTwinning Conference	5
December 2025	AI and Books: Shaping the Lesson of the Future	20

Table 1: Events that have taken place concerning the use of AI in education, where the partners from LSP 3 have participated and the Trust AI Lab platform has been presented.

Furthermore, the engagement strategy extended into deeper capacity building through specialized workshops focused on the maturity of teachers in the age of Artificial Intelligence. A subset of the registered users participated in these sessions, which explored the ethical, technical, and pedagogical implications of AI in education. By situating the platform within this broader framework of digital maturity, the FAITH project has not only gathered feedback on the tool itself but has also fostered a community of practice. These interactions have confirmed that the platform is perceived not just as a standalone utility, but as a catalyst for professional growth, helping teachers navigate the transition toward AI-supported instruction with confidence.

2.1.3 Users

The engaged stakeholders of the Education LSP include the end users, the AI developers and the educational stakeholders. The end users of the Education LSP are the students and the teachers.

Students interact with the Educational LSP AI solution and provide (anonymized) data in order to train the system. Teachers create inquiry-based scenarios using the system and use the system in order to monitor and improve student performance.

In addition to the end users there are the AI developers who handle all the technical details and deliver a trustworthy and accountable AI system, which supports the needs of both the students and the teachers.

Finally, other users of the Educational LSP AI solution are the educational AI champions and policy makers to whom the obtained data, and the developed system will be presented to. They will provide recommendations and future policy actions, so that AI can become a part of the school curriculum.

2.1.4 Evaluation process

The evaluation process consists of four distinct steps and takes place through consecutive iterations. The four steps that take place at each iteration, namely the creation of the inquiry-based scenario by the teacher, the implementation of the scenario and the acquisition of student data, the AI recommendations of the conversational agent, and finally the creation of an AI-enhanced scenario shall be discussed in detail next.

- ***Creation of an inquiry-based scenario by the teacher***: As it was extensively discussed in D6.1, the initial version of the inquiry-based scenario is prepared by the teacher, who is the one that created all the activities at each of the five inquiry-phases. In D6.1 a very complicated inquiry-based scenario concerning the photoelectric effect was presented and discussed explicitly. We note at this point that the system allows teachers from the same organization to work simultaneously in the same scenario, to give their input, cross check the process, and share the workload, especially in the very early stages of the creation, as the process is very demanding.
- ***Implementation and data acquisition***: Once the scenario is ready, it has to be implemented in the classroom with students. Based on the acquired data the AI agent will identify activities at risk, as shown in Figure 3, and provide recommendations. We note here that all student data are fully anonymized, while the student performance can be monitored by the teacher through dashboard analytics. Students and teachers are asked to fill in questionnaires at the end of each implementation run, so that the team of LSP 3 can get their feedback.



Figure 3: Activities at risk.

- ***AI recommendations***: Based on several metrics obtained through the acquired data, a number of recommendations are provided by the AI agent to the teacher(s) to better assess students' problem-solving competence and help students overcome difficulties. A recommendation can be a new activity, as shown in Figure 4, or to omit an activity as it might be trivial for the students and has no pedagogical meaning. Teachers are reviewing the recommendations of the AI agent, which they accept, reject or edit. In addition, teachers are asked to fill in further questionnaires about the quality of the AI recommendations and how relevant they are to the learning topic.

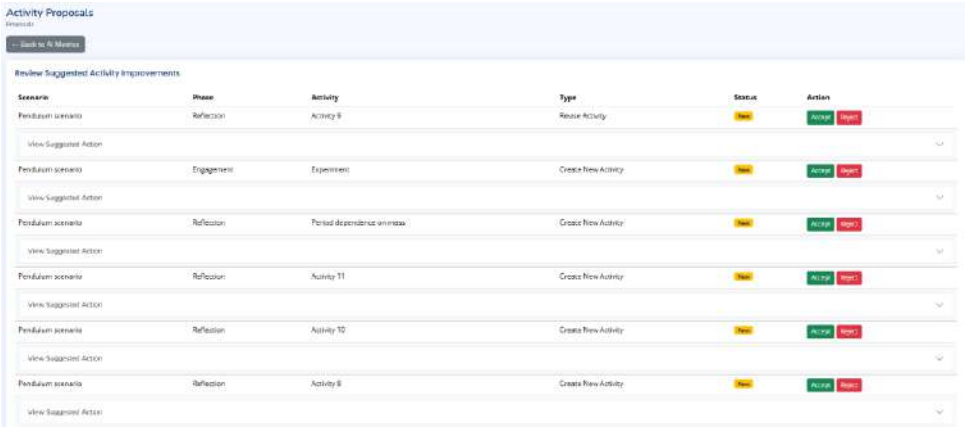


Figure 4: AI recommendations

- ***Creation of the AI-enhanced scenario***: As can be seen from Figure 4 above, the teacher has the option to accept, reject or edit a proposed activity. Once this process has been completed a new, AI-enhanced, version of the scenario has been created.

The process described above is the first iteration cycle, as shown in the left panel of Figure 5. The AI-enhanced scenario will become the basis for a new iteration, with 2 and 3 being the same as above, and eventually the creation of a further enhanced AI-scenario. It is important to stress at

this point that the teacher is present at all stages of the different versions of the AI-enhanced scenario, ensuring the *human-in-the-loop* procedure.

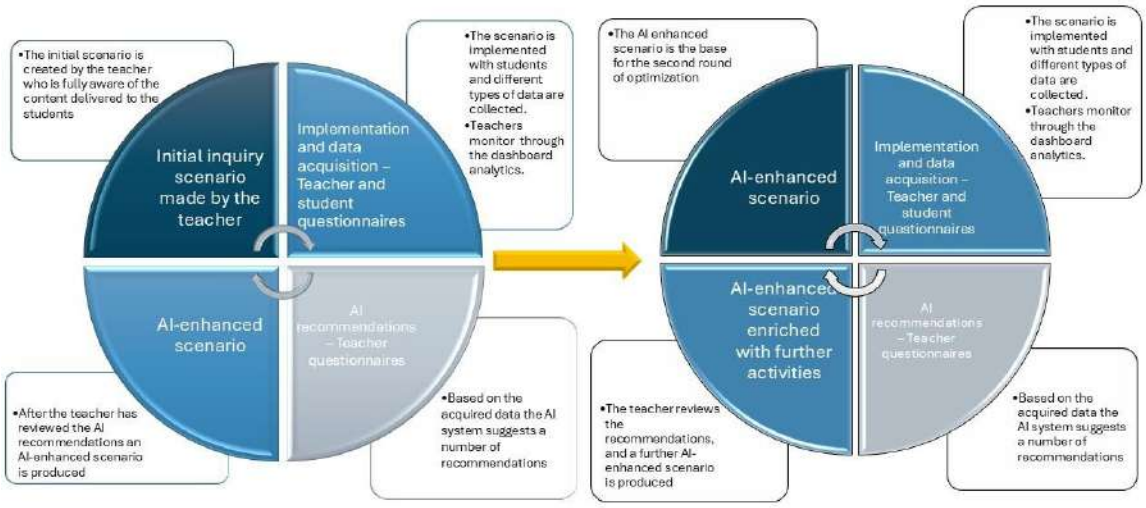


Figure 5: The iterative evaluation processes. The left panel illustrates how the inquiry-based scenario created by the teacher is the basis for the creation of the AI-enhanced scenario, through a process that is constantly monitored by the teacher. In the right panel, a new iterative cycle is shown, with the AI-enhanced scenario being the basis for the acquisition of data, new AI recommendations, and the creation of further AI-enhanced scenarios.

2.2 Status of the Replication Pilot preparation/initialization

During the replication pilot, we are going to perform the in-context trustworthiness evaluation of the Educational LSP AI solution in accordance with the FAITH_AI_TAF_v2 framework and using the FAITH_AI_TAF_v2 tools. First, we intend to attend the training workshops on FAITH_ AI-TAF tools and then we will evaluate the trustworthiness of the Educational LSP solution with the FAITH-AI-TAF tools.

The replication pilot of the FAITH project began in August 2025 (M20), and student implementations have begun, enabling the collection of data for the replication pilot. This phase marked an important step in validating and extending the project's Artificial Intelligence supported learning framework. As part of this phase, the project collaborated with the Balkan Physical Union Summer School to implement physics learning scenarios based on core topics of quantum mechanics, since 2025 was announced by UNESCO as the International Year of Quantum Science and Technology[1]. Three major experiments were selected as the focus of the activities: the photoelectric effect, the double slit experiment and the Franck Hertz experiment.

[1] https://www.unesco.org/en/years/quantum-science-technology

A total of seventy-five students participated, generating valuable interaction data that will be used to train the AI agent in order to provide recommended activities.

These implementations provided an opportunity to demonstrate the educational value of the FAITH scenario-based learning approach in real classroom conditions. At the same time, they created a substantial evidence base that will support the improvement of the recommendation system. The collected data will help refine the artificial intelligence engine so that it can offer more accurate and personalized guidance to learners. By examining the learning paths, difficulties and misconceptions that students encounter while working on quantum mechanics scenarios, the system will be able to provide better support to both teachers and students.

In addition to the activities of the Balkan Physical Union Summer School, an expanded implementation was carried out at Ellinogermaniki Agogi. During this phase, 125 students worked with the AI-enhanced version of the pendulum scenario. This large-scale deployment provided a broader dataset for evaluating the performance of the system in a more traditional mechanics context. Further implementations are already planned for 2026 at Ellinogermaniki Agogi as well as at other collaborating schools. These future activities will widen the range of student profiles, contributing to more robust and general recommendations.

2.3 Data collection throughout the 1st Pilot execution

During the initial pilot phase, the inquiry-based STEM scenarios were implemented repeatedly at Ellinogermaniki Agogi as well as in additional schools across Greece and other European countries. These implementations served both to validate the pedagogical approach and to generate the student activity data required for training the AI agent. A full description of the underlying pedagogical framework and the student assessment methodology was already provided in D6.1; here, we briefly summarize the essential elements to contextualize pilot progress.

Each inquiry-based scenario followed the five phases of the 5E instructional model (Engagement, Hypothesis, Experiment, Analysis, and Reflection), with each phase corresponding to a distinct problem-solving process as defined in PISA 2015. Students were categorized as Low, Moderate, or High performers at each phase as well as overall, using a weighted scoring scheme based on the PISA 2015 framework. This structured assessment enabled the generation of high-quality data traces across multiple stages of the learning process.

To properly train the AI agent and enable it to generate meaningful and context-aware recommendations, a minimum of 500 student implementations per inquiry scenario was required. This target was achieved for the simplest available inquiry scenario: the Simple Pendulum. Once the threshold of 500 implementations had been reached and processed, the AI agent produced a set of recommended student activities tailored to the structure and learning goals of the scenario. These recommendations were subsequently reviewed by the scenario author, who could accept, reject, or edit each proposed activity. This interactive process not only ensured pedagogical validity but also contributed to further refinement of the AI system through iterative feedback.

By the end of the initial pilot phase, this procedure resulted in the creation of a new, AI-enhanced inquiry-based scenario that incorporated the validated recommendations. This progress demonstrated the feasibility of combining large-scale classroom data collection with human-in-the-loop AI training and provided the foundation for the replication phase to follow.

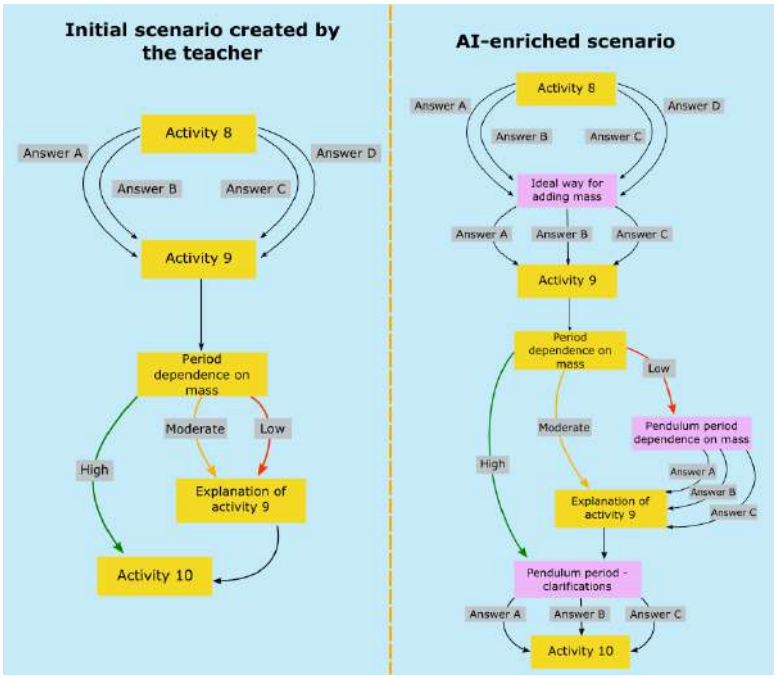


Figure 6: Initial scenario created by teachers s (left), AI-enriched scenario (right).

A primary focus of the initial phase involved direct interaction with the students, as explained above, to facilitate the acquisition of high-quality data. This data collection process was essential for training AI algorithms, allowing the system to learn from real-world educational patterns and refine its ability to suggest tailored learning trajectories. By involving students early in the lifecycle, the project ensured that the underlying models reflect the actual needs and behaviors of the learners they intended to serve.

In D6.1 a general time plan for the different implementations had been presented. Here we mention in detail all the implementations that have taken place by M24 in Table 2.

Month/Year	Inquiry-based scenario	Training the model/ AI-assisted	Number of implementations
May 2024 (M5)	The simple pendulum	Training the model	125
August 2024 (M8)	The photoelectric effect	Training the model	75
	Observations Galileo	Training the model	75
November 2024 (M11)	The simple pendulum	Training the model	150
	The photoelectric effect	Training the model	10

March-April 2025 (m15-M16)	Circuits	Training the model	100
	The pH	Training the model	100
	Photosynthesis	Training the model	100
June 2025 (M18)	The simple pendulum	AI-assisted	65
	The photoelectric effect	Training the model	35
	Observations of Galileo	Training the model	20
	Photosynthesis	Training the model	30
August 2025 (M20)	The photoelectric effect	Training the model	75
	The double-slit experiment	Training the model	75
	The Franck-Hertz experiment	Training the model	75
November 2025 (M23)	The simple pendulum	AI-assisted	125
	The photoelectric effect	Training the model	30
	The double-slit experiment	Training the model	30

Table 2: A detailed time plan of the implementations that have taken place within LSP 3 until M24 of the FAITH project. During the first pilot most of the implementations aimed at training the model, nevertheless AI-assisted implementations took place for the simple pendulum scenario.

The implementation results are available to the authorized teachers in the dashboard analytics of the Trust AI Lab platform. In D6.1 discussed how the results of an implementation can be visualized through a Sankey plot. Here we present Sankey plots from student data for the simple pendulum scenario and the photoelectric effect scenario. The plots are shown in Figure 7 and Figure 8, for the simple pendulum and the photoelectric effect respectively.

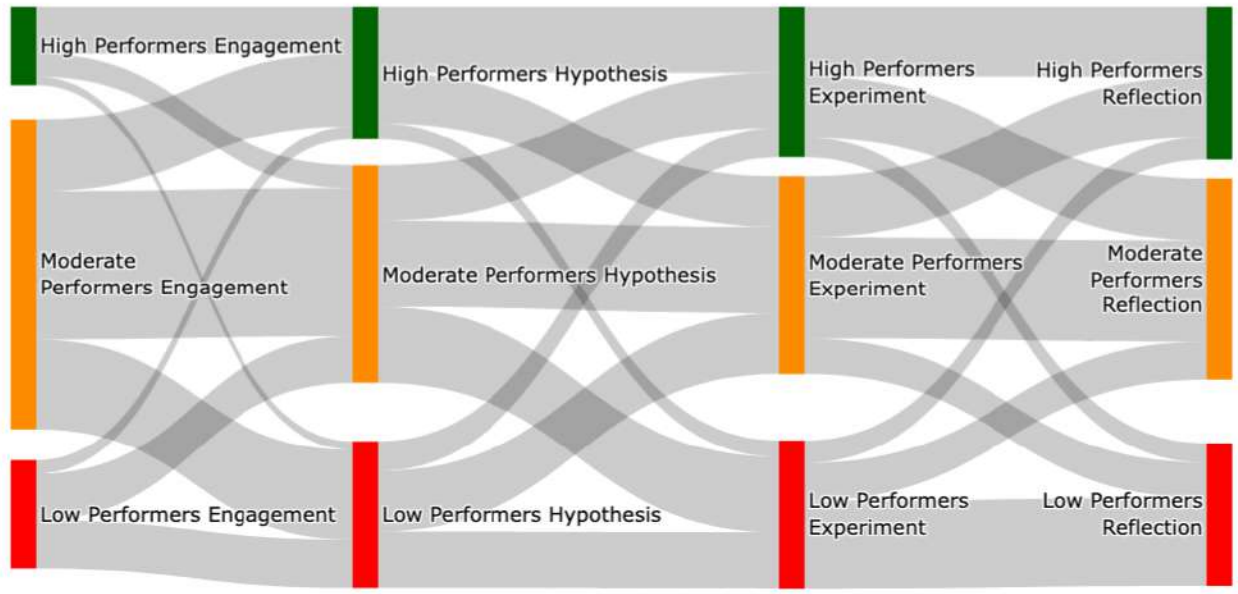


Figure 7: Sankey plot for simple pendulum scenario, showing how the students transition between the Low (red), Moderate (orange) and High (green) performance levels, as they move through the different inquiry-phases.

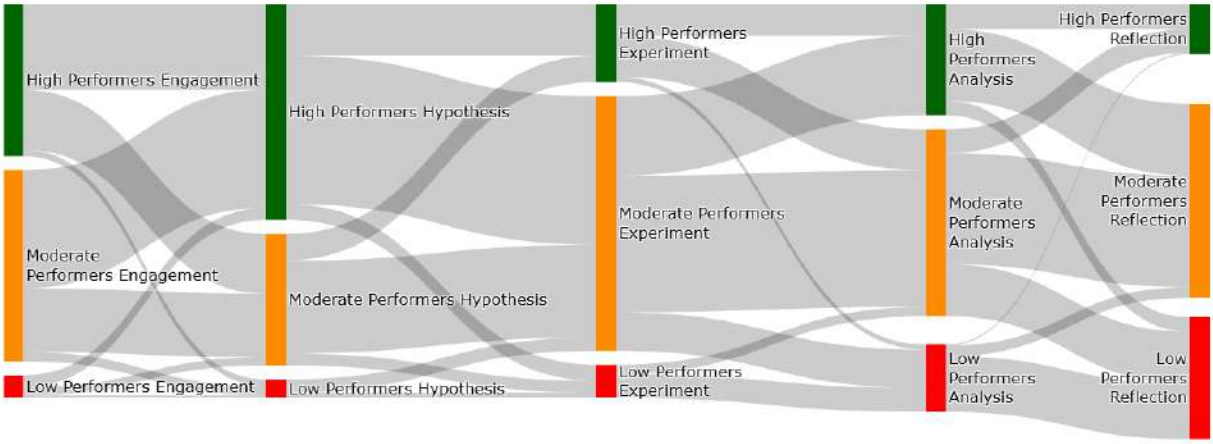


Figure 8: Sankey plot summarizing the results of the photoelectric effect implementations.

Further implementations have been scheduled from M21 until the end of the FAITH project and will be discussed on D6.3.

2.4 Integration into LSP workflows: Lessons learned

The Education Large Scale Pilot (LSP) examined how trustworthy AI could be meaningfully and responsibly integrated into real school environments. Its main focus was to introduce, test, and refine an AI-based Digital Assistant that supported inquiry-based STEM learning, while ensuring that all dimensions of trustworthiness, such as safety, transparency, explainability, privacy protection, and fairness were respected throughout the AI lifecycle. This work aligned with the FAITH project's evaluation methodology, which drew on NIST and ENISA guidance for assessing AI systems under real-world conditions.

The pilot was implemented in three phases: an initial coordination and setup phase, the first round of classroom implementation, and a broader replication phase across additional schools. Throughout these stages, the pilot collected continuous evidence regarding how reliably the AI behaved in classroom activities, whether teachers and students were able to understand and interpret its outputs, and whether the safeguards designed during the development phase functioned effectively during real educational use. This included reviewing the clarity of model explanations, evaluating the robustness of the system in daily classroom conditions, and ensuring full compliance with data protection and ethical requirements outlined in the WP6 work description.

During the early stages of implementation, several practical difficulties emerged that affected the quality of the recorded data. Many teachers initially misunderstood the level of interaction expected during the AI-supported activities and did not introduce the taught subject or actively engage with their students while the system was being used. As a result, very low interaction times were captured in the first implementations. After teachers reached out to the LSP3 team for clarification, targeted guidance was provided, instructing them to always give a proper introduction to the topic, interact with learners throughout the activity, and support the inquiry-based process as intended. These clarifications helped stabilize the implementation quality across classrooms.

In addition, the assessment mechanism required revision. Initial observations showed that students who were identified as low or moderate performers during the evaluations intended for all students occasionally exited the feedback loop as high performers. This misalignment increased the time spent by high performers in phases that were not designed for them and contradicted the instructional design of the scenario, where low and moderate performers should remain longer in early inquiry stages and high performers should progress more rapidly to later phases. Adjustments to the assessment process were therefore necessary to safeguard the pedagogical integrity of the learning scenarios, thus the primary evaluations were introduced to label a student's performance in each inquiry phase.

The pilot produced a clear, evidence-based understanding of how trustworthy AI could enhance inquiry-based STEM learning and documented the system's strengths, limitations, and pedagogical value in classroom settings. It also generated refined, reusable learning scenarios that combined sound pedagogical design with responsible AI practices. Furthermore, the pilot established a sustainable community of practice among participating teachers, enabling continued exchange of experiences, resources, and best practices. This community strengthened long-term adoption and ensured that the integration of AI in education continued to follow ethical and pedagogically grounded standards beyond the duration of the FAITH project.

3 Application of FAITH Tools

3.1 Application of the FAITH AI TAF V1 methodology

The in-isolation trustworthiness assessment of the Educational LSP AI Solution was conducted using the FAITH AI TAF V1 methodology. In alignment with FAITH TAF v1, the AI assets (human and AI) of the Educational LSP AI Solution are assessed in isolation at the asset level during this phase of the FAITH journey. The assessment focuses on the intrinsic trustworthiness properties of each individual AI asset without yet considering the wider context of the connected socio-technical system in which those assets operate. The broader systemic assessment, which will examine the interactions, dependencies, and emergent risks arising from the integration of human and AI assets within the full Educational LSP ecosystem, will follow from Milestone M33 in the next iteration of this deliverable, applying TAF V2.

The FAITH AI TAF V1 methodology consists of six sequential phases, as presented in Figure 9: (1) Cartography, (2) Threat Analysis, (3) Consequence Assessment, (4) Vulnerability Analysis, (5) Risk Analysis, and (6) Countermeasures / Risk Management. FAITH tools were used to support all TAF phases and tasks.

As the FAITH tools were being prepared for use by the pilots, a substantial volume of preparatory work was conducted by the LSP3 team in advance of the formal tool-supported trustworthiness assessment. This preparation step was necessary to make the following tool-based evaluation effective and high fidelity.

First, Educational LSP AI Solution dry run execution took place. Section 3.3.1 describes the dry run activities which gave the LSP3 team an operational perspective of the components of the AI solution and its data flows, before the formal assessment.

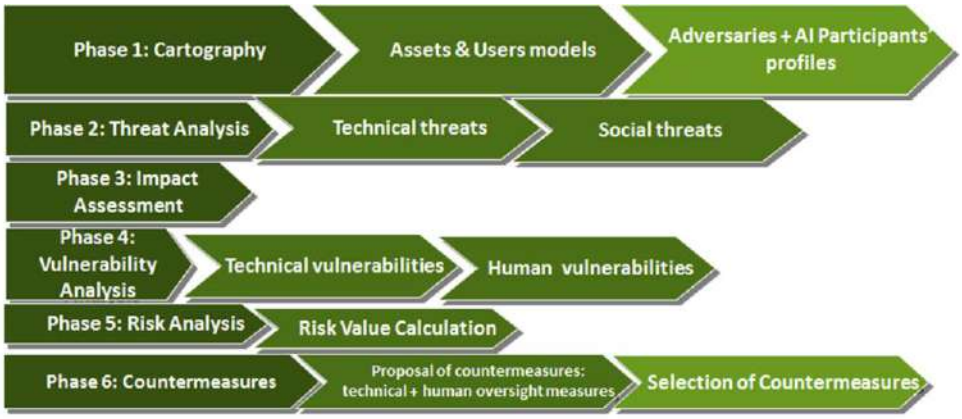


Figure 9: FAITH AI TAF V1 methodology

Besides the dry run, the LSP3 team has performed a preliminary identification exercise to list threats applicable to the Educational LSP AI Solution, as well as vulnerabilities, possible impacts, and

mitigation controls. This pre-assessment work had been recorded in a formal Excel workbook, which was the base input to the formal threat register that was to be further utilized in the TrustGuard tool. Figure 10 shows the Excel developed by the LSP3 team to show the breadth and depth of the initial analysis done before the in-isolation trustworthiness was analyzed by FAITH tools.



Figure 10: Preliminary Threat, Vulnerability, and Control Measure Register

The risk assessor conducted the in-isolation trustworthiness assessment for the educational LSP AI solution using the TrustGuard tool, which implements the FAITH AI TAF V1 methodology. To conduct the assessment, the TrustSense tool was also used to measure the AI maturity of the participants, and the FAITH AI Model Hub was used to create the models and datasets for the educational LSP AI solution.

3.1.1 Cartography

The Cartography phase represents the initial phase of the TAF V1 methodology, which aims at creating a full list of all AI assets in the Educational LSP AI Solution and describing each of the assets concerning the role, lifecycle stage, and criticality level. During the cartography phase, the risk assessor described the AI assets of the educational LSP solution, including its models and datasets. Figure 11 presents the Cartography of the Educational LSP AI Solution.

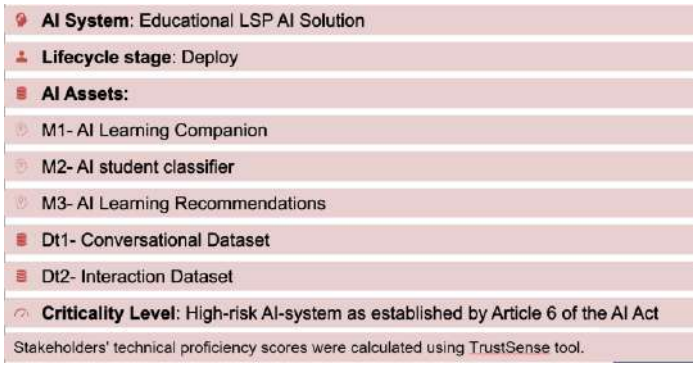


Figure 11: Cartography of the Educational LSP AI Solution

To accomplish this phase, the risk assessor applied three tools of FAITH, FAITH AI Model Hub, the TrustGuard and TrustSense tool. All models and datasets that constitute the AI assets of the Educational LSP AI Solution were registered and described in the FAITH AI Model Hub. The Model Hub is the official source of AI asset data in the FAITH project that stores data like model architecture, training data provenance, version history and intended use. Using TrustGuard requires registration

of assets in the Model Hub since all datasets and models developed in the FAITH AI Model Hub were directly available inside TrustGuard to be selected at the Cartography phase.

After that, the risk assessor used the TrustGuard tool to describe each AI asset of the Educational LSP AI Solution and conducted the assessment on a per-asset basis. 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 assessor to determine whether each AI asset falls within the scope of high-risk AI system provisions and to calibrate the rigor of subsequent assessment phases accordingly.

Additionally, the maturity level of all AI participants involved with the Educational LSP AI Solution was assessed using the TrustSense tool. TrustSense measures the AI maturity of human participants in an AI system. The maturity scores derived from TrustSense were subsequently entered into TrustGuard by the risk assessor on a per-AI-asset basis, enabling the maturity of the human participants associated with each asset to be incorporated into the overall risk calculation in Phase 5.

3.1.2 Threat Assessment

During the Threat Assessment phase, the risk assessor determined the threats associated with each AI component of the Educational LSP AI Solution using the TrustGuard tool. The initial threat identification performed by the LSP3 team in the preparatory phase gave a systematic starting point. The in-isolation trustworthiness assessment using TrustGuard expanded and confirmed this preliminary work in the form of an individual AI asset.

Prior to the in-isolation trustworthiness assessment of the educational LSP AI solution using the TrustGuard tool, a list of threats was compiled. The identified threats include the following:

- The first identified threat is "***Failure or malfunction of ML application"***. In what concerns the AI system developed within LSP3, the threat relates to the failure of the system to:
    - Accurately assess the level of competency of the student. Each educational scenario is intended to both assess and foster a student's problem-solving competence. The assessment model has been explicitly discussed in D6.1, thus we have to make sure that the system can accurately assess students' problem-solving competence and accordingly provide recommendations for the creation of the AI-enchanced scenario.
    - Based on the acquired data the AI system will provide several recommendations (student activities) to enhance the scenario in order to get a more accurate assessment, as well as help the students to develop their problem-solving skills as they move through the different inquiry phases. An identified threat here is that the recommended activities are not relevant to the context of the inquiry-phase or even worse the subject taught.
- Accordingly, the threat of "**Human error**" can be seen. In the context of LSP 3, the specific threat can be referred to different situations of misuse of the AI system:

  - First is the over-reliance of the teacher on the AI system, which means that although the teacher might be present in the process, he does not have control of the process. This contradicts the human in the loop strategic approach towards the trustworthiness of AI systems.
  - Second, the threat might refer to the teacher's lack of experience to identify that the recommendations of the AI system are not relevant to the scope of the scenario or even irrelevant. We remind the reader that all the recommendations of the AI system should be first reviewed by the teacher, who can either accept, reject or edit a recommendation. By doing so the teacher is training an additional AI system, thus it is crucial that the teacher is aware of the educational goals that the system has been designed to meet.
- The next threat is "**Data poisoning**". For the case of LSP 3 this treat is connected with the implementation parameters. In order to train the AI system to serve the needs of students at different competency levels (low, moderate, high) the implementation parameters should make sure that students from all three competency levels are present, otherwise the data will reflect only a specific competence level, with the repercussions that this can have for the AI system.
- Moving forward, the "**Data disclosure**" threat refers to the leakage of student personal data. In our case we have made sure that no personal student data exists on the system and the students log in to the system with anonymized credentials provided by the teacher.
- The next identified threat is "**Bias in Personalized Learning Recommendations**." At this stage, the educational scenarios hosted in our AI system cover a variety of STEM subjects and do not include personal opinions. As a result, the risk associated with this type of bias does not arise.
- The threat "**Model drift over time**" exists, whereby the use of the AI model can deteriorate in performance when the data distribution changes and thus loss in accuracy and reliability can take place unless the model is continuously monitored, updated and retrained.
- "**Teacher concerns**" is a threat that is associated with the teachers' fear that their role might become obsolete due to the AI system.
- "**Lack of Explainability**" is resulting in concerns by teachers because they are not aware of how the AI system works to produce its outputs, and how they can be benefitted by it.

3.1.3 Consequence Assessment

During the Consequence Assessment phase, the risk assessor — with the support of teachers from the LSP3 educational context — evaluated the potential consequences of each identified threat to the various dimensions of trustworthiness defined within the FAITH AI TAF V1 framework. The assessment was conducted per AI asset.

The consequence level was rated and stored in TrustGuard, based on the domain knowledge of teachers, to contextualize the potential impact in the educational context, per AI asset. The inclusion of teachers in this step is due to the human nature of the FAITH approach, in that the consequences ratings are based on the practical implications of the relevant consequences to the students, teachers, and the community of learners in general and not based on a technical analysis. The results of the consequence assessment per AI asset are detailed in Section 3.4.

The consequences of the above-mentioned threats are discussed here for each of the three AI models explained above (AI Learning Companion, Student Assessment and AI Learning Recommendations). The dimensions that are considered are *safety, security and resilience, explainability and interpretability, privacy, fairness and bias, accountability and transparency* and *validity and reliability*.

4.3.1 AI Learning Companion

- ***Safety***: The AI Learning Companion provides students with real-time guidance, feedback, and adaptive support during inquiry-based learning activities. Its purpose is strictly to assist, not replace, human decision-making, and teacher involvement remains essential for its safe and effective use. All educational content delivered through the AI Learning Companion is always reviewed and approved by the teacher beforehand. If safe operation is compromised, the AI Learning Companion risks losing teacher trust, which would likely lead to the discontinuation of the use of the Trust AI Lab platform. Unsafe functioning may also generate business risks, such as fears among teachers that the system is attempting to replace them or over-reliance on AI that reduces teacher engagement. Ethical risks may include negative impacts on learning quality and the erosion of trust in AI-supported education.
- ***Security and resilience***: The security and resilience of the AI Learning Companion are crucial for maintaining user trust, even though the system does not store sensitive student data and therefore poses no severe consequences in the event of a breach. Because the companion is designed to support student learning rather than issue grades or formal evaluations, it is not an attractive target for malicious attacks. Nevertheless, if its security were compromised, such as through unauthorized manipulation of feedback or content, teachers and students could, as was the case before, lose confidence in its effectiveness and may stop relying on it. Resilience is also essential to ensure continuous, reliable operation and protect the credibility of AI-assisted learning. Should the system become systematically vulnerable to attacks, educators might question its reliability, leading to reduced adoption in classrooms. Ethical risks would also emerge, as broader trust in AI within education could diminish, potentially hindering the integration of beneficial technologies into teaching and learning processes.
- ***Explainability and interpretability***: The AI Learning Companion must be interpretable to ensure that teachers understand why the system provides specific guidance, feedback, or instructional prompts. Although no severe consequences are expected if explainability is limited, since teachers oversee the system and its outputs are not used for high-stakes

decisions, interpretability remains essential in an educational context. When teachers can clearly understand the reasoning behind the AI's support, they are better able to explain it to students, validate it, or override it when necessary. If the Learning Companion's behavior becomes difficult to interpret, educators may lose trust in the system, become less willing to rely on it during classroom activities, and find it harder to address errors or explain the AI's suggestions. This loss of clarity could lead to reduced adoption, even in the absence of direct harm. Ensuring transparency in how the AI generates its feedback is therefore central to maintaining confidence and effective integration into everyday teaching.

- ***Privacy:*** There are no severe consequences in this case, as the system does not store sensitive student data.
- ***Fairness:*** The AI Learning Companion must operate fairly to ensure that all students receive equitable guidance and support during the learning process. Although no severe consequences are expected if fairness issues arise, since teachers supervise the system closely and it is not used for grading, any perceived bias in the Companion's feedback could cause teachers to lose trust and discontinue its use. Fairness is especially critical in educational contexts, where unequal treatment can undermine students' confidence and distort learning opportunities. If teachers detect biased patterns in how the AI responds to different groups of students, they are likely to view the system as unreliable and choose not to integrate it into daily classroom practice. Ensuring fair, consistent, and unbiased interactions is therefore essential for maintaining trust and supporting inclusive learning experiences.
- ***Transparency and accountability:*** Transparency and accountability are essential for the AI Learning Companion to function effectively in educational environments. Although no severe consequences are expected if transparency is limited, teachers must clearly understand how the system generates feedback, what data it uses, and how it supports students' problem-solving processes. This understanding strengthens trust and enables teachers to verify, explain, or override the Companion's suggestions when necessary. If the system lacks transparency, educators may doubt the reliability of its guidance, leading to reduced trust and reluctance to adopt or continue using it in the classroom. Such uncertainty can undermine the educational value of the tool and create ethical challenges, as teachers may struggle to explain AI-generated support to students. Ensuring transparent behavior and clear accountability therefore remains vital for sustaining teacher confidence and integrating the Companion responsibly into instructional practice.
- ***Validity and reliability***: The validity and reliability of the AI Learning Companion are essential to ensure that it supports the learning process consistently and accurately. Although no serious consequences are expected, since the system has already been used by real students and has demonstrated strong performance, it must remain robust to function effectively in classroom settings. Because it is used by groups of students in real time, it is crucial that the system does not freeze, provides accurate guidance, and

remains stable even under unusual or unpredictable conditions. If the Learning Companion is not robust, disruptions may occur during learning activities, leading to frustration among students and teachers. Such instability can generate negative feedback from schools and may result in decreased trust and eventual discontinuation of the system's use. Ensuring high reliability is therefore fundamental for maintaining smooth learning experiences and sustaining confidence in AI-supported learning tools.

4.3.2 Student Assessment

- ***Safety***: The Student Assessment model analyzes student responses and behavior patterns across the phases of inquiry-based learning to estimate their level of problem-solving competence. It is designed exclusively as a formative support tool that helps teachers identify student needs. It does not issue grades or make summative judgments. The assessment process relies heavily on teacher oversight, ensuring that data-driven insights support rather than replace professional judgment. If safe operation is compromised, teachers may lose trust in the accuracy and fairness of the model, leading them to abandon its use. Business risks include concerns that automated assessment could be misinterpreted as a threat to teacher authority, while unsafe or misleading outputs could negatively affect classroom decision-making. Ethical risks involve the possibility of students receiving incorrect feedback, which may distort their learning trajectory, and a broader loss of trust in AI-enhanced assessment practices. Ensuring the model operates safely, reliably, and transparently is therefore essential to sustaining its credibility and educational value.
- ***Security and resilience***: The Student Assessment model operates without storing sensitive student data and is used only to provide formative insights rather than formal grades, which significantly reduces the incentive for malicious interference. For this reason, no severe consequences are expected from typical security threats. However, if the integrity of the assessment model were compromised, for instance, if unauthorized changes affected how student performance is interpreted, users could lose trust in the system's reliability. Security and resilience therefore remain fundamental to ensuring that teachers feel confident in the accuracy and stability of the model's outputs. If persistent vulnerabilities were present, educators might distance themselves from using the assessment component, leading to reduced adoption by schools. Ethical risks would also surface, as compromised or untrusted assessment results could weaken confidence in AI-enhanced evaluation tools across the education sector.
- ***Explainability and interpretability***: The Student Assessment model relies heavily on interpretability, as teachers must understand why the system assigns particular competence levels or identifies certain problem-solving patterns. While the model does not produce high-stakes evaluations such as grades, and thus poses no severe consequences, explainability is crucial to help teachers trust and meaningfully use the insights it produces. When educators understand the basis of the AI's assessment, they can better support students, explain the results, or correct inaccuracies. If the system becomes opaque or its reasoning cannot be understood, teachers may question its

reliability and feel reluctant to incorporate its assessments into their instructional planning. This loss of trust could decrease the system's classroom usage and limit the perceived value of AI-supported formative assessment, even if no direct harm occurs. Interpretability is therefore a core requirement for ensuring that the assessment model functions as a trustworthy and effective instructional support tool.

- ***Privacy:*** The same that holds for the AI Learning Companion.
- ***Fairness:*** Fairness is particularly important for the Student Assessment model, which estimates students' problem-solving skills based on their interactions. While the system does not assign grades or make high-stakes decisions and thus poses no severe consequences in the traditional sense, any form of biased assessment could quickly erode teachers' trust. If teachers perceive that the model evaluates certain students unfairly, they will naturally question its usefulness and may stop relying on it altogether. This loss of trust represents a significant business risk, as adoption and sustained use depend on the system being viewed as impartial and reliable. Ethically, biased assessments could misinform students about their abilities and undermine confidence in AI-supported learning tools, making fairness a foundational requirement for meaningful and equitable educational support.
- ***Transparency and accountability:*** The Student Assessment model relies heavily on transparency to ensure that teachers understand how students' problem-solving skills are evaluated and what data informs the system's conclusions. While no severe consequences arise from a lack of transparency, given that teachers supervise the process and the system does not determine grades, clarity is still crucial for maintaining user trust. Educators must be able to explain the reasoning behind AI-generated assessments and feel confident that the model's outputs are grounded in understandable logic. If the system becomes opaque, educators may hesitate to rely on it, and schools may be reluctant to incorporate it into learning programs. This erosion of trust also carries ethical risks, as teachers would be unable to justify or interpret the assessment results, weakening confidence in AI-supported learning tools. Transparency and accountability are therefore indispensable for ensuring that the assessment model contributes positively and responsibly to teaching and learning.
- ***Validity and reliability***: The Student Assessment model depends on its validity and reliability to produce accurate interpretations of students' problem-solving skills. While no serious consequences are anticipated—given the model's strong performance in real-world trials and the fact that teachers supervise all outputs—robust functioning remains vital. The system must consistently produce stable and trustworthy results, avoid technical failures, and maintain accuracy even in typical learning scenarios. If the model becomes unreliable or unstable, the learning process may be disrupted, causing teachers to question the accuracy of the assessments and diminishing their willingness to use the tool. Negative experiences of this kind can result in poor feedback from schools and reluctance to adopt the system, despite its otherwise promising benefits. Reliability is

therefore a key factor in ensuring that the assessment model remains a dependable and credible support tool for educators.

4.3.3 AI Learning Recommendations

- ***Safety***: The AI Learning Recommendations system generates personalized learning pathways and suggests additional activities based on each student's performance and interaction history. These recommendations are never delivered directly to students without teacher approval, reinforcing that educators maintain full pedagogical control. Like the other models, this system is designed to support, not replace, teacher decision-making, and its safe functioning depends on continuous human oversight. If the system's safety is compromised, teachers may doubt the relevance, fairness, or appropriateness of its suggestions, resulting in a loss of trust and eventual discontinuation of its use. Business risks include educator concerns that automated recommendations might encroach on their instructional autonomy or reduce their role in lesson planning, while over-reliance on the system could diminish teacher involvement and negatively affect educational outcomes.
- ***Security and resilience***: The AI Learning Recommendations system, like the other models, does not handle sensitive student data and functions only as a supportive tool that proposes personalized learning pathways rather than making formal academic decisions. This greatly limits the consequences and attractiveness of potential attacks. Nonetheless, maintaining strong security and operational resilience is essential, as a compromised recommendation engine could generate inappropriate or unreliable suggestions, undermining teacher confidence. If educators perceive the system as vulnerable or easily manipulated, they may stop using it, resulting in reduced adoption of AI-based personalized learning tools in schools. Beyond business risks, ethical concerns would arise as trust in AI-driven educational support systems deteriorates, potentially impacting the willingness of schools to engage with innovative technologies that could otherwise benefit teaching and learning.
- ***Explainability and interpretability***: The AI Learning Recommendations system must provide clear and interpretable rationales for the personalized pathways and activity suggestions it generates. Although the system poses no severe consequences, given that teachers ultimately approve or reject every recommendation, explainability is essential for maintaining confidence in the tool. Teachers need to understand why a specific activity or learning path is recommended so they can justify it to students, adapt it if necessary, or override it when it does not align with pedagogical goals. If the system's recommendations are not interpretable, educators may distrust the underlying logic, hesitate to rely on the suggestions, and struggle to diagnose errors or unexpected outputs. This could lead schools to be cautious or unwilling to integrate the recommendation engine into everyday practice. As a result, ensuring transparent and understandable AI behavior is fundamental for supporting teacher decision-making and sustaining institutional trust in personalized learning technologies.

- ***Privacy:*** The same that holds for the AI Learning Companion and the AI Learning recommendations.
- ***Fairness:*** The AI Learning Recommendations system must provide unbiased and equitable learning pathways to ensure all students receive appropriate and fair opportunities for support and enrichment. Although the system does not influence grading and teachers ultimately approve all recommendations, limiting the severity of direct consequences, any bias in the learning suggestions can still damage trust. If educators observe unfair patterns in how recommendations are generated, they may lose confidence in the system and choose not to integrate it into their instruction. Such loss of trust poses clear business risks by reducing adoption and limiting the system's impact. Moreover, ethical risks are significant: unfair recommendations may misinform students, restrict learning opportunities, or contribute to the erosion of trust in AI across educational contexts. For these reasons, fairness is not only desirable but essential to the responsible and effective use of personalized learning technologies.
- ***Transparency and accountability:*** The AI Learning Recommendations system must be transparent so that teachers understand how personalized pathways and activity suggestions are generated, as well as the data sources behind these recommendations. Although the absence of transparency does not lead to severe consequences, since teachers review all suggestions before they reach students, it can significantly weaken trust in the system. Educators who cannot interpret or justify the origin of a recommendation are less likely to rely on the tool and may avoid using it altogether. This reluctance poses business risks for adoption and continued usage in schools and raises ethical concerns, as educators may be unable to explain AI-generated pathways to students or ensure their appropriateness. Strong transparency and accountability mechanisms therefore play a central role in maintaining trust and ensuring that the recommendations system supports rather than complicates the teaching process.
- ***Validity and reliability***: The AI Learning Recommendations system must be valid and reliable to ensure that the personalized pathways it generates are appropriate, accurate, and pedagogically meaningful. Although no severe consequences are expected—since the system has been tested with real students and teachers reviewing all recommendations, the robustness of the system remains essential. It must generate meaningful recommendations even when faced with unusual patterns of student interaction. If inaccurate recommendations are produced, the learning process may be interrupted, causing teachers to lose confidence in the tool. Such disruptions can lead to negative perceptions from schools and may ultimately result in reduced adoption. For these reasons, maintaining strong accuracy and ensuring the system's stability are vital to preserving trust and supporting effective personalized learning.

If the identified risks were to materialize, they could significantly affect the educational value and trustworthiness of the AI system. For example, biased personalized learning recommendations could result in learners receiving uneven or inappropriate instructional content. This could hinder

student engagement, distort their academic development, and introduce inequities in the learning process.

Such risks would also directly influence teachers' confidence in using the system. Educators rely on AI-supported tools to complement instruction, monitor progress, and support differentiated learning. If the system delivers inconsistent, biased, or unreliable recommendations, teachers may question its pedagogical validity and feel uncertain about integrating it into their classroom practice. This hesitation could lead to reduced system adoption, additional workload for teachers who must verify outputs, and concerns about classroom fairness.

Beyond the classroom, unresolved risks could impact on the broader educational ecosystem and the project's long-term sustainability. Institutions may face challenges related to transparency, accountability, and compliance with ethical AI principles. Failure to adequately address these issues could undermine institutional credibility, limit the system's scalability, and expose stakeholders to reputational or operational challenges. Ultimately, ensuring that risks are understood and mitigated is essential for maintaining trust among students, teachers, and all educational stakeholders.

3.1.4 Vulnerability assessment

The Vulnerability assessment phase is aimed at defining and describing the vulnerabilities of the Educational LSP AI Solution that may lead to the actualization of the identified threats. Vulnerabilities are measured on a technical and human basis, and at all AI lifecycle phases.

Technical vulnerabilities assessed include, but are not limited to:

- Software flaws and weaknesses in the AI model implementation or inference pipeline.
- Inadequate data governance practices, including insufficient controls over training data quality, provenance, and access.
- Network and infrastructure weaknesses that could expose AI assets to unauthorized access or manipulation.
- Insufficient monitoring and logging mechanisms that reduce the detectability of attacks or anomalous behavior.

Human vulnerabilities assessed include:

- Low Trustworthy AI maturity of developers, educators, and other participants involved in the design, development, and operation of the Educational LSP AI Solution, as measured by the TrustSense tool during the Cartography phase.

In each of the identified threats in Phase 2, the vulnerability of all AI components in both lifecycle stages was determined by TrustGuard, and the level of vulnerability was estimated and documented. Identified vulnerabilities along with related threats for the Educational LSP are presented in Figure 10.

3.1.5 Risk assessment

During the Risk assessment phase, TrustGuard's integrated risk engine calculated composite risk scores for each AI asset of the Educational LSP AI Solution, synthesizing the outputs of the preceding four phases into an overall risk profile.

The risks of all AI assets of the AI system will be estimated in terms of:

- Threat Level: The frequency of occurrence of a threat (Phase 2).
- Vulnerability Level: Determine the system's susceptibility based on control measures (Phase 4).
- Consequence Level: Consider the potential consequences of the threat to all dimensions of trustworthiness (Phase 3).
- Trustworthy AI maturity of teams, trustworthiness of AI participants and sophistication of potential adversaries (Phase 1).

The combination of these inputs within TrustGuard produced risk ratings per threat, per AI component, and per lifecycle stage, yielding a comprehensive and granular risk profile for each AI asset of the Educational LSP AI Solution. The results of the risk assessment are presented in Section 3.5.

3.1.6 Risk management

This phase encompasses the proposal, selection, and initial implementation of appropriate strategies to mitigate the risks identified during Phase 5. This phase also includes an initial review of the outputs of Phases 1 through 5 for human-centered validation of findings, providing an opportunity for the LSP3 team to verify the coherence and completeness of the assessment before committing to mitigation strategies.

The risk assessor using TrustGuard:

- Examined the consolidated risk results of Phases 1-5 and performed a preliminary human-focused validation of the threat, vulnerability, consequence and risk results with the concerned stakeholders
- Proposed a list of candidate control measures that would deal with the identified risks on the technical, organizational, and human levels
- Chose the best controls that were in terms of minimizing risks, their feasibility and suitability in the context of operations of the Educational LSP
- Documented the status of controls implementation, differentiating between those controls that are already in place, those that are going to be implemented, and those that are going to be evaluated further.

The output of this stage is a list of selected controls and their justification, a list of implemented controls and controls that will be implemented, and preliminary testing and evaluation reports of controls that are implemented. The results of this phase can be found in Section 3.5.

3.2 Tool Validation Planning & Training Processes

3.2.1 Dry-run

In the dry run execution of the deployment phase of the Educational LSP AI solution, the LSP team carefully filled the assessment spreadsheet, beginning with the Cartography of the system. The LSP team characterized the System Under Test (SUT) as the Educational LSP AI solution, which is an application that evaluates the problem-solving skills of students. Figure 12 presents the cartography phase of the Educational LSP AI solution. It consists of three models and two datasets. All components are currently in the deployment stage.

CARTOGRAPHY	*values with the ones that fit for your AI System. If need Insert new lines to add more AI Assets*			
	ANSWER	Comment		
AI SYSTEM NAME	Educational LSP AI Solution			
SECTOR	EDUCATION			
LIFECYCLE STAGE	DEPLOY			
AVG score of all stakeholders' technical proficiency (ONLY IF LIFECYCLE STAGE=DEPLOY) (please visit D2.1 for this)				
AI ASSET TYPE*	**AI ASSET NAME**	**LIFECYCLE STAGE**	**TEAM OWNER per AI Asset (only if lifecycle stage= DESIGN/DEVELOP)**	**AI MATURITY FOR TEAM (please visit D2.1 for this)**
AI ASSET -Models *(insert below more lines if necessary)*	M1- AI Learning Companion Model (RASA NLU)	DEPLOY		
AI ASSET -Models (insert below more lines if necessary)	M2- AI Student Classifier	DEPLOY		
AI ASSET -Datasets *(insert more lines if necessary)*	Dt1- Conversation Dataset	DEPLOY		
AI ASSET -Datasets *(insert more lines if necessary)*	Dt2- Interaction Dataset	DEPLOY		
AI ASSET -Datasets *(insert more lines if necessary)*				

Figure 12: Cartography of the Educational LSP AI solution.

Following the architectural mapping, the team documented the system's risk assessment in the Criticality Level sheet. The criticality level of the Educational LSP AI Solution is a high-risk AI-system as established by Article 6 of the AI Act. The LSP team defended this grading by highlighting the fact that the system measures the problem-solving performance of students, which is a critical operation that directly affects the progress of students in education and as such, reliability, fairness and accuracy are important requirements. Figure 13 shows the Criticality assessment of the Educational LSP AI solution.

CRITICALITY ASSESMENT			
Questions	Useful resources	Answer	Comment
Does the AI system falls under the qualification of a high risk AI-system as established by Article 6 of the AI Act, such as the high risks systems listed in AI Act Annex III If the AI system poses a very high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.	Annex III	Yes	
Does the AI system pose a high risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making? and /or Is the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an essential service as defined in Annex I of NIS 2 Directive?	Annex I of NIS 2 Directive	No	
Does the AI system pose substantial risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making If the AI system is used by an Operator of Essential services (OES) and the AI system is used for the provision of an important service as defined in Annex 2 of NIS2?	Annex 2 of NIS2	No	
Does the AI system pose medium risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making? Is the AI system for the provision of an important service as defined in Annex 2 of NIS2?		No	
Does the AI system pose low or no risk of harm to the health, safety or fundamental rights of natural persons, including influencing the outcome of decision making?		No	

Figure 13: Criticality assessment of the Educational LSP AI solution

The Business Consequence Level sheet further estimated the potential effects of system failures. One of the main negative effects that the pilot team marked as having a high severity level entailed the mis grading of students. The rationale given underscored the fact that improper grading would be a source of unfair treatment that will not only have negative impacts on the academic records of students but also their psychological health. Figure 14 illustrates the business consequences of the lack of trustworthiness in the Educational LSP AI Solution.

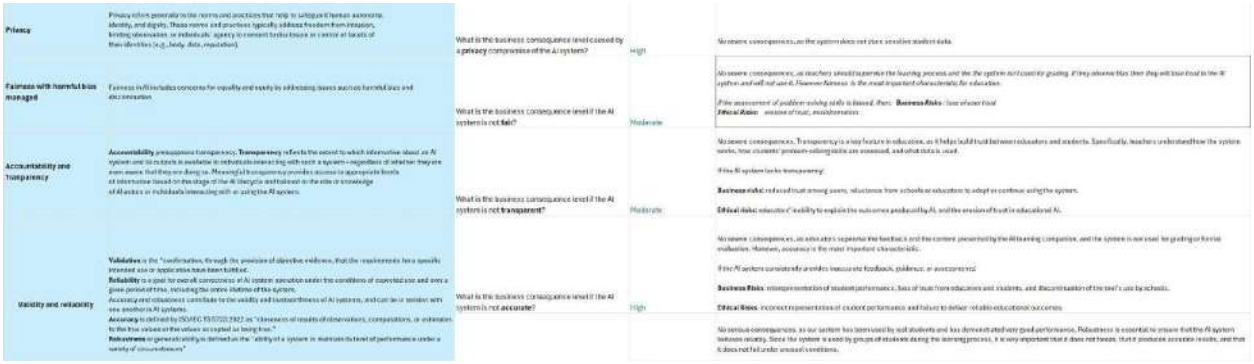


Figure 14: Business consequences of the lack of trustworthiness in the Educational LSP AI Solution

Lastly, the practical implementation of the assessment methodology was confirmed by the Control Questions sheet adapted to the Deploy stage. The team has answered specific questions about operational safeguards systematically, which means that they have been implemented, and their answers have special evidence, including system logs and monitoring reports, to demonstrate compliance. Figure 15 displays the control questions for the deployment phase of the Educational LSP AI solution.

Lifecycle Stage	Other threats relevant to the same vulnerability	Vulnerability (from Golden Table)	Control	Relevant AI Asset for the control question (Replace with the AI Assets of your SUT)	Template Control Questions (Please replace "Dtr", "M1", "M2" with the names of the AI assets declared in the cartography phase. Add more/remove lines if needed)	Answer
Deploy (Production)	Denial of service due to inconsistent data or a sponge example	Lack of consideration of real-life conditions in training the model	Ensure that the model is sufficiently resilient to the environment in which it will operate.	M1	In the context of M1, have you ensured that the model is sufficiently resilient to the environment in which it will operate?	Yes
				M2	In the context of M1, have you ensured that the model is sufficiently resilient to the environment in which it will operate?	Yes
	Failure or malfunction of ML application	ML application not integrated in the cyber-resilience strategy	Integrate ML applications into the overall cyber-resilience strategy	M1	Has M1 been integrated into the overall cyber-resilience strategy?	Yes
				M2	Has M2 been integrated into the overall cyber-resilience strategy?	Yes
	Failure or malfunction of ML application	Existence of unidentified failure scenarios	Conduct a risk analysis of the ML application	M1	Have you identified all possible failure scenarios for M1?	No
				M2	Have you identified all possible failure scenarios for M2?	No
	Failure or malfunction of ML application, Compromise of ML application components, Poisoning	Undefined indicators of proper functioning, making complex malfunction identification	Define and monitor indicators for proper functioning of the model	M1	In the context of M1, have there been defined indicators of proper functioning of the model?	Yes
				M2	In the context of M2, have there been defined indicators of proper functioning of the model?	Yes
	Failure or malfunction of ML application	Lack of explainability and traceability of decisions taken	Build explainable models	M1	Is the M1 model explainable?	Yes
				M2	Is the M2 model explainable?	No
	Failure or malfunction of ML application	Lack of security process to maintain a good security level of the components of the ML application	Implement processes to maintain security levels of ML components over time		NA	
	Failure or malfunction of ML application	Existence of several vulnerabilities because ML specificities are not integrated in existing policies	Ensure ML projects follow the global process for integrating security into projects	M1	In the context of M1, have it been ensured that the model has followed the global process for integrating security?	Yes
				M2	In the context of M2, have it been ensured that the model has followed the global process for integrating security?	Yes
	Compromise of ML application components, Failure or malfunction of ML application	Contract with a low security third party	Ensure ML applications comply with third parties' security requirements	M1	Have you ensured that the M1 model complies with third parties' security	Yes
				M2	Have you ensured that the M2 model complies with third parties' security requirements?	Yes
	Failure or malfunction of ML application	Application not compliant with applicable regulations	Assess the regulations and laws the ML application must comply with	M1	Have you assessed the regulations and laws the M1 model must comply with?	Yes
				M2	Have you assessed the regulations and laws the M2 model must comply with?	Yes
	Oracle, Model or data disclosure	The model allows private information to be retrieved	Ensure that models respect differential privacy	M1	Does the M1 model respect differential privacy?	Yes
				M2	Does the M2 model respect differential privacy?	Yes
	Oracle,			M1	*In the context of M1, have the output information given by the model been reduced*	No

Figure 15:Control questions for the deployment phase of the Educational LSP AI solution.

Importantly, the participants used the columns of feedback to review the assessment tool. The LSP team provided comments related to the questions. Through its involvement of these validation metrics, the LSP3 team offered the required information to make sure that TrustGuard questionnaire can be applied practically, understood, and mostly applicable to end-users who run deployed AI systems in the field of education.

3.2.2 Plan development

The Detailed implementation plan of the Education LSP contains the activities planned in the Full Deployment stage, which will take place between Month 23 and Month 36. The implementation of this plan is mainly motivated by ICCS and EA. The roadmap will be based on three main stages, which will include preliminary planning, general training, and actual application of the FAITH tools. The detailed implementation plan of the Education LSP is presented in figure 16.

The planning preparatory works have begun with the specification and formal agreement of the implementation roadmap. The ICCS and EA teams completed these initial tasks successfully in Month 25 which created a clear direction towards the next stages.

To make sure that the tools deployed were used effectively, a set of comprehensive training sessions were organized among the respective teams. The ICCS & EA team already attended the trustSENSE tool training in Month 23 and the trustGUARD training on Month 26. Moreover, the AI Model Passport and AI Data Passport training sessions were completed successfully. The team will be trained on the SSM tool later in the year in Months 32 and 33, and the team will also have an introduction to the FAITH STM Platform registration and workflows.

Detailed Implementation Plan LSP3 - Education	Phase	Deployme	Pis delete 1:		FULL DEPLOYMENT				*NOTE: You will need to answer trustGUARD questions as if the SUT is deployed in reality (not in the sandbo									
ICCS & EA																		
	Project Start:		M23				26-Jan	26-Feb	26-Mar	26-Apr	26-May	26-Jun	26-Jul	26-Aug	26-Sep	26-Oct	26-Nov	26-Dec
	Display Month:		1		M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36
TASK	ASSIGNED TO	PROGRES S	START	END														
Planning																		
Define plan	ICCS & EA	100%	M25	M25														
Agree plan	ICCS & EA	100%	M25	M25														
Training and support																		
Attend training for AI Model Passport (pythonic library)		100%																
Attend training for AI Data Passport		100%																
Attend training for trustSENSE	ICCS & EA	100%	M23	M23														
Attend training for trustGUARD	ICCS & EA	100%	M26	M26														
Attend training for SSM	ICCS & EA		M32	M33														
Attend introduction to FAITH STM Platform registration and workflows	ICCS & EA		M33	M33														
FAITH Tools Implementation and Iteration 2026																		
Deploy SUT	ICCS & EA	100%	M6	M20														
Use trustSENSE for team maturity assessment (early deployment period)	Risk Assessor + End Users	100%	M25	M25														
Provide formal feedback on Usability and Usefulness of trustSENSE	Risk Assessor	100%	M25	M25														
Consider trustSENSE mitigations	ICCS & EA	100%	M26	M27														
Use trustGUARD v1 for in-isolation AI Model Asset assessment @ deployment phase	Risk Assessor	100%	M28	M28														
Use trustGUARD v1 for in-isolation (input) Data Asset assessment @ deployment phase	Risk Assessor	100%	M28	M28														
Provide formal standardised feedback on Usability and Usefulness of trustGUARD v1 during deployment phase	Risk Assessor	100%	M29	M29														
Apply 1 or more Controls to AI Model Asset or Data Asset during deployment phase based on trustGUARD recommendations	ICCS & EA	100%	M29	M32														

Figure 16:Detailed Implementation Plan of the Educational LSP

The implementation and iteration of FAITH tools are aimed at the practical implementation, evaluation, and improvement of the FAITH tools. After the initial deployment of the System Under Test (SUT) in Month 5, the trustSENSE tool was used on Month 25 to evaluate team maturity of the end users and AI developers. At the same time, its usability and usefulness were also given formal feedback. During Months 26 and 27, the ICCS & EA team analyzed and considered the mitigation strategies suggested by the trustSENSE. Month 28, the Risk Assessor used trustGUARD v1 to make in-isolation the AI Model Asset and the Input Data Asset. In response, formal feedback on the usability of the tool was provided in Month 29 in a standardized format. According to the recommendations provided by trustGUARD the ICCS & EA team implemented specific controls in the AI Model and Data Assets in Month 29, which will be observed and revised until Month 32.

ATC will handle ongoing iteration of the trustGUARD v2, which will include constant Knowledge base updates according to LSP feedback within all the three lifecycle phases between Month 30 and Month 36. With the trustGUARD activities, Month 33 will also be the time to introduce the FAITH STM Platform and the SSM tool. The FAITH STM Platform will be used by the Risk Assessor and will also create standardized formal feedback where ICCS and UOS will beta test the platform between Month 34 and Month 36. The collaboration between ICCS, EA, and UOS that will produce the SSM system model will occur within the same period of Month 33. The Risk Assessor will in turn use the SSM to make an in-context SUT evaluation between Months 33 and 35 and give formal feedback on completion on Month 35. This feedback will drive the continuous development of SSM v2 by UOS, which will begin on Month 36 and continue to Month 45. The result of all these thorough assessment and refinement processes will culminate in the implementation of the SUT in the Replication Pilot, which is one of the major milestones that must be implemented by ICCS and EA Months 36 to 38.

3.2.3 Tool Training

To evaluate the trustworthiness of the Educational LSP AI Solution, the risk assessor and the LSP3 team participated in training workshops aimed at learning how to use the FAITH-AI-TAF tools to conduct the assessment. Table 3 provides an overview of the training workshops for FAITH tools that the LSP3 team participated in.

Workshop	**Workshop date**	**Participants**
TrustSense training	10 November 2025	WP2&LSP partners
Data & AI model Passport training	30 January 2026	FORTH&ICCS&EA
TrustGUARD training	19 February 2026	ATC&ICCS

Table 3: Overview of the training workshops attended by the LSP3 team

First, the LSP3 team participated in the TrustSense training workshop in November. The workshop with title “trustSense -Trustworthiness Assessment of AI Participants Teams” was organized by trustilio. The main purpose of the workshop was to train the LSPs on how to use the trustSense tool for assessing the trustworthiness of AI participant teams. During the workshop, Trustilio presented the TrustSense tool and explained how all LSPs could use it with the AI participants to assess the AI maturity of the end users and developers. At the end of the workshop, all participants used TrustSense and were asked to provide their feedback. As a result, the LSP3 team was trained to organize similar workshops to assess the AI maturity of the end users and developers of the Educational LSP AI Solution.

Additionally, the ICCS and EA team participated in the Data & AI Model Passport training workshop on January 30, 2026. After that, the LSP3 team was able to create datasets and AI models of the Educational LSP AI Solution in the FAITH AI Model Hub.

Finally, the ICCS team participated in TrustGUARD training on 19 February. As a result, the risk assessor was trained to use the TrustGUARD tool to conduct the cartography, threat assessment, and consequence and vulnerability assessment for the AI models of the Educational LSP AI solution.

3.3 FAITH Tool Validation Activities

3.3.1 TrustSense

The TrustSense tool used by the LSP3 team to assess the AI maturity of the AI participants of the Educational LSP AI Solution. The LSP3 team organized two workshops for the assessment of AI

maturity of teachers and developers. The first workshop, aimed at assessing the AI maturity level of the development team using TrustSense, took place on the 13th of January 2026 with the developers who implemented the Educational LSP AI Solution. Developers were introduced to the goals and process of the workshop, used the TrustSense tool by answering its questions, and subsequently provided feedback through a questionnaire. The trustworthiness level result of the technical team is very high, with a calculated score of 4.9, reflecting consistently high adherence to ethical standards, collaboration, and adaptability. Figure 17 presents the trustworthiness assessment results of the AI Technical Team (developers) as evaluated through the trustSense tool.

The TrustSense tool generated detailed mitigation actions per category, providing targeted recommendations and controls to sustain and further strengthen the team's trustworthiness profile. Most categories were rated at a high level. Only two categories were identified at a moderate level, resilience and resource accessibility. Overall, the mitigation actions recommended by the TrustSense tool reflect a comprehensive and structured approach to maintaining and improving trustworthiness within the developer team, with particular attention to the areas of resilience and resource accessibility where targeted intervention is advised.

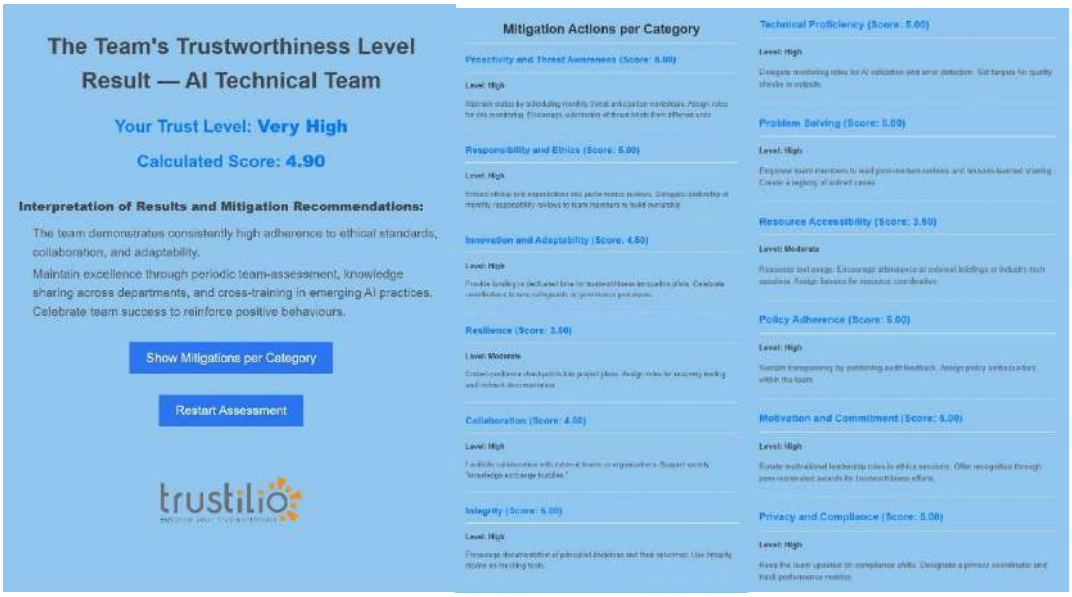


Figure 17: Trustworthiness level result of the technical team.

The second workshop, “AI Maturity Assessment for Teachers,” was conducted at EA on 14 January with the aim of evaluating teachers’ maturity, confidence, and readiness in working with AI-enhanced educational tools. The workshop brought together six teachers, from different STEM disciplines. All participating teachers had prior experience using the Educational LSP platform, ensuring that the discussions and assessments were grounded in real classroom interactions and hands-on familiarity with the system.

During the workshop, ICCS team presented the Educational LSP platform, outlined the workshop objectives, and introduced the TrustSense tool. After the presentation, teachers were invited to complete the Trust Sense tool, which includes a set of questions designed to gauge their understanding, trust levels, and perceived risks related to AI-supported learning. Their responses were collected, and individual scores were recorded to support a more detailed analysis of their current level of AI maturity. This allowed the team to identify trends, strengths, and potential areas where further guidance or training may be needed. The trustworthiness level of the teachers is very high. Figure 18 presents the trustworthiness assessment results of the Domain AI User Team (teachers) as evaluated through the trustSense tool. The team achieved a Trust Level of Very High, with a calculated score of 4.65, reflecting consistently high adherence to ethical standards, collaboration, and adaptability.

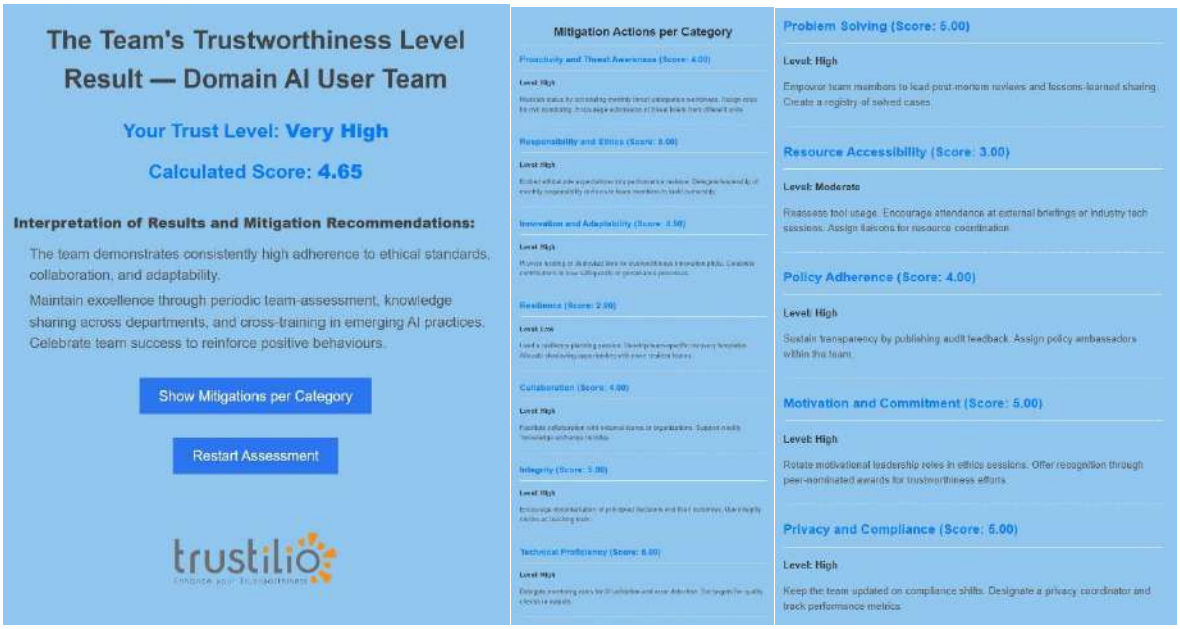


Figure 18: Trustworthiness level result of the teachers

The TrustSense tool generated detailed mitigation actions per category, providing targeted recommendations and controls to sustain and further strengthen the teaching team's trustworthiness profile. Most categories were rated at a high level. Two categories were identified as requiring more targeted attention: resource accessibility and resilience. The recommended actions include leading a resilience planning session, developing team-specific recovery templates, and allocating shadowing opportunities with more resilient teams. Overall, the mitigation actions recommended by the TrustSense tool for the teaching team reflect a structured and comprehensive approach to maintaining high trustworthiness standards, with particular emphasis on strengthening resilience and improving resource accessibility as the two areas requiring prioritized intervention.

3.3.2 Data & AI model Passport in the FAITH AI Model Hub

The FAITH AI Model Hub was used by the LSP3 team to create Data & AI Model Passports for all AI models and datasets belonging to the Education LSP. The Data & AI Model Passports creation was conducted to identify all the AI assets of the Education LSP AI Solution during the system cartography.

The datasets registered and documented in the FAITH AI Model Hub are the Learning Material Dataset and the Student Interaction Dataset. The Learning Material Dataset consists of all the educational scenarios created by teachers, while the Student Interaction Dataset contains data on student interactions with the implemented scenarios on the platform. Figure 19 presents a screenshot of the datasets created for LSP3 using the FAITH AI Model Hub.

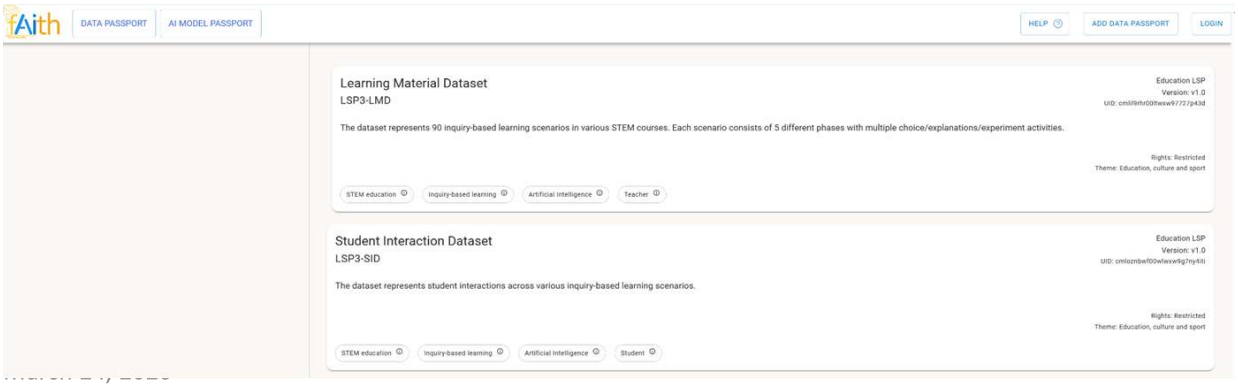


Figure 19: Datasets of the Educational LSP AI Solution created using the FAITH AI Model Hub

The AI Models registered and documented in the FAITH AI Model Hub were as follows:

- AI Learning Companion
- AI Student assessment
- AI Learning Recommendations

The LSP3 Team filled the Model Passport with useful metadata such as model architecture information, training data provenance, intended use context, and data sources, and governance information, in the case of each asset. Figure 20 shows a screenshot of the AI models created for LSP3 using the FAITH AI Model Hub.

The Model Hub is the focal point of lifecycle management as it is used to trace and hold everybody accountable throughout the development and deployment pathway. Formal documentation standard requirement is also met by the registration of models and datasets in the Model Hub, as it is ensured that the design of the model, sources of data, training process, and intended usage are properly reported.

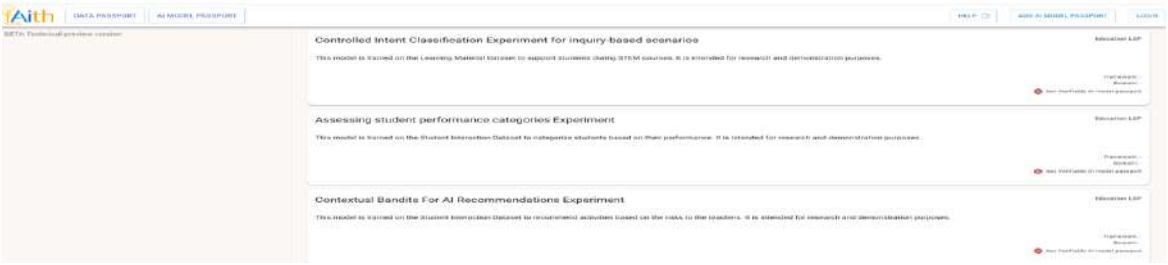


Figure 20: AI Models of the Educational LSP AI Solution created using the FAITH AI Model Hub

3.3.3 TrustGuard

TrustGuard has been used to perform the in-isolation trustworthiness assessment of the Education LSP AI Solution. TrustGuard evaluated the trustworthiness of all three of the AI assets, AI Learning Companion, Student assessment and AI Learning Recommendations for all the AI lifecycle phases (Design, development, deployment). Although the Educational LSP AI Solution is currently in the deployment phase, 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 results presented in the following sections focus on the deployment phase, as this reflects the current operational context of the solution and constitutes the most directly relevant evidence base for the ongoing pilot activities.



Figure 21: List of in-isolation trustworthiness assessment using the TrustGuard

The risk assessor conducted the in-isolation trustworthiness assessment using the TrustGuard tool for all the AI lifecycle phases. The process begins with the description of the AI model in the Cartography phase, followed by the completion of the assessments: (i) the Threat Assessment, covering the probability of occurrence of all applicable threat categories; (ii) the Consequence Assessment, evaluating the business impact levels across trustworthiness dimensions; and (iii) the Vulnerability Assessment, examining the level of current controls and protection. The results of the in-isolation trustworthiness assessment constitute the main evidence for risk assessment

results and countermeasure recommendations, which are reported below. Figure 21 presents the list of the assessments conducted for the Educational LSP AI Solution.

3.4. Threat Analysis

This section reports on the asset-level threats identified by TrustGuard across the three registered assets. Threats are organized by assets and include the assessed likelihood of occurrence together with the relevant justification provided during the assessment process.

3.4.1 Learning Companion

The Learning Companion model supports students by providing guidance and support on inquiry-based scenarios. The AI assets considered in the in-isolation assessment are the Learning Companion model and the learning materials dataset, which consists of inquiry-based scenarios created by teachers within the Educational LSP AI platform. The threat assessment was conducted across all phases of the AI lifecycle, namely design, development, and deployment. Figure 22 displays a screenshot of the threat assessment as filled in TrustGuard in the asset of Learning Companion to demonstrate the frequency choice of each of the applicable threat categories during the deployment phase.

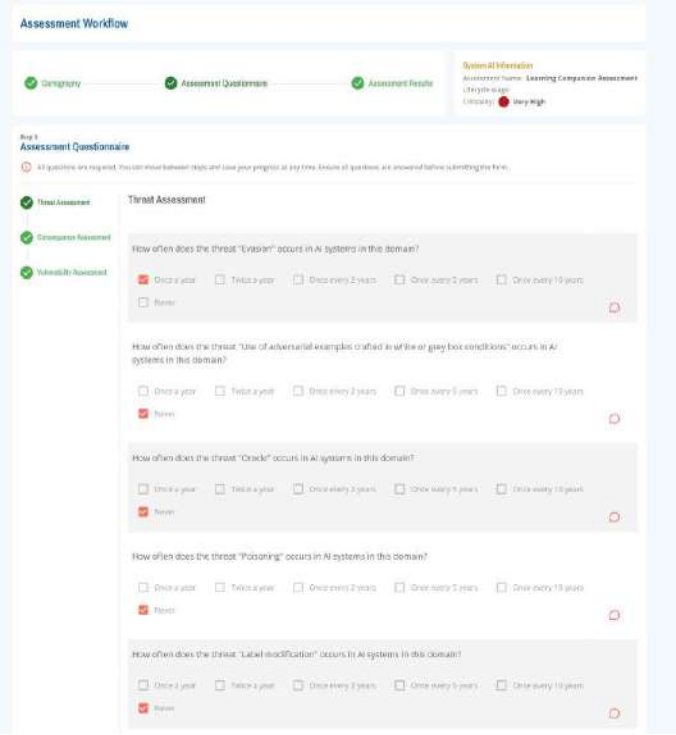


Figure 22: Threat Assessment for the Learning Companion

During the threat assessment, all potential threats were identified along with their frequency. The frequency was calculated based on both the likelihood of occurrence within the Educational LSP AI Solution and the observed occurrence of each threat since the system was deployed.

After completing the threat assessment, the risk assessor filled in the consequence assessment based on the input received from the educators regarding the potential consequences of the threat across all dimensions of trustworthiness. Figure 23 shows a screenshot of the completed consequence assessment in TrustGuard.

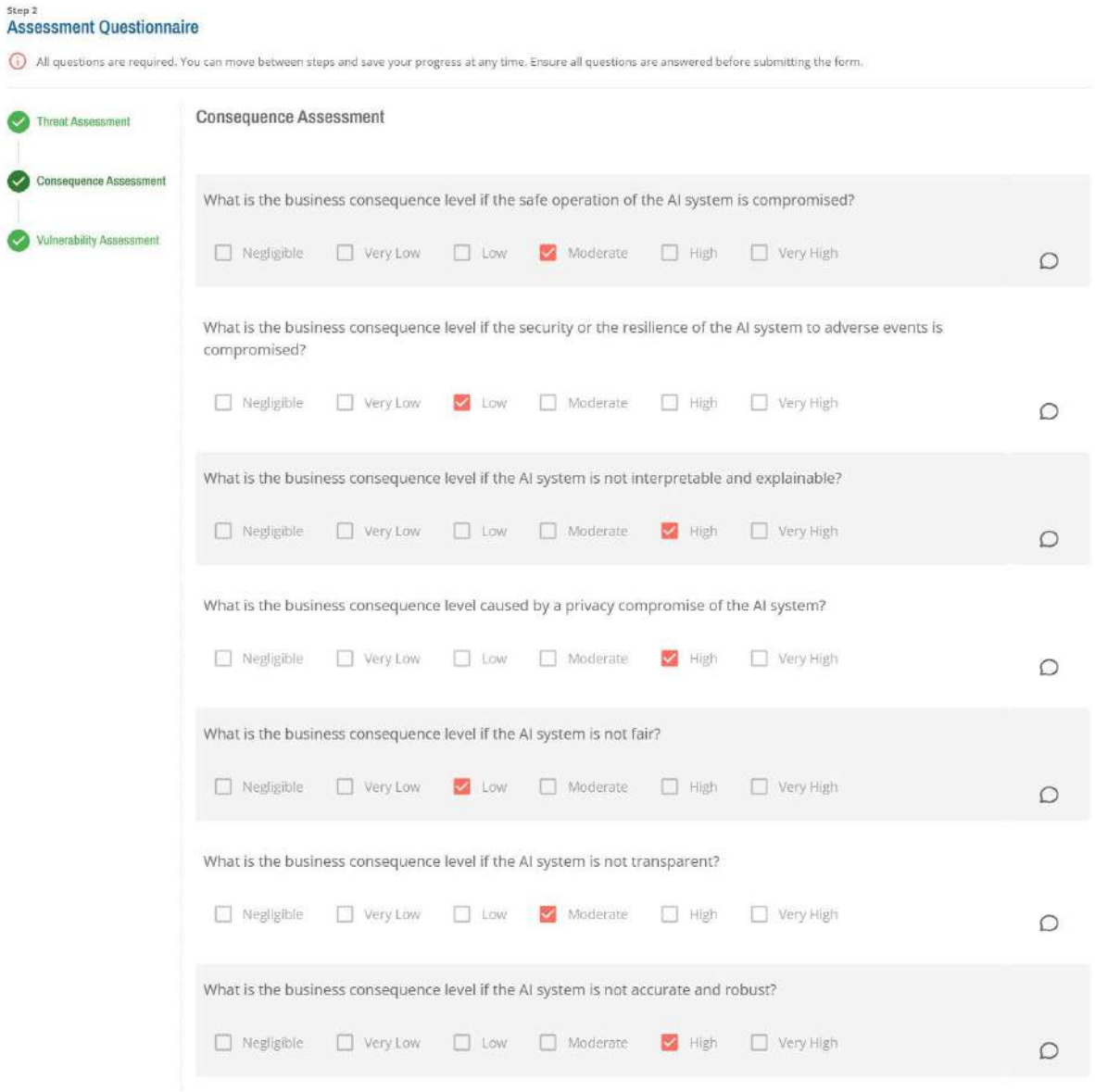


Figure 23:Consequence assessment for the Learning Companion

The next step is the completion of the vulnerability assessment for all models and datasets. The risk assessor answers questions related to the control measures that have been implemented. Figure 24 shows a screenshot of the completed vulnerability assessment in TrustGuard.

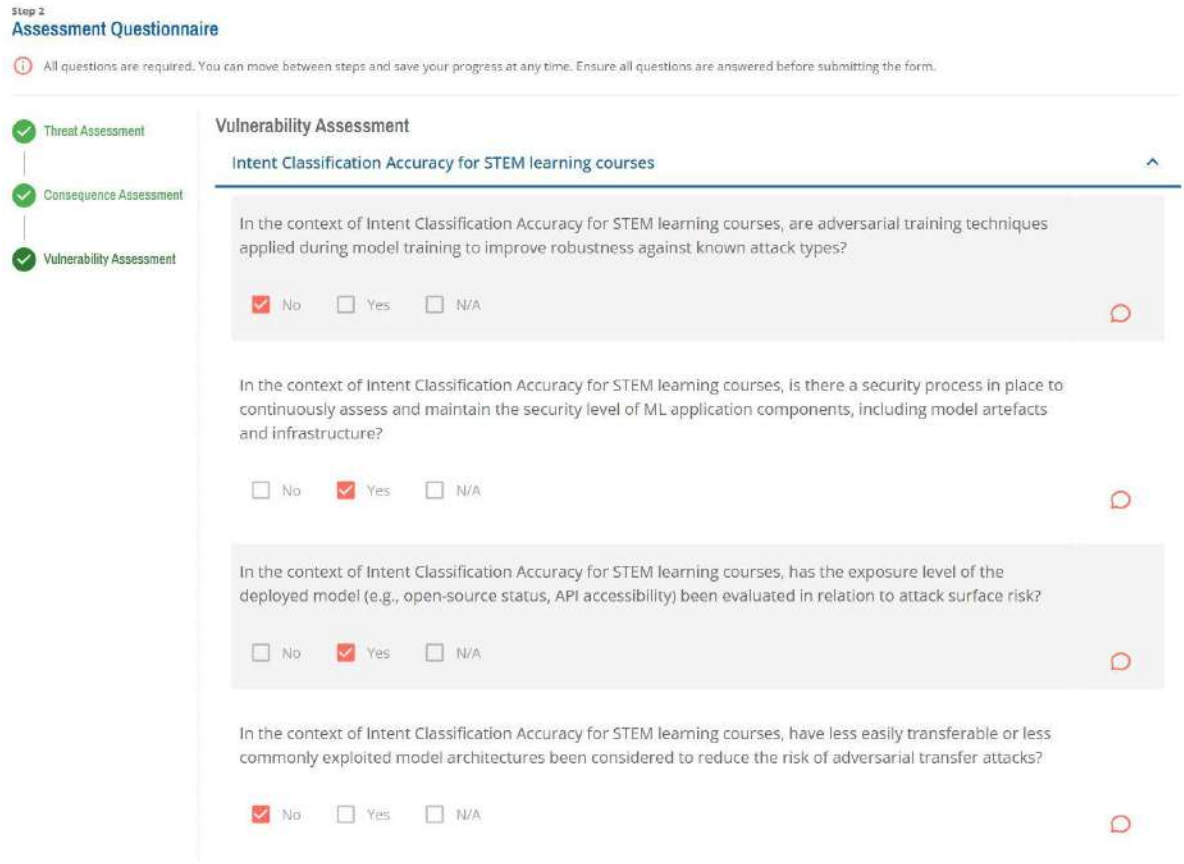


Figure 24: Vulnerability assessment for the Learning Companion

Table 4 presents all identified threats, along with their frequency and a corresponding explanation.

Threat	Frequency	Explanation
Evasion	Once a year	Students may employ response strategies that cause incorrect skill inference during normal use. This is treated as a recurring behavioural phenomenon rather than a discrete attack, warranting at least an annual frequency classification.
Failure or malfunction of ML application	Once a year	Occasional software bugs or unexpected runtime conditions may cause technical failures. Estimated at approximately once per year given the system's controlled operating environment.

Human error	Twice a year	Teachers or administrators may introduce errors during system setup, question uploads, or data handling. Estimated at approximately twice per year; errors are generally minor in scope.
Propagation of misinformation	Once a year	Occasional errors in AI-generated feedback or knowledge representation may lead to minor inaccuracies being conveyed to students. Estimated at once per year given testing procedures in place.
Model Drift Leading to Inaccurate Assessments	Once a year	Gradual shifts in student response patterns or experimental conditions could affect assessment accuracy over time. The risk is rare but plausible given the system's reliance on a consistent input distribution.
Poor quality of data used in training and testing	Once a year	Minor errors or inconsistencies may occasionally occur during dataset preparation or updates. A stable environment and teacher oversight limit frequency to approximately once per year.

Table 4: List of the identified threats, along with their frequency and a corresponding explanation for the Learning companion.

The table above lists all the threats that have a likelihood of occurring. Threats with a very low likelihood, characterized as ‘never,’ are not included in the table. Most threats were assessed as 'Never' occurring, reflecting the highly controlled and supervised nature of the classroom deployment. The threats related to Evasion, Failure or Malfunction, Human Error, Propagation of Misinformation, Model Drift and Poor Data Quality are due to operational aspects of any production AI system and are managed in the risk and countermeasures sections below.

3.4.2 Student Assessment

The AI Learning Recommendations model generates personalized learning pathways based on inquiry-based scenarios, using previous student interaction data. The AI assets considered in the in-isolation assessment are the AI Learning Recommendations model, the interaction dataset, and the learning materials dataset, which consists of inquiry-based scenarios created by teachers within the Educational LSP AI platform. The threat assessment was conducted across all phases of the AI lifecycle, namely design, development, and deployment. Figure 25 displays a screenshot of the threat assessment as filled in TrustGuard in the asset of the AI Learning Recommendations to demonstrate the frequency choice of each of the applicable threat categories during the deployment phase.

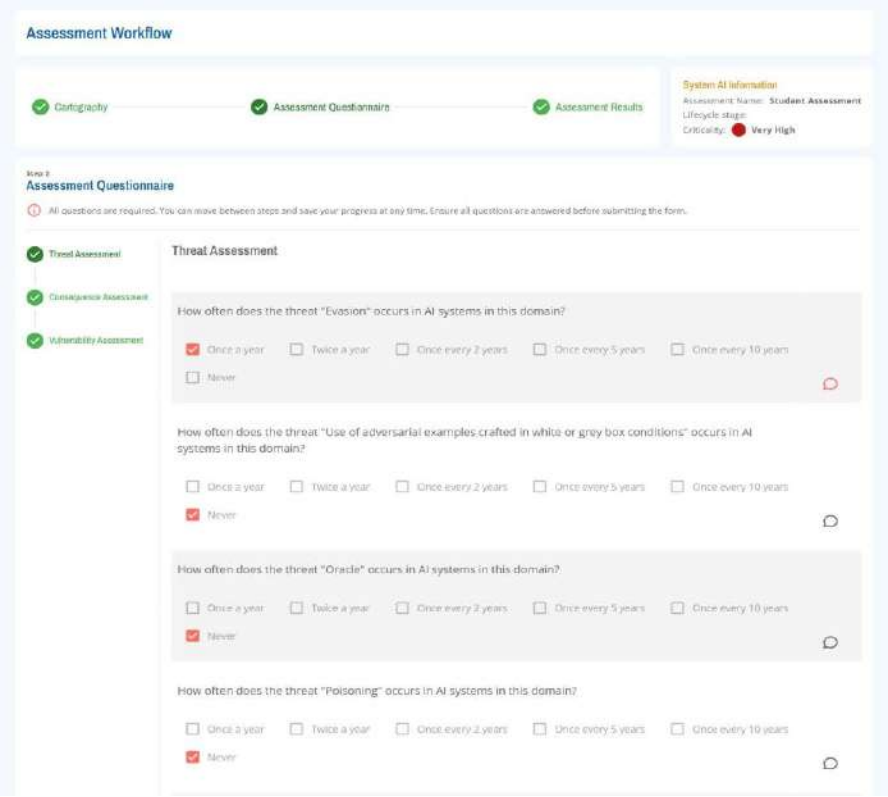


Figure 25: Threat Assessment for the Student Assessment

During the threat assessment, all potential threats were identified along with their frequency. The frequency was calculated based on both the likelihood of occurrence within the Educational LSP AI Solution and the observed occurrence of each threat since the system was deployed.

After completing the threat assessment, the risk assessor filled in the consequence assessment based on the input received from the educators regarding the potential consequences of the threat across all dimensions of trustworthiness. The next step is the completion of the vulnerability assessment for all models and datasets. The risk assessor answers questions related to the control measures that have been implemented. The combination of these inputs within TrustGuard produced risk ratings per threat, per AI component, and per lifecycle stage, yielding a comprehensive and granular risk profile for each AI asset of the Educational LSP AI Solution. Table 5 presents all identified threats, along with their frequency and a corresponding explanation.

Threat	Frequency	Explanation
Evasion	Once a year	Students may deliberately or inadvertently manipulate interaction patterns (e.g., guessing, rapid answering) to influence classifier outcomes.
Failure / Malfunction of ML Application	Once a year	Failures may occur due to software bugs, infrastructure issues, or incorrect model integration within the learning platform.
Human Error	Once a year	Mistakes during data preparation, labelling, configuration, or deployment can affect classifier reliability.
Model Drift Leading to Inaccurate Assessments	Once every 2 years	Evolving student behaviour, curriculum content, and learning environments may reduce model accuracy without periodic retraining.
Failure / Malfunction of ML Application	Once a year	Dataset-level failures (corrupted or incomplete records) may propagate to the model and degrade prediction quality.
Training Data Quality	Once a year	Real student interaction logs collected during educational pilots may include incomplete or inconsistent records.
Poisoning	Once every 2 years	Interaction logs may occasionally contain corrupt or maliciously influenced records; partial filtering is in place.
Limited Diversity in Training Data	Once every 2 years	Training data sourced from specific educational pilots may under-represent broader learner populations.

Changes in External Conditions	Once a year	Course design changes, new educational tools, or curriculum updates may alter input patterns and impact predictions.
Bias in Personalised Learning Recommendations	Once every 2 years	Algorithmic bias in personalised learning systems is a known structural risk, particularly with limited data diversity.
Ineffective Data Massaging	Once a year	Inadequate preprocessing or feature engineering may distort or fail to represent student behaviour patterns accurately.
Uniform Sampling Bias	Once every 2 years	Sampling strategies may not reflect the natural distribution of student performance levels, introducing systematic skew.
Poor Quality of Training/Testing Data	Once every 2 years	Noisy interaction logs, missing values, and incorrect labels can affect model performance.
Compromise of ML Application Components	Once every 5 years	Given the limited external-facing surface of the system, this is considered a low-frequency risk.

Table 5: List of identified threats, along with their frequency and a corresponding explanation for the student assessment.

3.4.3 AI Learning Recommendations

The AI Learning Recommendations model generates personalized learning pathways based on inquiry-based scenarios, using previous student interaction data. The AI assets considered in the in-isolation assessment are the AI Learning Recommendations model, the interaction dataset, and the learning materials dataset, which consists of inquiry-based scenarios created by teachers within the Educational LSP AI platform. The threat assessment was conducted across all phases of the AI lifecycle, namely design, development, and deployment. Figure 26 displays a screenshot of the threat assessment as filled in TrustGuard in the asset of the AI Learning Recommendations to demonstrate the frequency choice of each of the applicable threat categories during the deployment phase.

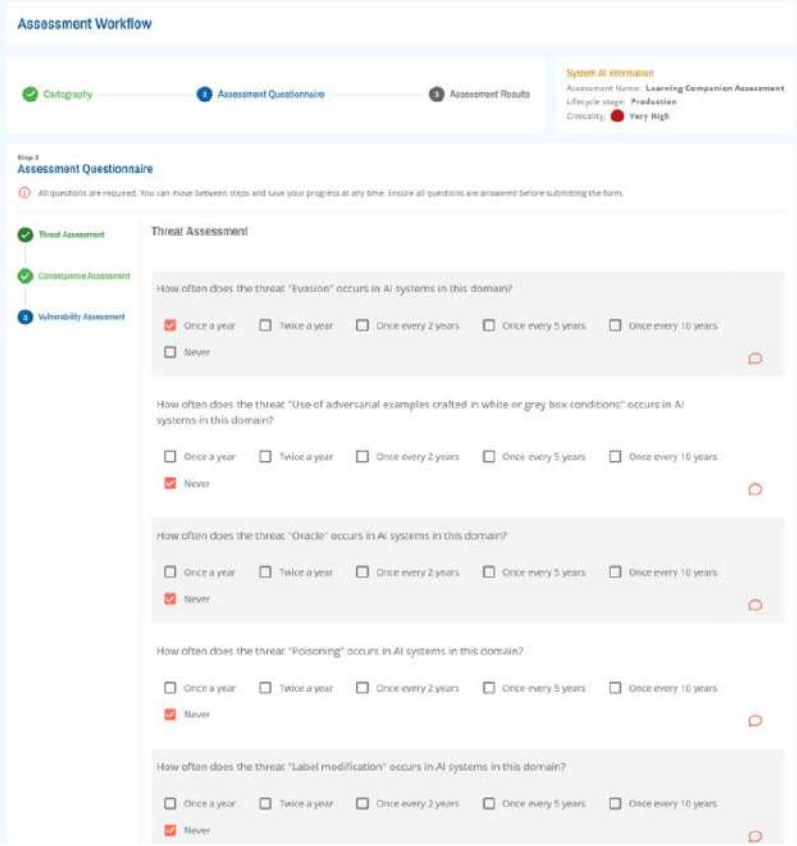


Figure 26: Threat Assessment for the AI Learning Recommendations

During the threat assessment, all potential threats were identified along with their frequency. The frequency was calculated based on both the likelihood of occurrence within the Educational LSP AI Solution and the observed occurrence of each threat since the system was deployed.

After completing the threat assessment, the risk assessor filled in the consequence assessment based on the input received from the educators regarding the potential consequences of the threat across all dimensions of trustworthiness. The next step is the completion of the vulnerability assessment for all models and datasets. The risk assessor answers questions related to the control measures that have been implemented. The combination of these inputs within TrustGuard produced risk ratings per threat, per AI component, and per lifecycle stage, yielding a comprehensive and granular risk profile for each AI asset of the Educational LSP AI Solution. Table 6 presents all identified threats, along with their frequency and a corresponding explanation.

Threat Category	Frequency	Explanation
Evasion	Once a year	Students may apply response strategies during normal system use that lead to incorrect skill inference by the model. This is classified as a recurrent behavioural phenomenon. The likelihood is classified as at least annual rather than a rare incident.
Failure or Malfunction of ML Application	Once a year	Occasional technical failures or malfunctions can occur due to software bugs or unexpected runtime conditions. Such failures are rare but plausible, estimated at approximately once per year.
Human Error	Twice a year	Human error by teachers is possible during review of the AI-generated recommendations. While errors are infrequent and generally minor, the likelihood is estimated at approximately twice per year.
Bias in Personalised Learning Recommendations	Once a year	Systematic bias affecting recommendation quality arising from unequal representation in training data.
Model Drift Leading to Inaccurate Assessments	Once a year	Model drift occurs when changes in student behavior and interaction data over time cause the model to rely on outdated patterns, leading to inaccurate assessments and reduced recommendation quality. Rare but plausible, estimated at approximately once per year.

Propagation of Misinformation	Once a year	Occasional errors in the review of learning recommendations may lead to minor propagation of misinformation. Given the controlled environment and testing procedures, estimated at approximately once per year.
Over-reliance in Model Predictions	Once a year	Teachers may trust AI-generated recommendations without sufficient human oversight, occurring approximately once per year.
Lack of Extensive Data for Training and Testing	Twice a year	Datasets are limited to specific schools, scenarios, or small pilot deployments, restricting generalisability.
Poor Quality of Data used in Training and Testing	Once a year	Noisy interaction logs and incomplete student activity records are filtered, but residual quality issues may persist. Estimated at approximately once per year.
Biased Training Data	Once a year	Educational datasets often reflect existing inequalities in student performance, access to resources, or participation patterns.
Lack of Effective Explanation Methods	Once a year	Some teachers may not fully understand the recommendations' explanations, reducing the effectiveness of human oversight.
Ineffective Data Massaging	Once a year	Data manipulation errors may be introduced by teachers during data preparation or review processes.

Training Data Quality due to Acquisition Procedure	Once a year	The training dataset is filtered, however, residual quality issues arising from the data acquisition process are plausible.
Changes in External Conditions	Once every 10 years	Shifts in curricula, learning environments, or technologies can degrade model performance over time. Assessed as a long-term risk occurring once every ten years.

Table 6: List of the identified threats, along with their frequency and a corresponding explanation for the AI Learning Recommendations.

The threats below were evaluated as never happening or not applicable to the Educational LSP deployment context.

Adversarial examples (when using white box or grey box settings) are not perceived as a threat since the teacher lacks the technical capability and expertise needed to design adversarial examples. Neither can Oracle attacks be relevant, since a user has no means, and motive to explore the system to acquire the concealed information. Attacks of poisoning are not considered as all the updates of the models are strictly regulated by the administrators, and teachers do not have access to the training data. On the same note, label modification cannot be done since the training, and evaluation labels are not available to the user.

There are reduced risks associated with model and data disclosure since model internals are not accessible and no sensitive data is kept in the system. Denial of service, general cybersecurity attacks, and timely attacks are also non-applicable in the present deployment scenario. The issue of regulatory and legal non-compliance is not an issue as the relevant regulatory frameworks are strictly observed, and all data management processes are centrally managed. Lastly, the entire domain-specific risks associated with other LSP domains include infrastructure inspection, routing, and underwater operations, which are completely inapplicable to Educational LSP domain.

3.5. Initial Risk Assessment results

This section describes the overall risk levels per AI Asset of the Educational LSP AI Solution identified by the Tools.

3.5.1 Learning Companion

The table below summarizes the overall risk levels identified by TrustGuard for the Learning Companion, with emphasis on the highest-risk findings. Risk scores are the product of the likelihood and consequence assessments performed during the TrustGuard evaluation.

Threat	Trustworthiness Dimension	Asset	Risk Score
Failure/malfunction of ML application	Safe, Valid and Reliable, Fair	Learning Companion	**Very High**
Evasion	Safe, Secure and Resilient	Learning Companion, Learning Material Dataset	**Very High**
Model Drift Leading to Inaccurate Assessments	Valid and Reliable	Learning Companion	**Very High**
Propagation of misinformation	Valid and Reliable, Explainable	Learning Companion	**Very High**
Human error	Explainable, Accountable	Learning Companion	**Very High**
Changes in external conditions	Valid and Reliable	Learning Companion	**High**
Failure/malfunction of ML application	Valid and Reliable	Learning Material Dataset	**High**
Compromise of ML application	Safe, Secure and Resilient	Learning Companion	**Low**
Poisoning	Safe, Secure and Resilient	Learning Companion	**Low**
Model or data disclosure	Secure and Resilient	Learning Companion	**Low**
Data disclosure	Secure and Resilient	Learning Companion	**Low**

Table 7: Overall risk levels identified by TrustGuard for the Learning Companion

The four Very High risks are Evasion, Failure or Malfunction of ML Application, Propagation of Misinformation, and Model Drift, which are the main risk exposure of the system. Human Error is also rated as very high, being mainly motivated by the fact that there is a possibility of errors in configuration or data-handling by the teachers or administrators when it is operated on. The asset management controls partially help in mitigating this risk, but additional procedural controls are advisable

3.5.2 Student Assessment

The table below summarizes the overall risk levels identified by TrustGuard for the Student Assessment model, with emphasis on the highest-risk findings. Risk scores are the product of the likelihood and consequence assessments performed during the TrustGuard evaluation.

Threat	Trustworthiness Dimension	Asset	Risk Score
Failure/Malfunction of ML Application	Valid & Reliable	Student Assessment Model	**Very High**
Model Drift Leading to Inaccurate Assessments	Valid & Reliable	Student Assessment Model	**Very High**
Human Error	Explainable, Accountable, Safe	Student Assessment Model	**Very High**
Evasion	Secure & Resilient, Safe	Student Assessment Model, Student Interaction Dataset	**Very High**
Failure/Malfunction of ML Application	Valid & Reliable, Fair	Student Assessment Model, Student Interaction Dataset	**High**
Poor Quality of data	Valid & Reliable	Student Interaction Dataset	**Low**

Table 8: Overall risk levels identified by TrustGuard for the Student Assessment model

3.5.3 AI Learning Recommendations

The table below summarizes the overall risk levels identified by TrustGuard for the AI Learning Recommendations, with emphasis on the highest risk findings. Risk scores are the product of the likelihood and consequence assessments performed during the TrustGuard evaluation.

Threat	Trustworthiness Dimension	Asset	Risk Score
Failure or Malfunction of ML Application	Valid & Reliable	AI Learning Recommendations	**Very High**
Model Drift Leading to Inaccurate Assessments	Valid & Reliable	AI Learning Recommendations	**Very High**
Human Error	Explainable, Accountable	AI Learning Recommendations	**Very High**
Bias in Personalized Learning Recommendations	Fair – Bias Managed	AI Learning Recommendations	**Very High**
Evasion	Secure & Resilient, Safe	AI Learning Recommendations, Learning Material Dataset, Student Interaction Dataset	**Very High**
Failure or Malfunction of ML Application	Valid & Reliable	Learning Material Dataset, Student Interaction Dataset	**High**
Failure or Malfunction of ML Application	Fair – Bias Managed	Student Interaction Dataset	**High**
Failure or Malfunction of ML Application	Fair – Bias Managed	AI Learning Recommendations	**High**

Propagation of Misinformation	Explainable & Interpretable	AI Learning Recommendations	High
Lack of Generalisation in Production Settings	Valid & Reliable	AI Learning Recommendations	Low
Failure or Malfunction of ML Application	Explainable & Interpretable	AI Learning Recommendations	Low
Reduced Operator Accountability due to Lack of Interpretability	Explainable & Interpretable	AI Learning Recommendations	Low
Reduced Ability of Operators to Identify Prediction Errors	Explainable & Interpretable	AI Learning Recommendations	Low
Insufficient Safeguards against Bias	Explainable & Interpretable, Fair	AI Learning Recommendations	Low
Changes in External Conditions	Fair – Bias Managed	AI Learning Recommendations	Low

Table 9: Overall risk levels identified by TrustGuard for the AI Learning Recommendations

The Validity dimension presents high risk of Failure or Malfunction of the ML Application and Model Drift That results in Inaccurate Assessments, and low risk of Failure or Malfunction of the ML Application. 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.

The rest of the threats were evaluated as Low risk. These are: Model or data disclosure, Data disclosure, Poisoning, Label modification, Oracle attacks, Regulatory and legal non-compliance,

Denial of service, Model disclosure, and other threats of fairness. Most of these low-risk findings do not need any mitigation measures due to the controlled educational deployment environment and the lack of sensitive stored data. It is important to note that a few threats were directly determined as not applicable to the Education LSP situation, and they are not included in the risk register.

3.6. Countermeasures

3.6.1 AI Learning Companion

This section is a report of the countermeasures proposed by TrustGuard related to the Learning Companion asset, as generated in the Mitigation Controls table after the in-isolation trustworthiness assessment is completed. The mitigation actions are presented by trustworthiness dimension and implementation status. Each control implemented has a narrative description describing the measure taken and how it will impact the identified risk. Postponed and ignored controls are also documented, along with the observations of the risk assessor on the justification and remaining risk. Figure 27 presents a screenshot from TrustGuard showing the mitigation controls selected by the risk assessor for threats related to the validity of the AI Learning Companion, indicating which controls were implemented, ignored, or postponed.

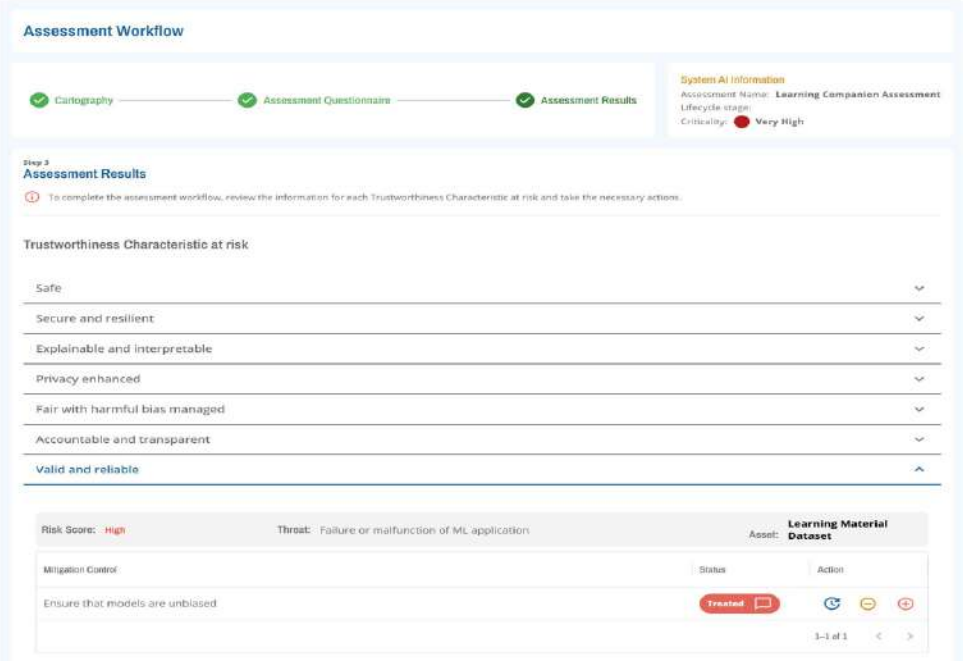


Figure 27: Mitigation controls selected by the risk assessor for threats related to the validity of the AI Learning Companion.

Among the mitigation measures suggested by TrustGuard in all six trustworthiness dimensions, some of these measures are known to be in place (Treated), some have been put off (Postponed), and others have been evaluated as not relevant or not justified in the present state of affairs (Ignored). Table 10 shows a summary of all mitigation measures suggested by TrustGuard regarding the Learning Companion asset, their implementation, and the risk leftover observations observed by the risk assessor.

Dimension	Threat	Mitigation Control	Status	Explanation
Accountable and Transparent	Human error	Include ML applications in asset management processes	**Treated**	ML components registered with defined ownership, versioning, and access control.
Safe	Failure or malfunction of ML application	Conduct a risk analysis of the ML application	**Treated**	Risk analysis completed..
Safe	Failure or malfunction of ML application	Define and monitor indicators for proper functioning of the model	**Treated**	Model performance is monitored via accuracy, precision, recall, and F1-score.
Secure and Resilient	Poisoning	Ensure reliable sources are used	**Treated**	All training data created by teachers.
Secure and Resilient	Data disclosure	Integrate ML-specific awareness into stakeholder strategy	**Treated**	ML-specific aspects incorporated into awareness strategy and disseminated to relevant stakeholders.
Valid and Reliable	Model Drift Leading to Inaccurate Assessments	Continuously monitor model performance	**Treated**	Performance monitoring in place.
Fair with Harmful Bias Managed	Failure or malfunction of ML application	Ensure that models are unbiased	**Treated**	Teachers create and review all scenarios, ensuring content is unbiased.
Fair with Harmful Bias Managed	Failure or malfunction of ML application	Assess regulations and laws the ML application must comply with	**Treated**	Compliance assessment conducted.

Privacy Enhanced	Data disclosure	Ensure that models respect differential privacy	**Treated**	Models confirmed to comply with differential privacy principles.
Privacy Enhanced	Model or data disclosure	Conduct a risk analysis of the ML application	**Treated**	Risk analysis completed covering data and model disclosure scenarios.
Valid and Reliable	Lack of generalisability in production settings	Adopt uncertainty quantification and drift detection tools	**Postponed**	

Table 10:Mitigation measures suggested by TrustGuard for the Learning Companion

3.6.2 Student Assessment

The countermeasures proposed by TrustGuard related to the Student Assessment asset are presented in this section, after the in-isolation trustworthiness assessment is completed. The mitigation actions are presented by trustworthiness dimension and implementation status. Each control implemented has a narrative description describing the measure taken and how it will impact the identified risk. Postponed and ignored controls are also documented, along with the observations of the risk assessor on the justification and remaining risk. Figure 28 presents a screenshot from TrustGuard showing the mitigation controls selected by the risk assessor for threats related to the validity of the Student Assessment, indicating which controls were implemented, ignored, or postponed.

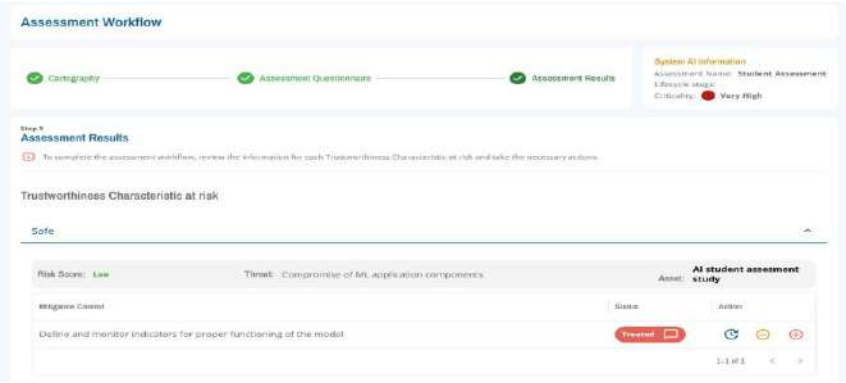


Figure 28: Mitigation controls selected by the risk assessor for threats related to the validity of the Student Assessment

Among the mitigation controls suggested by TrustGuard, some of these controls are known to be in place (Treated), some have been put off (Postponed), and others have been evaluated as not relevant or not justified in the present situation (Ignored). Table 11 shows a summary of the mitigation controls related to the Student Assessment asset, the status, and an explanation created by the risk assessor.

Mitigation Control	Current Status	Explanation
Conduct a risk analysis of the ML application	**Treated**	Risk analysis of the ML application was conducted.
Define and monitor indicators for proper functioning	**Treated**	Accuracy, precision, recall, and F1-score monitored continuously.
Ensure that models are unbiased	**Treated**	Completed.
Integrate ML specifics into awareness strategy	**Treated**	Completed.
Include ML applications into asset management processes	**Treated**	Completed. Defined ownership, access control, and versioning in place.
Continuously monitor model performance	**Treated**	Completed.
Ensure reliable sources are used	**Treated**	Completed. Noisy and incomplete interaction logs filtered at ingestion.
Embed fairness criteria into CI/CD pipelines	**Postponed**	
Use automated fairness testing tools within the model development lifecycle	**Postponed**	
Regularly test for bias regressions during system updates	**Postponed**	

Table 11: Mitigation controls related to the Student Assessment

3.6.3 AI Learning Recommendations

This section is a report of the countermeasures proposed by TrustGuard related to the AI Learning Recommendations asset, as generated in the Mitigation Controls table after the in-isolation trustworthiness assessment is completed. The mitigation actions are presented by trustworthiness dimension and implementation status. Each control implemented has a narrative description describing the measure taken and how it will impact the identified risk. Postponed and ignored controls are also documented, along with the observations of the risk assessor on the justification and remaining risk. Figure 29 presents a screenshot from TrustGuard showing the mitigation controls selected by the risk assessor for threats related to the validity of the AI Learning Recommendations, indicating which controls were implemented, ignored, or postponed.

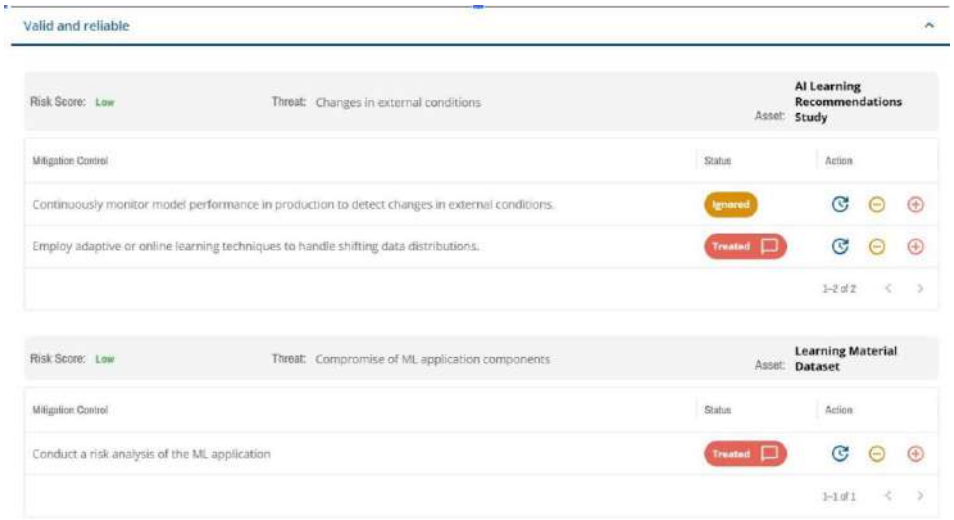


Figure 29: Mitigation controls selected by the risk assessor for threats related to the validity of the AI Learning Recommendations

Among the mitigation measures suggested by TrustGuard in all six trustworthiness dimensions, some of these measures are known to be in place (Treated), some have been put off (Postponed), and others have been evaluated as not relevant or not justified in the present state of affairs (Ignored). Table 12 shows a summary of all mitigation measures suggested by TrustGuard regarding the AI Learning Recommendations asset, their implementation, and the risk leftover observations observed by the risk assessor.

Mitigation Control	Dimension	Status	Explanation
Conduct a risk analysis of the ML application	All dimensions	**Treated**	A risk analysis was conducted.
Include ML applications into asset management processes	Accountable / Explainable	**Treated**	ML components are included in asset management with defined ownership, access control, and versioning. Teachers are trained to review recommendations
Define and monitor indicators for proper functioning	Safe / Valid / Accountable	**Treated**	Recommendation quality is monitored through Teacher Acceptance Rate and a Q-value table, which tracks and updates the effectiveness of recommended actions over time.
Ensure reliable sources are used	Fair / Safe / Valid	**Treated**	The dataset is created by teachers
Ensure that models are unbiased	Fair / Valid	**Treated**	Model fairness is evaluated across student performance categories by tracking acceptance rates and Q-values. Disparities are monitored and corrected to ensure unbiased recommendations across all categories.
Employ adaptive / online learning techniques	Fair / Valid	**Treated**	Adaptive learning techniques are employed through continuous updates of the Q-value table based on teacher feedback, allowing the system to adjust recommendations in response to changing student performance patterns and data distributions.

Engage legal counsel / privacy experts	Fair / Privacy	**Treated**	GDPR compliance confirmed
Track data provenance	Fair / Accountable	**Treated**	Origin and preprocessing recorded
Integrate ML specificities to awareness strategy	Secure / Privacy	**Treated**	All stakeholders informed
Ensure differential privacy	Privacy	**Treated**	No personal data stored
Continuously monitor model performance	Valid	**Treated**	Model performance is continuously monitored using metrics such as Teacher Acceptance Rate, Q-value trends, and category-specific outcomes to detect drift, maintain accuracy, and ensure reliable recommendations.
Implement bias detection and mitigation techniques, regular fairness audits.	Fair	**Postponed**	Planned to be addressed during the Replication Phase
Automated fairness testing	Fair / Explainable	**Postponed**	Planned to be addressed during the Replication Phase
Adopt assessment tools to detect possible malfunctioning of the system (e.g., uncertainty quantification strategies, outliers detection, drift detection tools) and inform the user when there can be potential misbehaviour of the system.	Valid	**Postponed**	Planned to be addressed during the Replication Phase

Adversarial training / resilient model design / adversarial detection	Safe / Secure	Ignored	Input is strictly limited to teacher-provided content, leaving no open interface for malicious injection.
Entropy-based measures / token-level probability	Explainable	Ignored	Output quality is ensured by human review.
Continuously monitor model performance for external condition changes	Valid / Fair	Ignored	

Table 12: Mitigation measures suggested by TrustGuard regarding the AI Learning Recommendations

Conduct a risk analysis of the ML application (Treated) – A formal threat analysis was performed on all the three assets of the AI Learning Recommendations assessment. This control has been adopted to deal with the threat of model and data disclosure, and the compromise of the components of ML applications.

Include ML applications into asset management processes (Treated) – ML elements of the AI Learning Recommendations system are officially integrated into the asset management model of the organization. This includes assigned ownership, access control policies and versioning of model artefacts and datasets. Moreover, teachers who are engaged with the system have been trained to critically examine AI-generated suggestions, which creates a layer of human oversight as an additional organizational control. This step directly caters to the Very High risk of Human Error on the Accountable and Transparent and Explainable and Interpretable dimensions.

Define and monitor indicators for proper functioning of the model (Treated) – Indicators for proper model functioning have been defined and are actively monitored in the production environment. In particular, the quality of recommendations is monitored by two complementary systems: the Teacher Acceptance Rate, which is the percentage of AI-generated recommendations accepted and implemented by teachers, and a Q-value table, which constantly monitors and updates the efficiency of recommended learning activities over a period of time based on the observed results of students

Ensure reliable sources are used (Treated) – All three assets have been guaranteed reliable data sources. The Learning Material Dataset is created by teachers themselves, which is a controlled and expert-tested source of training materials. The Student Interaction Dataset is sourced only to the working educational setting. Direct impact of this control is that the risk of poisoning in all sub-assets is reduced; the model is trained on the data of proven origin.

Ensure that models are unbiased (Treated) – Model fairness has been actively evaluated as part of the operational monitoring framework. For the Student Interaction Dataset, fairness is

assessed across student performance categories by tracking acceptance rates and Q-values per category. Disparities identified through this monitoring are corrected to ensure that recommendations remain unbiased across all performance groups. For the Learning Material Dataset, unbiasedness is maintained through the dataset's teacher-created origin, which is subject to pedagogical review. Teachers also perform ongoing reviews of AI-generated recommendations to detect and correct potential biases before they propagate to students. This control addresses the Very High risk of Bias in personalized Learning Recommendations.

Employ adaptive or online learning techniques to handle shifting data distributions (Treated) – Adaptive learning methods have been used to make sure that the recommendation model is in sync with the present patterns of student performance as time progresses. This is done by continuously updating the Q-value table with teacher input: when teachers are accepting or rejecting recommendations, the Q-values are modified to show what actions best work in the given circumstances. It is an adaptive mechanism that enables the system to react to slow changes in data distributions even without being re-trained with the whole model, directly satisfying the Low risk of Changes in External Conditions and also helping to mitigate the Model Drift.

Engage legal counsel or privacy experts to review data sources and terms of service (Treated) – Both Learning Material Dataset and Student Interaction Dataset have been analyzed in terms of legal and privacy review of the data sources. Even though neither of the datasets includes sensitive personal information in the traditional meaning, this review was conducted to ensure that the relevant regulatory provisions, such as the General Data Protection Regulation, are adhered to. The review ensured that any data handled during operation is as per the applicable legislative and governance provisions. This control deals with Low risk of Regulatory and Legal Non-Compliance in the two data assets.

Track data provenance (Treated) – Data provenance is systematically tracked for the Student Interaction Dataset to ensure full traceability. The origin, history and any preprocessing steps to each data item are logged and help in data integrity, auditing and accountability during model development and evaluation.

Integrate ML specificities to awareness strategy and ensure all ML stakeholders are receiving it (Treated) – ML-specific risks, limitations, and operational guidelines have been incorporated into the awareness strategy of all three sub-assets in the organization. This helps to make sure that all the stakeholders concerned are aware of the risks that the AI Learning Recommendations system poses and can responsibly engage with it. This control considers the threats of Data Disclosure, Oracle attacks, and Poisoning in the sub-assets by making sure that the users are aware of the limits of acceptable interaction with the system.

Ensure that models respect differential privacy (Treated) – Privacy by design principles have been exercised on the system architecture. The model has been set to avoid detailed output - hiding the scores of confidences, internal states, and debugging data - to minimize the possibility of inference attacks. The model does not process sensitive or identifiable data, making

conventional differential privacy techniques not applicable; however, the output restriction controls provide an equivalent protective effect within the system's operational context.

Continuously monitor model performance (Treated) – The combination of the measures Teacher Acceptance Rate, Q-value trends, and category-specific outcome tracking is continuously used in production to monitor the model performance. The main applied control against the Very High risk of Model Drift Leading to Inaccurate Assessments is this monitoring framework, allowing to detect the performance degradation early enough and making an informed decision regarding the model updates or retraining. The monitoring also offers evidence of operation in measuring whether the model is still fit in the various student groups that have been experienced in pilot schools.

Controls assessed as Postponed include:

- Implement bias detection and mitigation techniques, regular fairness audits. The risk assessor notes that these postponed controls represent a significant residual gap given the Very High risk of Bias in personalized Learning Recommendations. Their implementation should be prioritized in subsequent development cycles.
- Adopt assessment tools to detect possible malfunctioning of the system (uncertainty quantification, outlier detection, drift detection).
- Bias audits and fairness testing,
- Use automated fairness testing tools within the model development lifecycle,
- Regularly test for bias regressions during system updates, and
- Embed fairness criteria into continuous integration/continuous delivery pipelines.

These controls have been classified as Postponed as they are planned to be addressed during the Replication Phase of the FAITH Educational LSP pilot. The Very High-risk rating assigned to Bias in personalized learning recommendations is fully acknowledged by the risk assessor, and the implementation of bias detection and mitigation techniques, fairness audits, and the integration of automated fairness testing into the development lifecycle are designated as priority actions for the Replication Phase.

The following controls have been assessed as Ignored:

- Choose and define a more resilient model design,
- Include ML applications in detection and response to security incident processes,
- Implement tools to detect adversarial examples, and
- Use adversarial training techniques.
- Introduce entropy-based measures and Estimate token-level probability.
- Integrate ML applications into the overall cyber-resilience strategy
- Continuously monitor model performance in production to detect changes in external conditions.

These controls have been classified as “Ignored” based on the conducted risk assessment because the risks they address are of low significance within the current architecture and operational context of the Educational LSP AI solution. Although the system employs a large

language model (LLM) component for the generation of learning recommendations, the LLM operates within a tightly constrained input space: it receives as input only the educational material explicitly provided by the teacher. This architectural constraint significantly limits the attack surface for adversarial manipulation, as the system does not expose an interface through which malicious inputs could be freely injected. Furthermore, all AI recommendations are subject to mandatory teacher review before they are applied — only recommendations explicitly accepted by the teacher are acted upon, providing a robust human oversight layer that further mitigates the risks these controls are designed to address. Entropy-based measures and token-level probability estimation are therefore not considered necessary at this stage given the controlled and bounded nature of the LLM's operational context and the human-in-the-loop validation mechanism in place. These controls will be re-evaluated during the Replication Phase.

Any remaining privacy risks on the three sub-assets are rated as Low because of the controlled deployment environment and the lack of personally identifiable data.

3.7 Formal Tool Feedback

In this section, the formal feedback process implemented by the LSP3 team to the three FAITH tools utilized in the in-isolation trustworthiness evaluation of the Educational LSP AI Solution: the FAITH Model Hub, TrustGuard, and TrustSense is reported. For each tool, the feedback process is comprised of two stages: an initial informal feedback exchange, and the subsequent completion of a formal feedback questionnaire. The timing of each stage, the parties involved, and a high-level assessment of tool usefulness and usability are reported below.

Detailed analysis of the usefulness and usability findings will be reported in the dedicated WP2 deliverable. The present section limits its scope to characterizing the feedback process and providing a brief, high-level assessment of each tool.

3.7.1 FAITH Model Hub

Informal feedback on the FAITH Model Hub was communicated on 12 February, following the team's initial use of the tool for the registration of models and datasets associated with the Educational LSP AI Solution. The formal feedback questionnaire was finally completed on 7 April by the ICCS team, as the main user of the tool during the Cartography phase of the TAF V1 assessment.

Overall, the FAITH Model Hub was assessed as user-friendly and useful. The tool has been effective in enabling the LSP3 team to design and maintain the AI asset catalogue of the Educational LSP AI Solution to allow smoother integration with the TrustGuard assessment workflow. The fact that all Model Hub assets were available directly in TrustGuard at the Cartography phase was specifically appreciated.

On a high level, the feedback showed that there is one major area of improvement, which is the possibility of users editing datasets and models once they are created. Allowing post-creation editing would be a valuable addition to usability of the tool, especially in iterative assessment cases when metadata of assets might need to be improved during the assessment.

3.7.2 TrustSense

The formal feedback questionnaire for TrustSense was completed in December by the LSP3 team, following the tool's initial deployment to measure the Trustworthy AI maturity of participants involved in the Educational LSP AI Solution.

Additional informal feedback was sent on 19 January following the workshops, during which the TrustSense tool was used with teachers participating in the Educational LSP. The workshop feedback provided a valuable perspective from end-users who represent a key participant group for AI maturity measurement within the FAITH Educational LSP context. Also, the formal feedback questionnaire was completed on 27 April by the end users and the developers of the Educational LSP platform.

TrustSense was assessed as user-friendly overall. The tool was successfully used to measure the Trustworthy AI maturity of participants and to provide the maturity scores required as input to the TrustGuard risk calculation during the Cartography phase of TAF V1.

At a high level, the feedback identified some improvement areas. Several questions were noted as requiring rephrasing to specifically address teachers and the educational context. Also, teachers suggested that the addition of an example for each question would help them better understand what is being asked. Moreover, it was observed that the order in which questions appear in TrustSense varies between login sessions. Finally, the ability to export results in CSV format would be helpful to facilitate further analysis of AI maturity data by the LSP team and support reporting activities.

3.7.3 TrustGuard

Informal feedback on TrustGuard was communicated on 3 March, following the risk assessor's use of the tool to conduct the in-isolation trustworthiness assessment across all phases of TAF-v1. The formal feedback questionnaire was completed on 7 April by the risk assessor of the Educational LSP AI Solution, who was the main user of TrustGuard throughout the assessment process.

TrustGuard is a user-friendly tool that effectively supports the risk assessor in performing the trustworthiness assessment. The structured workflow of the TrustGuard tool implements the FAITH TAF-v1 framework and facilitates the assessment process.

The feedback highlighted that the inclusion of brief explanations for all threats within the tool would be beneficial. Providing contextual clarifications for each threat would assist risk assessors in applying consistent and well-informed judgements when estimating threat frequencies and selecting applicable threats per AI component.

4 AI System Evaluation

4.1 Quantitative metrics on pilot data

4.1.1 Learning Companion

We monitored the performance of the Learning Companion model that runs on Rasa 3.6.6, an open-source conversational AI framework that combines intent classification with dialogue management. The system was trained specifically on the platform's educational interactions and evaluated using a held-out test set.

The NLU component interprets student input and maps it to the appropriate instructional intent. On the test set, it achieved 100% accuracy across all 11 domain intents (85 test examples). Key intents in the learning flow, such as "*ask_me*" (readiness to engage), "*provide_answer*" (submitting a response), "*request_hint*" (seeking guidance), "*confirm_read*" (acknowledging content), and "*delete_data*" (resetting progress), all reached a precision, recall, and F1-score of 1.0, as displayed in table 13. No confusion between intent classes was observed. These results indicate that the conversational layer reliably understands student actions, allowing the Learning Companion to respond consistently throughout the learning process.

Intent	Precision	Recall	F1-Score	Support
ask_me	1.0	1.0	1.0	21
user_response	1.0	1.0	1.0	28
provide_answer	1.0	1.0	1.0	3
request_hint	1.0	1.0	1.0	5
confirm_read	1.0	1.0	1.0	3
affirm / deny / greet	1.0	1.0	1.0	13
start_scenario / scenario_selection / delete_data	1.0	1.0	1.0	12
Overall	1.0	1.0	1.0	85

Table 13: Rasa Intent and its equivalent F1 score and support.

This ensures that at every conversational turn, the model correctly understands student intent and triggers the appropriate pedagogical response, whether presenting the next activity, delivering a hint, processing an answer, or adapting the learning path.

Support represents the number of test examples available for each intent in the evaluation dataset. In Rasa, it shows how many times an intent appears in the test data and helps determine

how reliable the precision, recall, and F1-score are for that intent. In the reported results, some intents such as provide_answer and confirm_read have low support values (three samples). This is expected, since these intents are triggered through structured commands rather than free-form natural language input. As a result, they appear less frequently in the evaluation dataset.

The system does more than recognise intents. It also adapts the learning path based on student performance through a custom action (action_handle_answer). After each student response, the system records the answer with a timestamp, updates the correct and incorrect counters for the activity, and recalculates the student's cumulative scenario score. Based on this updated score, the next activity is selected using three performance tiers (high, moderate, low), routing each student to content that matches their current level. This score-driven routing allows the Learning Companion to adjust the instructional path in real time. Students who struggle are guided to reinforcement material, while higher-performing students move on to more challenging activities.

4.1.2 Student Assessment

Student performance categorization was implemented using an XGBoost classifier trained on data from 759 students who completed the Pendulum activity. For each student, 29 features were extracted across four activity phases. These included normalized scores per phase, penalized scores, log-transformed response times, and binary indicators capturing phase completion and timing quality. Additional aggregate features were also included, such as total score, mean penalized score, number of phases answered, and overall participation rate.

The initial categorization followed the PISA framework, which classifies students as Low, Moderate, or High achievers based solely on total score. This approach was extended by introducing engagement-based penalties. A phase was marked as disengaged when a student's response time exceeded the per-phase threshold or when the phase was not attempted. Students with two or more disengaged phases were downgraded by one category level, while students with three or more disengaged phases were downgraded by two levels.

This adjustment affected 44 students (5.8% of the dataset). Among them, 12 students initially classified as High were reassigned, 11 moved to Moderate and 1 to Low. Additionally, 32 students originally classified as Moderate were reassigned to Low. After applying the engagement-based correction, the final distribution was 323 Low (42.6%), 359 Moderate (47.3%), and 77 High (10.1%).

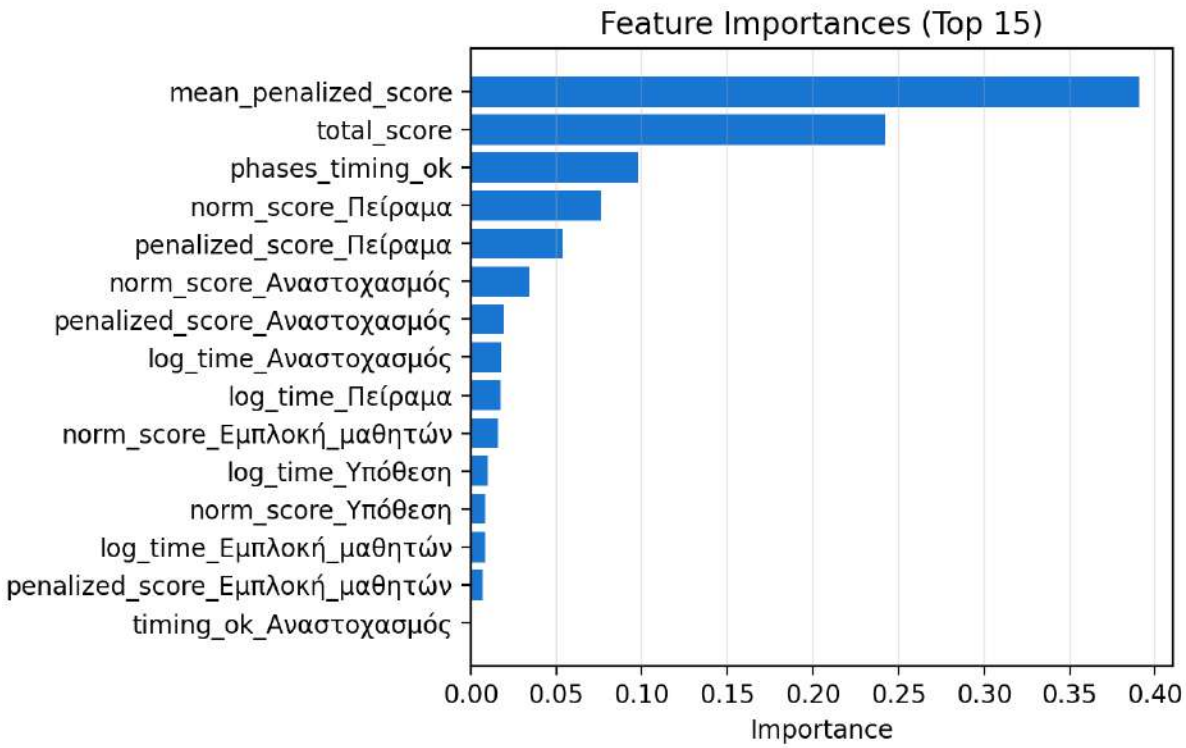


Figure 30: Feature Importances of XGBoost Classifier.

The XGBoost classifier was trained on 80% of the engagement-adjusted dataset (607 students) and evaluated on a stratified hold-out test set of 152 students that were not used during training. To ensure robust estimation, five-fold stratified cross-validation was performed on the training data. Figure 30 depicts that mean_penalized_score and total_score are the most important features when classifying students into the three performance categories.

Metric	CV (training set)	Hold-out test set
Accuracy	0.9423	0.9671
F1 Macro	0.9233	0.9668
F1 Weighted	0.9420	0.9671
Precision	0.9305	0.9763
Recall	0.9202	0.9556

Table 14: XGBoost metrics for classifying students in the three categories

Category	Precision	Recall	F1	Support
Low	0.98	0.95	0.97	65
Moderate	0.95	0.99	0.97	72
High	1.00	0.93	0.97	15

Table 15: Per category results on the hold-out test set

The performance of the XGBoost classifier is summarized in Table 14, showing high accuracy and F1-scores across both cross-validation and hold-out test sets. The per-category results, presented in Table 15, indicate consistently strong performance across all student performance categories. As seen in Figure 31, the High category achieved perfect precision (1.00), meaning no student was incorrectly classified as High. The recall for High was slightly lower (0.93), which indicates that a small number of genuinely high-performing students were conservatively assigned to the Moderate category.

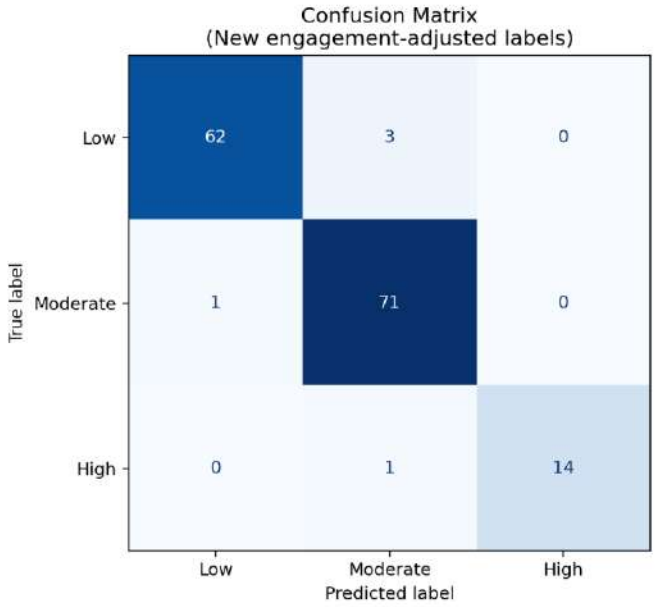


Figure 31: Confusion Matrix between XGBoost and PISA's actual categorization in the platform.

Overall, the model correctly categorized 96.7% of previously unseen students, suggesting strong generalization beyond the training data.

4.1.3 AI Learning Recommendations

The AI Learning Recommendations component helps teachers improve their scenarios based on how students perform. It monitors flagged activities where students in a specific performance group (High, Moderate, or Low) show signs of struggle, disengagement, or misconception, and

then generates proposals for instructional improvement. These recommendations are created automatically in the background and delivered directly to the teacher's authoring interface, allowing teachers to review and refine their scenarios based on observed learning behaviour.

The recommendation pipeline operates as a sequence of three components. It begins by analysing student answers and timing data to generate ActivityFlags, each representing a specific problem type. These include patterns such as Dynamic correctness threshold, Extreme correctness pattern, Timing discrepancy, and Engagement risk (too fast or too slow), identified for each student performance group. Once an activity is flagged, the system builds a Retrieval-Augmented Generation (RAG) index from the scenario's reference PDFs using ChromaDB vector embeddings. This provides grounded pedagogical context drawn directly from the teacher's own material.

A large language model (Qwen2.5:32b, served locally through Ollama) then receives the flagged activity, student performance insights, relevant PDF excerpts, scenario and phase summaries, and a recommendation derived from historical teacher feedback. Based on this information, the model generates a structured proposal suggesting whether the teacher should create a new activity, revise the existing one, or skip it. Each proposal is stored in the database and presented to the teacher for review. The teacher can accept the recommendation, modify it, or reject it. This decision is recorded and immediately fed back into the system's learning layer.

The recommendations generated by the system are not drawn from a fixed set of classes. Because of this, traditional classification metrics such as accuracy, precision, and recall are not suitable for evaluating correctness. Instead, recommendation quality is assessed through a reinforcement learning feedback mechanism based on Q-learning. The platform maintains a Q-value table indexed by *(flag_type, category, action)*. Each entry represents the learned value of selecting a particular action when a specific flag appears for a given student performance group.

To evaluate recommendation quality, the system tracks Acceptance Rate by Teachers, defined as the percentage of AI-generated recommendations accepted during review. Each proposal is assigned a status (new, accepted, rejected) and linked to the reviewing teacher and timestamp. This makes the acceptance rate fully auditable at scenario, teacher, or system level.

In the pilot deployment, the system achieved an acceptance rate of 80%, meaning four out of five recommendations were considered appropriate without requiring substantial revision. Rejected proposals generate a negative reward in the Q-learning layer, gradually steering the system away from less suitable actions for each contextual combination.

This represents an early baseline. As more teacher feedback is collected across scenarios, the Q-value table converges, and the acceptance rate is expected to improve. Over time, the recommendation policy prioritizes actions that have historically been validated by teachers for each *(flag type, student performance category)* context.

Flag Type	Category	Create	Revise	Skip
Dynamic correctness threshold	High	0.00	0.000	0.0
	Moderate	0.00	0.360	0.0
	Low	0.79	0.360	0.0
Extreme correctness pattern	High	0.00	0.000	0.0
	Moderate	0.00	0.590	0.0
	Low	0.00	0.000	0.0
Engagement risk (too slow, question)	High	0.20	0.360	0.0
	Moderate	0.20	0.000	0.0
	Low	0.00	0.360	0.0
Engagement risk (too fast, question)	High	0.00	0.000	0.0
	Moderate	0.00	0.000	0.0
	Low	0.00	0.360	0.0
Engagement risk (too slow, non-question)	High	0.00	0.738	0.0
	Moderate	0.20	0.360	0.0
	Low	0.00	0.000	0.0
Engagement risk (too fast, non-question)	High	0.00	0.000	0.0
	Moderate	0.00	0.000	0.0
	Low	0.00	0.360	0.0
Timing discrepancy under extreme correctness	High	0.00	0.000	0.0
	Moderate	0.00	0.000	0.0
	Low	0.00	0.000	0.0

Table 16: Initial Q-Table from the first set of teachers.

Additionally, teachers evaluated the trustworthiness of the AI Learning recommendations during workshops. Teachers evaluated the system based on its trustworthiness characteristics. The

feedback indicated that 80% agree with the statements that the AI learning recommendations that are provided by the system are valid and accurate and 60% of the respondents stated that they trust the AI recommendations. 60% of the respondents stated that they are sure they are responsible for the learning outcomes as they read the AI recommendations. 70% stated that they are certain that the AI learning recommendations they read have clear justifications. In terms of fairness and privacy, half of them stated that the AI recommendations are fair and free of prejudice, 70% stated that only teachers can see the recommendations, which does not compromise student privacy, and 80% affirmed that AI recommendations are processed in a manner that protects sensitive student information. Lastly, 80% of the teachers claimed that they are not worried about the fact that AI recommendations may substitute their work. In general, the responses show that the perception of the system in terms of its trustworthiness, fairness, and privacy protection is generally positive.

The screenshots of the survey questionnaires and the response graphs are the evidence that supports the user feedback. Teachers assessed the trustworthiness of AI learning recommendations across seven Trustworthiness Attributes (TWAs), as defined by the FAITH-AI-TAF framework: Validation & Reliability, Security & Resilience, Transparency & Accountability, Explainability & Interpretability, Safety, Fairness, and Privacy. The figures present teachers' perceptions of each trustworthiness attribute associated with the AI learning recommendations.

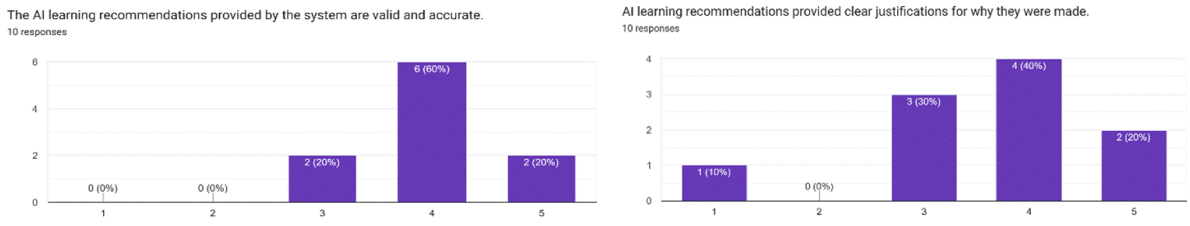


Figure 32: Validity (left) and Explainability (right) of the AI Learning Recommendations as evaluated by teachers

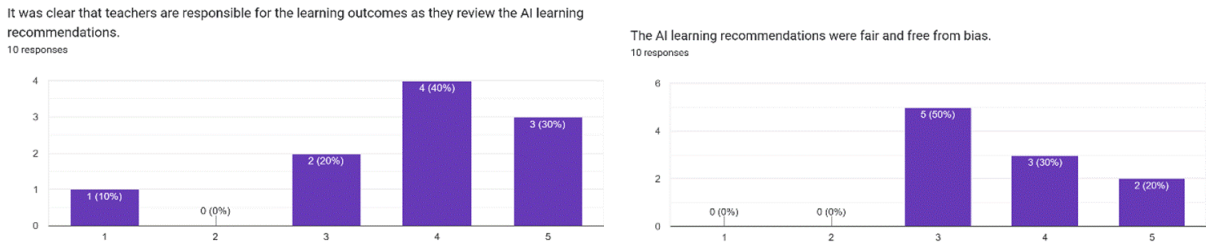


Figure 33:Transparency (left) and Fairness (right) of the AI Learning Recommendations as evaluated by teachers.

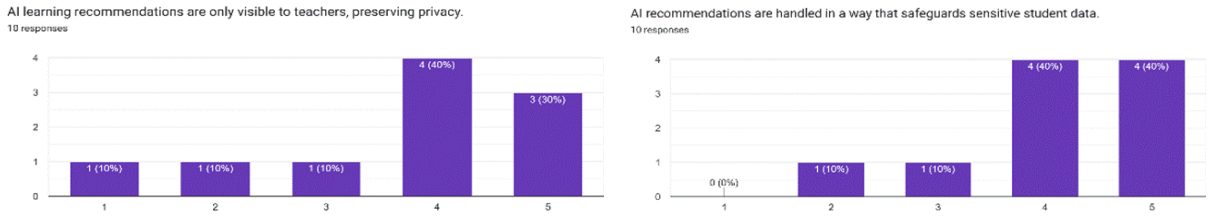


Figure 34:Privacy (left) and Security (right) of the AI Learning Recommendations as evaluated by teachers

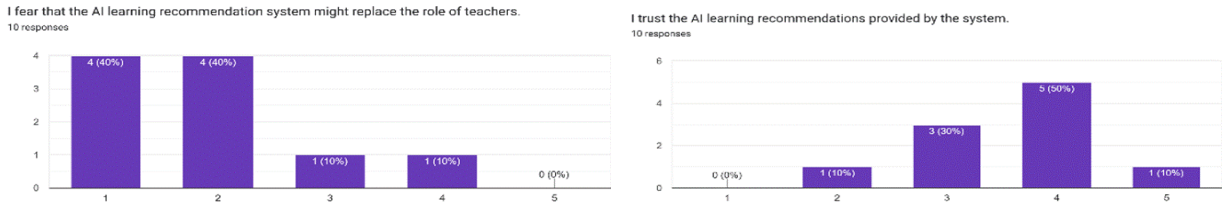


Figure 35: Safety (left) and Trust (right) of the AI Learning Recommendations as evaluated by teachers

4.2 Trustworthiness Assessment

This section reports on the insights generated by the FAITH tools regarding the performance of the AI assets of the Educational LSP AI Solution against the trustworthy AI dimensions defined within the FAITH AI TAF V1 framework. The assessment is grounded in the metrics and measurement approach established in FAITH Deliverable D2.1.

Given the nature of the Educational LSP AI Solution — an AI system providing personalized STEM learning support and adaptive recommendations to teachers— not all seven TWA dimensions carry equal relevance to this case. Two dimensions are assessed as more important for the Educational LSP context are validation and reliability and explainability and interpretability. The remaining dimensions — Fairness, Transparency and Accountability, Security and Resilience, Safety, and Privacy — were assessed within the TAF V1 risk evaluation through TrustGuard and are cross-referenced below but are not addressed as standalone TWA dimensions in this section.

Through workshops, teachers evaluated the trustworthiness of the system's AI recommendations, highlighting which characteristics are most essential for establishing trust in classroom AI tools. Based on the feedback collected by teachers, validity and explainability are the most important trustworthy dimensions. Validity and explainability primarily impact the pedagogical outcome, affecting student engagement if the content is unclear or scientifically incorrect, leading to questioning and potentially underestimating the procedure. Validity is also highlighted with respect to source referencing, and explainability is key for understanding.

Privacy and security are considered the most critical risks due to increasingly strict privacy regulations, monitoring by the Ministry of Education regarding student data usage in programs,

and the need for students and parents to trust that the data of their children remains private and secure. However, the Educational LSP AI solution has been designed with privacy by design and does not store any personal student data, so these risks are significantly mitigated. Teachers highlighted that privacy, safety, and security are major concerns and must be taken seriously.

The trustworthiness assessment integrates the structured risk-based assessment conducted using TrustGuard across all six TAF V1 phases, and the teacher evaluation of AI recommendation trustworthiness conducted through workshops.

4.2.1 Validation & Reliability

Validity and Reliability was the most important trustworthiness dimension rated by teachers (together with Explainability) As it directly relates to the pedagogical results because false or inaccurate recommendations may confuse teachers and negatively affect student interaction. In Section 4.1 there are quantitative performance measures which convincingly indicate model reliability. The AI Learning Companion has a 100% accuracy (F1 = 1.00), the Student Assessment Model has an accuracy of 96.7% and the AI Learning recommendations have a teacher recommendation acceptance rate of 80%. The results significantly boost the evidence base of reliability.

Additionally, teacher evaluation reinforces these findings. Based on teachers' responses to questionnaires during workshops, 80% of teachers agree that AI recommendations are valid and accurate, and 60% report trusting them. The 20-percentage-point gap between perceived validity and expressed trust is a consistent signal that technical performance alone is insufficient. Teachers require explainability and source referencing in addition to accurate outputs before fully extending trust to the system.

The absence of continuous monitoring of model performance was identified as a threat in the trustworthiness assessment using the TrustGuard tool and was rated as high consequence across multiple trustworthiness dimensions. The countermeasure ‘continuously monitor model performance’ was implemented to address threats related to the validity and reliability of the Educational LSP AI solution.

4.2.2 Explainability & Interpretability

Explainability generated the most substantive teacher feedback of the entire first pilot cycle. Teachers emphasized that explanations are important to understand why the Educational LSP AI Solution proposes specific learning recommendations and to be educationally grounded. The 70% agreement that recommendations have clear justifications is a positive signal, but the fact that only 60% of teachers report trusting the recommendations indicates that clarity of justification alone is not yet sufficient to bridge the gap to full trust.

To overcome the shortage of explanation, explainability dashboards were introduced to give transparent information on why specific learning recommendations are produced to allow users

and stakeholders to gain a better insight into the underlying decision-making processes of the system. Nevertheless, the control measure, which was determined as being at risk, according to TrustGUARD, was the control measure: Integrate ML specificities into the awareness strategy and ensure that all ML stakeholders are receiving it. This risk implies that even with such technical solutions as dashboards, it is possible that there are gaps in the effective communication of machine learning-specific concepts, risks, and responsibilities to all concerned stakeholders. To counter this, workshops were held to inform and educate all the involved stakeholders to ensure that they are well aware of ML-specific considerations and able to interpret and use the tools of explainability that are available. Consequently, these joint efforts enhance the baseline of knowledge, encourage reasonable usage, and improve the general management and credibility of the system.

4.2.3 Security, Transparency, Safety, Fairness and Privacy

The further TWA dimensions — Security and Resilience, Transparency, Safety, Fairness, and Privacy — were assessed within the TAF V1 risk evaluation conducted using TrustGuard and are noted here for completeness. These dimensions are not addressed as standalone sections in this report. Their full technical treatment is found in Section 3 “Application of FAITH Tools”.

- **Security and Resilience:** Relevant threats were assessed across all AI assets and lifecycle stages using TrustGuard. Countermeasures have been selected and are documented in the risk register. No security incidents were observed during the first pilot cycle. The anonymization architecture substantially reduces the attack surface for student data-related security threats.
- **Transparency**: Teachers confirmed they were clearly informed about the AI system's role as an advisory tool subject to their professional oversight, consistent with the human oversight requirements of Article 14 of the EU AI Act. The Ethics Committee approval and the TrustGuard audit trail provide strong institutional and procedural accountability foundations.
- **Fairness**: The 50% fairness agreement is the most concerning survey result. While the anonymization architecture prevents individual-level discriminatory profiling, the structural risk of bias remains an open concern. Explicit fairness monitoring is a priority for the Replication Phase evaluation plan.
- **Safety:** Safety threats were assessed through TrustGuard's consequence and vulnerability analysis phases. Human oversight by educators is the primary safety control mechanism, consistent with Article 14 of the EU AI Act. 80% of teachers stated they are not worried that the AI system will substitute their role, which is a strong positive finding that reflects the system's positioning as a support tool rather than a decision-maker.
- **Privacy:** Privacy is addressed through the privacy-by-design architecture of the Educational LSP AI Solution, whereby teachers create anonymous student accounts and distribute them randomly, preventing the system from identifying individual students. Only anonymized learning activity data, performance metrics, and interaction logs are

stored. Privacy by Design implementation substantially mitigates the privacy risk profile of the solution.

4.3 Preliminary Impact Evaluation

Early evidence from the simple pendulum scenario shows that AI-enhanced inquiry learning has a positive effect on how students engage with the different phases of an inquiry-based scenario. The adaptive support offered by the AI system helps learners move past difficulties instead of getting stuck, while additional tasks proposed during phases such as the Reflection encourage deeper thinking, especially for students already performing well. This creates a more supportive and responsive learning experience that keeps students active throughout the inquiry process.

A clear impact is seen in how moderate performers respond to the AI-enhanced design. Their average completion time increased from about 13.44 minutes in the non-AI training runs to roughly twenty minutes in the AI-enhanced version (20.03 minutes in one implementation and 19.32 minutes in another). This increase reflects stronger persistence, greater involvement, and more meaningful interaction with the activities, even though it also shows that these students often need steady guidance as they work through the scenario.

Another encouraging finding comes from the non-AI implementations, where overall Low performers showed two distinct completion times: one around 5 minutes and another around 25 minutes. Rather than being a problem, this pattern is useful because it allows the system to distinguish between students who disengage immediately and those who genuinely try but face difficulties. This separation is essential, as it enables the AI to offer different types of supporting reengagement prompts for the first group and more structured guidance for the second.

Several areas still need attention. Moderate performers remain a diverse group with different needs, meaning that a single type of support cannot reach all of them effectively. Certain phases, especially in more demanding topics, continue to show mixed performance, and learners with limited prior knowledge may struggle without additional preparation or explanation. Gathering more balanced data across implementations will help refine future AI suggestions, and the system should better identify very short or very long working times as signals of confusion or disengagement. Teachers remain central to the process, adjusting AI-suggested activities and ensuring that learning stays meaningful. With richer data, clearer guidance, and thoughtful scaffolding, the AI-enhanced environment can develop into an even more supportive and inclusive learning space.

5 Regulatory and Ethical Assessment

In this section, the regulatory and ethical evaluation of the Educational LSP AI Solution implemented in the FAITH project is described. It discusses the categorization of the Educational LSP AI Solution, its responsibilities in accordance with the EU AI Act and its compliance with the General Data Protection Regulation (GDPR).

The assessment draws on the outputs of the in-isolation trustworthiness assessment conducted using the FAITH AI TAF V1 methodology and complements that technical risk-based evaluation with an explicit examination of the regulatory and ethical dimensions applicable to the Educational LSP AI Solution.

5.1. Alignment with the EU AI Act (High-Risk AI requirements)

The Educational LSP AI Solution has been evaluated and rated as a high-risk AI system according to the EU AI Act. This categorization was established at the Cartography stage of the FAITH AI TAF V1 test, as per Article 6 of the EU AI Act, as interpreted alongside Annex III.

In particular, Annex III, Point 1 of the AI Act defines the AI systems that are to be applied in the educational or vocational training settings as high-risk when they are applied to: identify access to or admissions in educational institutions; assess learning results and measure competencies; track and identify prohibited behaviour. Educational LSP AI Solution, using machine learning models to personalize learning paths, provide suggestions to teachers, and assess the problem-solving development of students, is in this category.

Being a high-risk AI system, the Educational LSP AI Solution must comply with the obligatory requirements established in Chapter III, Section 2 of the AI Act (Articles 9-15) before it is placed on the market or put into service along with the responsibilities to the deployers provided in Chapter III, Section 3.

The table below aligns each of the relevant high-risk AI system requirements in the EU AI Act with its status and activities in the Educational LSP AI Solution.

Article	Requirement	Obligation Summary	Educational LSP Status & Actions
Art. 9	**Risk Management System**	A documented, iterative risk management system must be established and maintained throughout the AI system lifecycle.	Established via the FAITH AI TAF V1 methodology. TrustGuard implements the six-phase risk assessment cycle across all AI assets and lifecycle stages.

Art. 10	**Data and Data Governance**	Training, validation, and test data must be subject to appropriate data governance practices, including examination for biases.	Data governance practices applied to Educational LSP datasets; datasets registered and described in FAITH Model Hub; bias threat identified and assessed.
Art. 11	**Technical Documentation**	Comprehensive technical documentation must be drawn up before the system is placed on the market.	Technical documentation has been completed.
Art. 12	**Record-Keeping / Logging**	High-risk AI systems must be capable of automatic logging of events throughout their lifetime.	Implementation of automatic event logging has been conducted.
Art. 13	**Transparency & Provision of Information**	High-risk AI systems must be transparent and provide sufficient information to deployers.	Explainability dashboards have been implemented.
Art. 14	**Human Oversight**	High-risk AI systems must be designed to enable effective human oversight during operation.	Human oversight mechanisms embedded in the Educational LSP AI Solution design; teacher involvement mandated in consequence assessment and validation.
Art. 15	**Accuracy, Robustness & Cybersecurity**	High-risk AI systems must achieve appropriate levels of accuracy, robustness, and cybersecurity.	Robustness and security threats assessed across all AI assets and lifecycle stages

Art. 26	**Obligations of Deployers**	Deployers must use high-risk AI systems in accordance with instructions for use and ensure human oversight.	Deployer obligations communicated to LSP3 educational institutions; human oversight protocols established.

Table 17: Mapping of EU AI Act High-Risk AI Requirements in the Educational LSP AI Solution.

5.2. GDPR Compliance Assessment

The GDPR compliance profile of the Educational LSP AI Solution is largely determined by the anonymization architecture of the solution. Since the system is set in such a way that it does not reveal the names of individual students, teachers create anonymous accounts and share them randomly, so that the system records and the actual student identities cannot be linked, the solution goes largely beyond the scope of the GDPR requirements regarding the data on students. The system does not retain, process or deduce personal data that can be identified with specific students.

Regardless of the anonymization model, a review of compliance with GDPR was performed to make everything fully regulated. The review looks at the remaining GDPR implications concerning the Educational LSP AI Solution, such as the data relating to teacher accounts and the institutional data flows involving the implementation of the system. The GDPR and the EU AI Act are complementary documents: the data governance (Article 10), transparency (Article 13) and human control (Article 14) of the AI Act are consistent and supportive of the principles of data minimization, purpose limitation and transparency of the GDPR, which are reflected in the anonymization-by-design approach.

The anonymization structure of the Educational LSP AI Solution is a successful and intentional Privacy by Design. The solution removes the major type of privacy risk posed by AI systems implemented in the education context, ensuring that the system is structurally unable to determine individual students. The institutional level data governance mechanisms around the process of creating and distributing anonymous accounts, such as the security of the teacher-to-student account mapping stored outside the system, lie with the participating school, and corresponding protocols have been set up in consultation with the school data protection guidelines.

5.3 Ethics Committee Approval

The approval process of the Ethics Committee was based on the description made by the LSP3 team of the following aspects: the purpose and intended functionality of the Educational LSP AI Solution; the pedagogical context and the roles of all human participants (students, teachers, developers); the anonymization architecture and the specific types of data that the system will store; the measures provided to ensure the protection of student privacy and wellbeing. This documentation was thoroughly reviewed by the Ethics Committee of EA, who assured that the actions of the pilot are ethical and that the safeguards are sufficient.

The Educational LSP AI Solution has also been evaluated with respect to the seven essential criteria of trustful AI outlined by the European Commission in its High-Level Expert Group on Artificial Intelligence (HLEG) in the Ethics Guidelines on Trustworthy AI (2019), in addition to the formal approval of the Ethics Committee. These are the ethical requirements that the regulatory framework of the EU AI Act is based on and form a significant supplementary aspect of the trustworthiness examination.

By aligning the seven HLEG requirements: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-Discrimination and Fairness; Societal and Environmental Wellbeing; and Accountability directly to the trustworthiness dimensions measured in the FAITH AI TAF V1 approach, it offers a coherent and consistent ethical evaluation scheme. The alignment findings in Table 18 below are further supported by the Ethics Committee approval.

HLEG Principle	Ethical Requirement	Educational LSP Alignment
Human Agency & Oversight	AI systems should support human autonomy and decision-making, not replace it.	Educators retain full decision-making authority; AI outputs are advisory only. Human oversight is embedded in system design and was validated during the consequence assessment phase. The Ethics Committee review confirmed that the design adequately preserves educator professional autonomy.
Technical Robustness & Safety	AI systems should be technically robust, secure, and safe, with contingency plans for failures.	Robustness and safety threats assessed across all AI assets and lifecycle stages under TAF V1. Countermeasures selected to mitigate identified risks. The system's robustness characteristics were documented as part of the materials submitted to and approved by the Ethics Committee.
Privacy & Data Governance	Privacy must be protected and data governance standards upheld.	Privacy by Design is implemented through the anonymisation architecture: teachers create anonymised student accounts distributed randomly, preventing the system from identifying individual students. Only anonymised learning activity data, performance metrics, and interaction logs are stored. This approach was explicitly described in the system documentation submitted to the Ethics Committee of EA and received formal approval.

Transparency	The capabilities and limitations of AI systems must be communicated to users.	Transparency and explainability are identified as trustworthiness dimensions. Explainability dashboards have been implemented to help educators understand the reasoning behind learning recommendations. The Ethics Committee submission included a detailed description of the system's capabilities, limitations, and the nature of the data stored, supporting informed transparency with all stakeholders.
Diversity, Non-Discrimination & Fairness	AI systems must not perpetuate or amplify bias or discrimination.	Fairness reviewed as a trustworthiness dimension; educator validation used to contextualise impact. The anonymisation model additionally mitigates the risk of discriminatory profiling of individual students, as the system cannot associate outputs with identifiable individuals.
Societal & Environmental Wellbeing	AI systems should benefit individuals and society as a whole.	The Educational LSP AI Solution is designed to support personalized learning and improve STEM educational outcomes for all students. Societal benefit — through improved STEM competence and reduced attainment gaps — is a primary objective of the FAITH Educational LSP and was presented to the Ethics Committee as a core justification for the pilot.
Accountability	Mechanisms to ensure responsibility and accountability for AI systems must be in place.	The risk assessor role is formally defined; the FAITH TAF V1 provides an auditable, structured accountability mechanism; and institutional responsibilities are clearly allocated. Critically, the Ethics Committee of EA reviewed the full system and data description and granted formal ethical approval, providing an independent accountability checkpoint for the pilot.

Table 18: Educational LSP AI Solution — Alignment with HLEG Trustworthy AI Principles

These seven ethical principles are operationalized in the FAITH AI TAF V1 in-isolation trustworthiness assessment, which is performed with the help of TrustGuard and its organized six-step approach, which transforms abstract ethical requirements into specific, measurable assessment results. Together with the formal Ethics Committee approval obtained through EA, the trustworthiness assessment thus can be viewed as a regulatory compliance tool, an ethical due diligence tool, and an independently validated ethical assurance tool.

6 Conclusions and Roadmap for the Final Pilot Phase

6.1 Summary of achievements in the second period

The initial pilot of the Education LSP has been successfully completed. The Educational LSP AI Solution was deployed and applied in a real secondary school environment, where teachers and students engaged in structured interactions with the AI system across repeated inquiry-based learning cycles. In addition, the in-isolation assessment has been finalized, covering the full six-phase trustworthiness evaluation of all AI assets using the FAITH AI TAF V1 methodology via TrustGuard, spanning all lifecycle stages (design, development, and deployment) from Cartography through Risk Management. Furthermore, all models and datasets have been registered and fully described in the FAITH Model Hub, creating an auditable AI asset catalogue and enabling seamless integration with TrustGuard, while the Trustworthy AI maturity of all relevant participant groups has been measured using TrustSense, with the resulting scores incorporated into risk calculations.

Moreover, a series of workshops with participating teachers provided structured qualitative feedback on trustworthiness dimensions. Additionally, both formal and informal feedback was submitted for all three FAITH tools (FAITH Model Hub, TrustGuard, and TrustSense) throughout the reporting period. In parallel, formal ethical approval was obtained from the Ethics Committee of EA following the submission of comprehensive system and data documentation, offering independent external validation of both the AI solution and the pilot methodology. Finally, the regulatory and ethical assessment has been completed, including compliance mapping with EU AI Act high-risk requirements, and assessment of GDPR compliance.

Taken together, these achievements demonstrate that the Educational LSP has successfully executed its first pilot phase in alignment with the objectives of Task 6.2 and with the FAITH project's broader mission of developing, validating, and deploying a comprehensive trustworthiness assessment ecosystem for real-world AI systems. The iterative, teacher-center approach adopted — grounded in the inquiry-based and project-based learning principles that guide the project — has generated rich qualitative evidence regarding the trustworthiness of the Educational LSP AI Solution and has established a clear improvement agenda for the Replication Phase.

6.2 Plan for Replication Phase

During the Replication Pilot, we are going to perform the in-context trustworthiness evaluation of the Educational LSP AI solution in accordance with the FAITH_AI_TAF_v2 framework and using the FAITH_AI_TAF_v2 tools. This stage is a crucial step in the evolution of the already completed in-isolation assessment, which allows assessing the system in real working conditions and in its actual educational environment of implementation.

The first step will be to conduct specific training workshops on the FAITH AI TAF v2 tools to make sure that the methodology is applied uniformly and informed by the training participants. Based

on this capacity, we will evaluate the trustworthiness of the Educational LSP solution with the FAITH-AI-TAF_v2 tools, with real-world interactions, stakeholder feedback, and contextual factors affecting how the system will behave in the various learning contexts.

In August 2025 (M20), the Replication Pilot of the Educational LSP was launched, and student implementations were already being done. These activities assist in the systematic gathering of the quantitative and qualitative data that would be required to make the in-context assessment, such as the performance of the system, user interaction, and trustworthiness-related indicators in all the dimensions considered.

This step is one of the stepstones in the process of validating, refining and extending the AI-supported learning framework of the project. It allows not only validating the earlier results in real circumstances but also detecting context-specific threats, weaknesses, and enhance opportunities, which eventually leads to a more resilient, scalable, and reliable Educational LSP AI Solution.

6.3 Identified risks for the Replication Phase

The Replication Phase of the FAITH Educational LSP advances the pilot from the in-isolation, asset-level trustworthiness assessment conducted in the initial pilot phase to an in-context assessment of the Educational LSP AI Solution as a connected socio-technical system, embedded within and interacting with the wider human, institutional, and pedagogical environment of the replication site. This transition brings with it a qualitatively different risk profile. Risks that were manageable in the controlled conditions of the first pilot may become more significant when the system operates within a broader institutional setting, interacting with a new group of educators, a more diverse student population, and the organizational processes of the replication site. The following risks have been identified based on the first pilot findings and the specific characteristics of the in-context replication deployment.

A first risk concerns educator AI maturity and the quality of human oversight at the replication phase. The first pilot demonstrated that the quality of human oversight is directly dependent on the Trustworthy AI maturity of the educators operating the system. Teachers with lower AI maturity are less confident in evaluating and acting on AI recommendations. At the replication site, where educators will not have benefited from the capacity-building experience of the first pilot cycle, lower baseline AI maturity is a plausible starting condition. Insufficient educator preparation could reduce the quality of field validation, weaken the human oversight safeguards that are central to the system's high-risk compliance framework, and limit the usefulness of teacher-evaluated trustworthiness evidence collected during the replication cycle. TrustSense AI maturity assessment of all replication site educators prior to deployment, followed by a tailored training program, will address this risk directly. Additionally, in case the teacher feedback is not satisfactory in terms of volume and quality, the risk could be mitigated through performative simulation.

A second risk concerns anonymization protocol compliance at scale. The Privacy by Design architecture of the Educational LSP AI Solution — whereby teachers create anonymous student accounts and distribute them randomly, preventing the system from identifying individual students — is a foundational privacy protection measure and a key element of the GDPR compliance approach. This architecture depends on consistent procedural compliance by all educators at the deployment site. In the in-context assessment, operating across a broader institutional environment with a new group of teachers who may be less familiar with the protocol, maintaining consistent anonymization compliance represents a more significant operational risk than in the controlled first pilot setting. Detailed procedural guidance will be embedded in educator onboarding materials, anonymization compliance will be included in the replication governance checklist, and monitoring mechanisms will be in place throughout the deployment period.

A third risk concerns the evolving regulatory and compliance landscape during the Replication Phase. The Educational LSP AI Solution operates within a dynamic regulatory environment. The EU AI Act is being progressively implemented through implementing acts, delegated acts, harmonized standards, and possible updates or clarifications during the Replication Phase may introduce additional documentation, governance, or compliance requirements. This is particularly relevant given the system's high-risk classification, which already places it under the most demanding tier of regulatory obligations. While no fundamental regulatory barrier to the Replication Phase is anticipated, practical challenges may arise in areas such as technical documentation, and formal evidence requirements as the system moves toward a more mature deployment stage. Regular regulatory monitoring and a compliance review at the commencement of the Replication Phase will ensure that any new requirements are identified and addressed promptly.

Overall, the identified risks for the Replication Phase can be summarized as follows: variable educator AI maturity and oversight quality at the replication site; the procedural demands of maintaining anonymization protocol compliance across a new group of educators; and an evolving regulatory environment requiring ongoing monitoring. These risks confirm that the in-context deployment should be approached as a controlled validation and refinement step in which technical improvement, educator preparation, and trustworthiness governance must proceed in parallel.