Ref. Ares(2026)4619138 - 06/05/2026

D4.2

Report on the initial pilot - media

Related Work Package	**WP4 – Implementation of Media LSP**
Related Task	Task 4.2 – The initial pilot
Lead Beneficiary	FH
Contributing Beneficiaries	ATC
0Document version	v.1.0
Deliverable Type	R
Distribution level	PU
Contractual Date of Delivery	30/04/2025
Actual Date of Delivery	06/05/2026

Authors	Cristina Guseth (FH), Teodora Musat (FH), I. Kanellos, (ATC), T. Tsafonis (ATC), G. Giotis (ATC), E. Tsalapati (ATC)
Contributors	
Reviewers	D. Moroni (CNR), G. Riccardo Leone (CNR), Andrea Carboni (CNR), Luca Volpicella (MME), Cristiano Tatini (MME)

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
v.0.1	TOC	26/02/2026
v.0.2	First draft	14/04/2026
v.0.9	Draft for internal revision	21/04/2026
v.1.0	Version updated with internal reviewers' comments	06/05/2026

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 activities performed for the initial pilot phase of the Media Large-Scale Pilot (LSP) based on the work done in Task 4.2.

This report presents the results of the first MediaLSP), implemented by Freedom House Romania, which evaluated the AI Intelligent Coaching (https://fcompanion.lab.atc.gr/) application designed to support journalists in identifying disinformation, logical fallacies, and hate speech. The pilot involved journalists from across Romania, ensuring diverse perspectives on the use of AI in real newsroom environments. The main objective was to assess how AI can assist journalistic work while maintaining accuracy, credibility, and ethical standards. The tool was tested both in structured workshops and during daily professional activities. Through a broad number of users, this task provided a contextual assessment of LLM trustworthiness characteristics and the relative risks, impacts, costs, and benefits.

Furthermore, during T4.2, the in-isolation assessment was successfully completed through the dry-run process, encompassing the full six-phase trustworthiness evaluation of all AI assets of the AI Intelligent Coaching application in accordance with the FAITH AI TAF V1 methodology. This process contributed to the enhancement of the FAITH Golden Table through the inclusion of LLM-oriented threats, vulnerabilities, and mitigation actions, an activity that remains continuously evolving throughout the project lifecycle.

In parallel, all relevant models and datasets were registered and comprehensively documented within the FAITH Model Hub. The Trustworthy AI maturity levels of the participating stakeholder groups were also assessed using TrustSense, providing additional input to the overall risk evaluation process. Subsequently, the in-isolation trustworthiness assessment was conducted using TrustGuard, leading to the identification and analysis of the various risks associated with the AI Intelligent Coaching application across the different trustworthiness dimensions.

In addition, both formal and informal feedback was collected and submitted for all three FAITH tools, namely the FAITH Model Hub, TrustGuard, and TrustSense, contributing to their ongoing refinement and improvement.

Table of Contents

1 Introduction 9
1.1 Structure of the document 10
2 Progress on the AI system deployment and execution of the Pilot 12
2.1 AI System under test 12
2.1.1 Claim Analysis 13
2.1.2 Fallacy Detection 15
2.1.3 Hate Speech Detection 16
2.1.4 Technical Implementation 17
2.2 Status of the 1st Pilot Execution 19
2.2.1 Deployment Context & Domain Specifics 22
2.2.2 Training 23
2.2.3 Users 24
2.2.4 Evaluation Process 25
2.3 Status of the Replication Pilot preparation/initialization 27
2.4 Deviations from the original deployment plan 27
2.5 Data collection throughout the 1st Pilot execution 28
2.6 Integration into LSP workflows: Lessons learned 29
3 Application of FAITH Tools 34
3.1 Application of the FAITH AI_TAF V1 methodology 34
3.1.1 Phase 1: Cartography 35
3.1.2 Phase 2: Threat Analysis 38
3.1.3 Phase 3: Impact Assessment 55
3.1.4 Phase 4: Vulnerability Assessment 58
3.1.5 Phase 5: Risk Analysis 64
3.1.6 Phase 6: Countermeasures 66
3.2 Tool Validation Planning & Training Processes 68
3.2.1 Dry-run 68
3.2.2 Plan development 71

3.2.3 Tool Training 72
3.2.4 Data & AI model Passport in the FAITH AI Model Hub (ATC) 74
3.2.5 TrustGuard 76
3.2.6 TrustSense 79
3.3 Threat Analysis 84
3.4 Initial Risk Assessment results 85
3.5 Countermeasures 86
3.6 Formal Tool Feedback 89
3.6.1 Data & AI model Passport in the FAITH AI Model Hub 89
3.6.2 TrustGuard 89
3.6.3 TrustSense 91
4 AI System Evaluation 93
4.1 Detailed user feedback 93
4.1.1 TRUSTWORTHINESS PROPERTY: SECURITY 94
4.1.2 TRUSTWORTHINESS PROPERTY: VALIDITY & RELIABILITY (ACCURACY) 94
4.1.3 TRUSTWORTHINESS PROPERTY: VALIDITY & RELIABILITY (ROBUSTNESS) 95
4.1.4 TRUSTWORTHINESS PROPERTY: PRIVACY 97
4.1.5 TRUSTWORTHINESS PROPERTY: FAIRNESS WITH HARMFUL BIAS MANAGED 97
4.1.6 TRUSTWORTHINESS PROPERTY: TRANSPARENCY 99
4.1.7 TRUSTWORTHINESS PROPERTY: ACCOUNTABILITY 99
4.1.8 TRUSTWORTHINESS PROPERTY: INTERPRETABILITY 100
4.2 Trustworthiness Assessment 101
4.2.1 Validation & Reliability 101
4.2.2 Fairness 102
4.2.3 Privacy 102
4.3 Preliminary Impact Evaluation 103
5 Conclusions and Roadmap for the Final Pilot Phase 105
5.1 Summary of achievements in the second period 105
5.2 Plan for Replication Phase 105
6 References 107

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
AI RMF	FAITH AI Risk Management Framework
AI_TAF	Artificial Intelligence Trustworthiness Assessment Framework
ATC	Athens Technology Center
AI Act	Artificial Intelligence Act
EU	European Union
FAITH	Fostering Artificial Intelligence Trust for Humans
FH	Freedom House
GDPR	General Data Protection Regulation
ISO	International Standardization Organization
LLM	Large Language Model
LSP	Large Scale Pilot
LSP1	Large Scale Pilot 1 (Media)
OES	Operator Essential Services
SUT	System Under Test
TRL	Technology Readiness Level
VMS	Video Management System

List of Tables

Table 1. Dates and participant numbers for workshops conducted as part of the first Media Large Scale Pilot (LSP) 20
Table 2. Threats' frequencies relevant to the Media LSP 55
Table 3. Risk Calculator Matrix (D2.1) 65
Table 4. Final Risk estimation-2 (D2.1) 66
Table 5. TrustSense dimensions with corresponding levels and scores 80
Table 6. Threats relevant to the four functionalities of the AI Intelligent Coaching service. 84
Table 7. Countermeasures denoted as "Treated" in TrustGuard. 86
Table 8. Risk levels of the trustworthiness dimension "Validation & Reliability" per associated threat, for each functionality of the system. 101
Table 9. Risk levels of the trustworthiness dimension "Fairness" per associated threat, for each functionality of the system. 102
Table 10. Risk levels of the trustworthiness dimension "Privacy" per associated threat, for each functionality of the system. 103

List of Figures

Figure 1. Claim Detection. The Intelligent Coaching application detects multiple claims. Upon selection, the decontextualized claim appears in a pop-up window. The user then can select to analyse the claim either by selecting the “Match to debunked Claims” or the “Match Claims to Other Articles” functionality. 13
Figure 2. The “Match to Debunked Claims” functionality, returns the decontextualized claim, the URL of the stored fact-checking report and (if available) a short summary. 14
Figure 3. The “Match to Other Articles” functionality returns the decontextualized claim and a brief report generated from the detected resources, along with their URL and a matching score. 14
Figure 4 The “Logical Fallacy Detection” functionality returns the parts of the text that contain certain types of fallacy, along with their type. 15
Figure 5. Hate Speech Detection Result. 16
Figure 6 Workshop session conducted on 09.10.2025. 20
Figure 7 Workshop session conducted with journalists from the PRESShub network on 03.02.2026. 21
Figure 8 Workshop session held at the Journalism Department of Ovidius University of Constanța on 16.03.2026. 21
Figure 9 Workshop session held at the Media, Communication, and Journalism Department of Alexandru Ioan Cuza University of Iași on 18.03.2026. 22
Figure 10 FAITH AI_TAF V1 Risk Assessment Phases according to D2.1 34
Figure 11 Dry-run exercise for the Hate Speech detection tool. 69
Figure 12 Overview of threats, vulnerabilities, and mitigation actions identified during the Media LSP dry-run 71
Figure 13 Datasets used for the Hate Speech Detection Model in AI Data Passport. 74
Figure 14 Models of Claim Analysis functionality in AI Model Passport. 75
Figure 15 Cartography phase in TrustGuard for the Claim Detection Model. 76
Figure 16 Consequence Assessment in TrustGuard in Claim Detection filled in by FH. 77
Figure 17 Vulnerability Assessment for the Claim Detection model. 78
Figure 18 Risk Assessment and Mitigation Actions suggested by TrustGuard. 78
Figure 19 TrustSense assessment results for the AI User Team domain (FH). 81
Figure 20 Mitigation actions per category, showing scores and levels for Proactivity, Threat Awareness, and Responsibility and Ethics. 81
Figure 21 Mitigation actions per category, showing scores and levels for Proactivity, Threat Awareness, and Responsibility and Ethics. 82
Figure 22 Mitigation actions per category, showing scores and levels for Resilience, Collaboration, Integrity, and Technical Proficiency 82
Figure 23 Mitigation actions per category, showing scores and levels for Problem Solving, Resource Accessibility, Policy Adherence 83

Figure 24 Mitigation actions per category, showing scores and levels for Motivation and Commitment, and Privacy and Compliance..... 83
Figure 25 Aggregated user ratings of trust in the security of the AI Intelligent Coaching application 94
Figure 26 Aggregated user ratings of the AI system's Claim Analysis performance 95
Figure 27. Aggregated user ratings of the AI system's Hate Speech Detection performance 95
Figure 28 Aggregated user ratings from the feedback forms regarding the consistency of the AI system's Claim Analysis results..... 96
Figure 29. Aggregated user ratings from the feedback forms regarding the perceived consistency of the Hate Speech Detection component. 96
Figure 30. Aggregated user ratings from the feedback forms regarding their trust in the AI system's handling of data privacy..... 97
Figure 31. Aggregated user ratings on the inclusiveness of the Claim Analysis component. 98
Figure 32. Aggregated user ratings on the inclusiveness of the Hate Speech Detection component. 98
Figure 33. Aggregated user ratings on the clarity of the AI nature of the system. 99
Figure 34. Aggregated user ratings regarding the clarity of the responsibility for the AI system's outputs..... 100
Figure 35. Aggregated user ratings on the intelligibility of the Claim Analysis results. 100
Figure 36. Aggregated user ratings on the intelligibility of the Hate Speech Detection. 101

1 Introduction

The first pilot iteration in the Media Large Scale Pilot (LSP) focuses on testing and assessing the AI Intelligent Coaching application, a Large Language Model (LLM)-based service that aims to assist media professionals in assessing the validity of the articles they read or write. This pilot is run by Freedom House Romania and involves a total of 75 journalists from various regional newsrooms across Romania, contributing to a coordinated, large-scale evaluation effort. This cohort includes young journalists, journalism professors, and freelance journalists. This variety of users allowed the pilot to record a variety of viewpoints regarding AI usability, reliability, and ethical considerations. The pilot aims to offer evidence-based insights to direct the adoption of reliable AI in journalism and explores how AI may be used responsibly while maintaining credibility in actual journalism and fact-checking processes.

On digital platforms, disinformation spreads quickly, often faster than it can be disproven. AI can assist by spotting trends in the spread of messages, spotting coordinated campaigns, and grouping related content from different languages and nations. However, according to the Joint Research Center of the EU Commission[1], in order to effectively counteract disinformation, public institutions must communicate in a transparent and trustworthy manner; civil society must be involved; individuals must be media and AI literate; and **research** and policy must continue to optimize the advantages of AI while reducing its risks[2].

In this context, to ensure that detrimental biases are appropriately addressed, the purpose of this pilot is to specifically evaluate the AI system against key trustworthiness attributes. These are safety, security and resilience, explainability and interpretability, privacy, fairness, accountability and transparency, validity, and reliability. Both planned activities, such as webinars, and independent exploration during professional activity, were used by participants to assess the AI Intelligent Coaching application. They investigated the system's various features, including Claim Analysis and Hate Speech Detection.

Within this phase, journalists' inputs were essential because they offered context and useful views on how AI impacts their work. They assisted in defining the limits of AI use, and encouraged conversations on the conflicts that arise between technology and human judgment, examining how AI enhances, or occasionally contradicts, human intelligence and decision-making.

In addition to providing a solid foundation for evaluating the AI tool's usability, dependability, and practical usefulness in journalistic workflows, the pilot is designed to produce both

[1] https://joint-research-centre.ec.europa.eu/jrc-explains/ai-friend-or-foe-disinformation_en

[2] https://joint-research-centre.ec.europa.eu/jrc-explains/ai-friend-or-foe-disinformation_en

qualitative and **quantitative findings** that guide future developments. These findings will contribute to improving the system and mitigating identified risks, ensuring that the AI Intelligent Coaching application is ready to progress to the next piloting phase (T4.3). Furthermore, the implementation of the Initial Pilot and the integration of the AI Intelligent Coaching application into journalistic workflows yielded a series of lessons learned regarding the Media LSP implementation, which will inform the preparation of the Replication Pilot. In this subsequent phase, an additional group of 10 journalists at national level will further assess the tool under conditions of responsible and reliable use.

Journalists used the AI coaching application to examine and report incidents of disinformation, providing insightful information about the system's operation and the risks they encountered. While senior journalists offer critical assessments based on years of experience, early findings show that young journalists are generally more open to AI tools, ensuring that the system's development takes into account a variety of user needs and practical limits, ultimately supporting a comprehensive assessment of risks, impact, costs, and benefits associated with the AI application.

In addition to the user-centered pilot activities, Task 4.2 involved the application and formal evaluation of the FAITH AI_TAF V1 methodology and the FAITH tool prototypes (TrustGuard, TrustSenses, and FAITH Model Hub) to facilitate the iterative refinement of the FAITH ecosystem. Finally, within Task 4.2, the Media LSP partners established a roadmap for the Final Pilot Phase, transitioning toward a holistic risk management and a comprehensive socio-technical system trustworthiness assessment based on FAITH AI_TAF V2.

1.1 Structure of the document

The document is structured as follows:

Chapter 1 (Introduction): introduces the Media Large Scale Pilot (LSP), outlining the scope and objectives of the pilot focused on the AI Intelligent Coaching application, the involvement of journalists, and the key trustworthiness attributes guiding the evaluation.

Chapter 2 (Progress on the AI system deployment and execution of the Pilot): describes the deployment and execution of the AI Intelligent Coaching application, including its functionalities, technical setup, user engagement, training activities, and integration into journalistic workflows, as well as data collection and lessons learned.

Chapter 3 (Application of FAITH Tools): presents the application of the FAITH AI_TAF V1 methodology, covering all risk assessment phases and the use of tools such as TrustGuard, TrustSense, and the AI Model Hub, along with identified risks and mitigation actions.

Chapter 4 (AI System Evaluation): provides the user-centric evaluation of the system across key trustworthiness dimensions, combining quantitative results and qualitative feedback, and includes a preliminary assessment of system impact and performance.

Chapter 5 (Conclusions and Roadmap for the Final Pilot Phase): summarizes the main outcomes of the pilot, outlines the plan for the replication phase, and highlights key risks and challenges for future development and adoption.

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

2.1 AI System under test

The AI system under evaluation, namely the AI Intelligent Coaching application, is described in detail in Deliverable D4.1. It consists of three main functionalities: 1) claim analysis, 2) logical fallacy detection, and 3) hate speech detection. A user can assess an article by selecting one or multiple functionalities at the same time.

Building on the feedback collected from end-users during the pilot activities, the user interface (UI) has undergone significant refurbishment in order to improve usability, accessibility, and overall user experience.

In addition, the functionality of the Intelligent Coaching application has been further enhanced so that it is now capable of detecting multiple claims within an input text, rather than focusing solely on the main claim. This improvement better reflects the complexity of real journalistic content and strengthens the system's applicability in disinformation analysis scenarios.

Furthermore, support for the Romanian language is being continuously refined and improved through iterative user testing and feedback cycles, ensuring that the application becomes increasingly aligned with the linguistic and contextual needs of the participating journalists and regional newsrooms.

For completeness, a brief description of the various functionalities of the AI Intelligent Coaching application is provided in this section.

2.1.1 Claim Analysis



Figure 1. Claim Detection. The Intelligent Coaching application detects multiple claims. Upon selection, the decontextualized claim appears in a pop-up window. The user then can select to analyse the claim either by selecting the "Match to debunked Claims" or the "Match Claims to Other Articles" functionality.

Claim Analysis: This functionality consists of 3 modules: 1) The "Claim Detection" module (Figure 1), which detects the core claims in the text and decontextualizes them (i.e., it adds from the input article all required information for the claim to stand on its own), 2) the "Match to Debunked Claims", which searches the locally stored database with fact-checks if there is a fact-checking report that assesses a claim *semantically similar* to the detected claim selected for analysis and generates a short report based on the detected fact check (Figure 2), 3) the "Match to Other Articles", which searches sources in the Web that debunk the detected claim selected for analysis and generates a short report based on the detected sources (Figure 3).

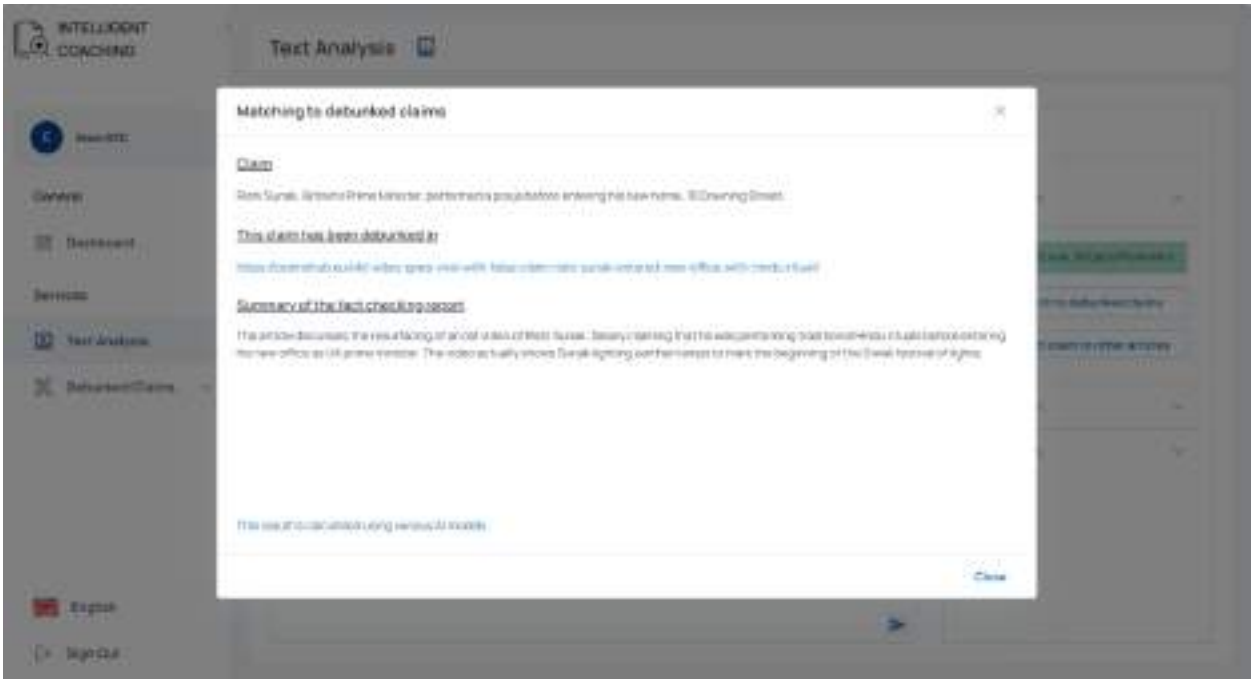


Figure 2. The "Match to Debunked Claims" functionality, returns the decontextualized claim, the URL of the stored fact-checking report and (if available) a short summary.

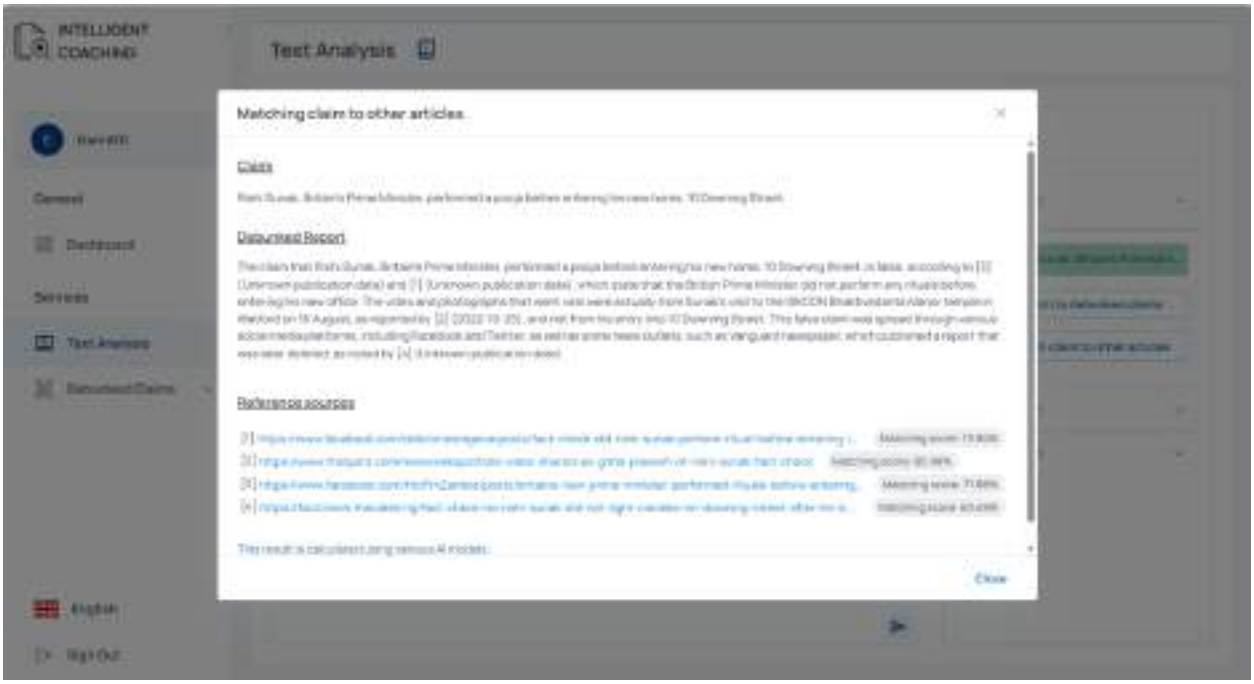


Figure 3. The "Match to Other Articles" functionality returns the decontextualized claim and a brief report generated from the detected resources, along with their URL and a matching score.

2.1.2 Fallacy Detection

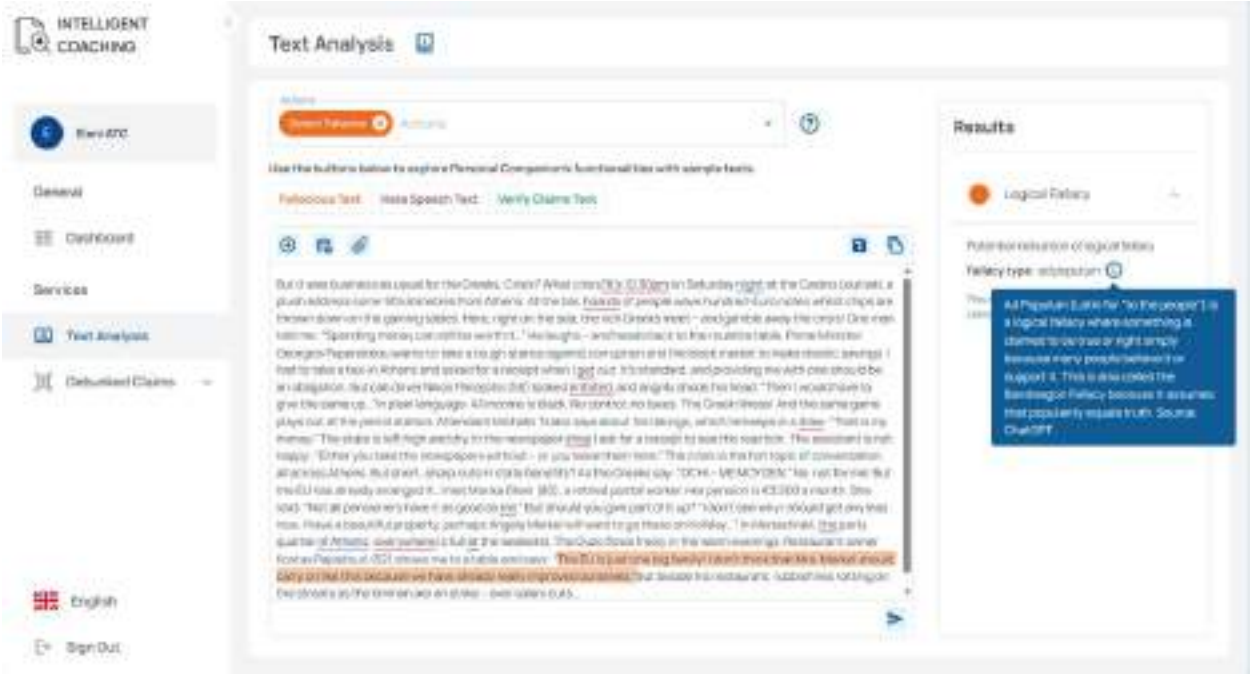


Figure 4 The "Logical Fallacy Detection" functionality returns the parts of the text that contain certain types of fallacy, along with their type.

Logical Fallacy Detection: It is a classifier that takes as input an article and checks whether it contains logical fallacies. It highlights the fallacious sentence and returns the logical fallacy type and, upon request, its meaning (Figure 4). The AI Intelligent Coaching application detects mainly the following 12 types of fallacies:

- **Ad hominem:** an excessive attack on an individual or a group instead of addressing someone's argument or opinion.
- **Ad populum:** claiming a truth or affirming something is good or correct because many people think so.
- **Circular reasoning:** a premise explicitly or implicitly assumes that its conclusion is true rather than providing any supporting statements.
- **Deductive fallacy**: the text has an error in the formal logical reasoning.
- **Equivocation:** a word or phrase is used in one sense in one premise and in another sense in some other needed premise or in the conclusion.
- **Fallacy of credibility:** an appeal is made to some form of ethics, authority, or credibility.
- **Fallacy of extension:** an argument that attacks an exaggerated or caricatured version of your opponent's position.

- **False dilemma:** presenting only two options or sides when there are many options or sides.
- **False causality:** assuming that two correlated events must also have a causal relation.
- **Faulty generalization**: drawing a broad conclusion based on a limited sample of population.
- **Intentional fallacy:** an argument has some element that shows the intent of a speaker to win an argument without actual supporting evidence.
- **Red herring (fallacy of relevance):** the argument supporting the claim diverts the attention to issues which are irrelevant to the claim at hand.

2.1.3 Hate Speech Detection

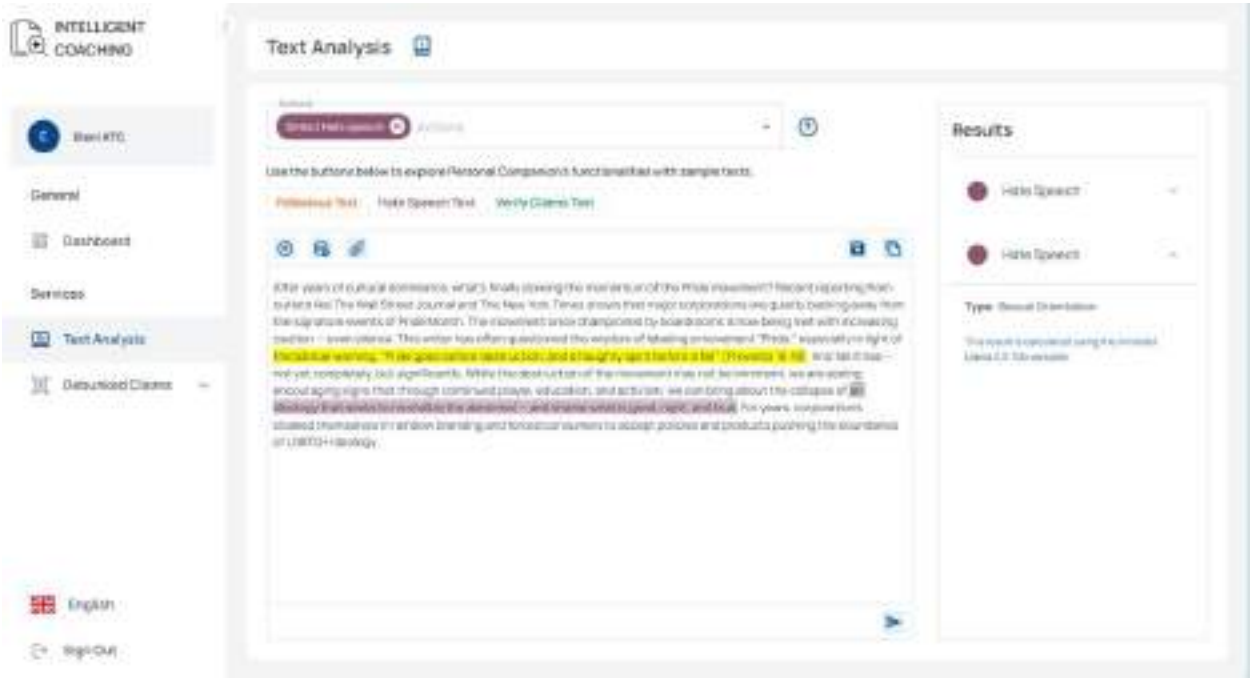


Figure 5. Hate Speech Detection Result.

Hate speech detection: Takes as input an article and checks if it contains certain forms of hate speech. For this task, the definition of hate speech follows the framework provided by the United Nations. According to this definition, hate speech is understood as:

"Any form of communication—spoken, written, symbolic, or digital—that incites hatred, discrimination, or violence against a specific group or individual based on protected characteristics. These characteristics include race, ethnicity, nationality, disability, religion or

belief system, gender, gender identity, or sexual orientation." The hate speech detector classifies a text into the following classes:

- No Hate: Text does not contain any hate speech.
- Racism: Text contains language that expresses hatred based on race, ethnicity, or disability. (Note: Disability was included in this category based on the annotation guidelines of the evaluation datasets.)
- Sexism: Text contains language that expresses hatred based on gender.
- Religious: Text contains language that expresses hatred based on religion.
- Sexual Orientation: Text contains language that expresses hatred based on sexual orientation.

The hate speech detector highlights the sentence with hate speech and returns the hate speech type (Figure 5).

2.1.4 Technical Implementation

To better understand the vulnerabilities, threats, and risks that the AI system may be exposed to, next, we provide the technical implementation of the modules assessed in this deliverable.

Claim Detection: It consists of a pipeline of transformations (prompt chains) on the input text using LLMs, to finally derive the text snippets containing the claims. This pipeline consists of the following steps:

1. **Summarization**: initially, the LLM is instructed to summarize the input text. To produce this summary, the LLM is prompted to focus on verifiable facts, figures, and data, analyse the sender's main intention, and focus on keyword indicators carrying domain-specific information (in line with the guidelines presented in Section 3). The summarization step is useful in the case of larger texts, reducing the context window for subsequent tasks, and enabling the focus on the most important details included in the text.
2. **Claim detection from the Summary:** the LLM is provided the previously generated input text summary and is instructed to write the claims included in it, along with the same guidelines provided for the summary generation.
3. **Merge/Rank Claims by Importance:** in this step the LLM is presented with the original input text along with the previously extracted claims. The instruction given aims to enhance the claims with any missing pieces of information, to make them valid standalone sentences (decontextualization), and to merge semantically overlapping ones. Additionally, the LLM is instructed to write the resulting claims in order of importance.

4. **Match Claims to Snippets:** in this final step the LLM is presented with the claims produced in step 3, along with the original input text. The instruction is to provide for each claim the passage in the input text, verbatim, where the claim is expressed.

Match to Debunked Claims. The claim matching service consists of a vector DB (which is the transformation of the Debunked Claims database into an embedding space), an API, and an AI Agent developed using LangGraph and LangChain. For a given claim, the agent first communicates with the vector DB to retrieve a set of top-k most relevant documents. Note, there is no similarity threshold cutoff, i.e., the most similar documents to the claim are retrieved, regardless of whether they actually are related to the claim. Following this, the LangGraph Agent decides on the relevance of the retrieved documents and produces a judgement about whether the examined claim is debunked based on the documents found.

Match to Other Articles. The implemented workflow expects a claim (i.e., a sentence) as its input, then performs web search on various search engines and web resources, based on this claim and on search keywords produced by generative AI. Then, it categorizes the collected results into different classes leveraging LLMs as classifiers, an approach that has been popular in the literature recently and has been shown to sometimes outperform specialized classifiers (Wang, Pang, & Lin, 2024). In particular, the web resources collected are categorized into the following classes: 'debunking' (web resources that debunk the claim), 'supporting' (web resources which support the claim), and 'related' (web resources may only partially address the claim, or relate to the claim in other ways, covering other similar topics, etc.).

Hate Speech Detection. It is a classification task split into two sequential prompts for improved accuracy:

- **Hate Speech Presence Detection Prompt**: A binary classification few-shot example prompt instructs the model to determine whether the input contained hate speech ("Hate" vs "No Hate").
- **Hate Speech Classification Prompt**: If the first LLM call detects hate speech, a second few-shot example prompt is issued to classify the hate speech into the appropriate category (Racism, Sexism, Religious or Sexual Orientation). Multiple hate categories can be detected within a single input; each linked to one or more specific text snippets. The prompt also instructs the model to highlight the specific phrases or sentences that triggered the detection. This approach raises the number of LLM calls to two per input if hate speech is detected in the first call. Otherwise, if no hate speech is detected, the calls per input are limited to one.

This two-step design aims to reduce false positives during category classification by first confidently filtering non-hate content. It also allows for more focused instructions in the Hate Speech Classification Prompt, tailored specifically to category identification, leading to improved overall classification accuracy.

2.2 Status of the 1st Pilot Execution

To take part in the pilot activities, a core working group of approximately 10-15 journalists from PRESShub has been formed. These individuals were the platform's most regularly active users, and they have actively participated in testing and offering input on its features. In order to increase participation and guarantee a broader diversity of viewpoints in the evaluation process, the group was eventually expanded to include an additional of approximately 60 journalists. In total, overall 75 journalists were involved in assessing the Intelligent Coaching application during the initial pilot phase.

The AI Intelligent Coaching tool was used by the journalists both individually in their professional activity and during organized webinars that were arranged as part of the pilot activities. In a number of instances, users used the tool in the context of their own work, creating independent reports that detailed potentially troublesome stories and the problems found throughout the analytic process. Although the platform is generally seen as accessible and user-friendly, several participants observed issues with the dependability of specific outputs, especially when applied to more complex or confusing scenarios, according to preliminary findings. As the pilot unfolded, these results provided insights for system improvements and provided useful insight for the ongoing evaluation process.

In addition to senior journalists, the pilot has also been extended to include junior journalists as well.

In addition to the input received from professional participants, this interaction offers more viewpoints on usability, accessibility, and the usefulness of the AI system inside journalistic processes.

Workshop date	**Number of participants**
03.07.2025	9
24.09.2025	16
06.10.2025	9
09.10.2025	12
03.02.2026	7
05.02.2026	5

09.02.2026	8
12.02.2026	15
17.02.2026	14
02.03.2026	9
11.03.2026	16
11.03.2026	9
12.03.2026	8
12.03.2026	11
16.03.2026	16
18.03.2026	11
24.03.2026	19

Table 1. Dates and participant numbers for workshops conducted as part of the first Media Large Scale Pilot (LSP)

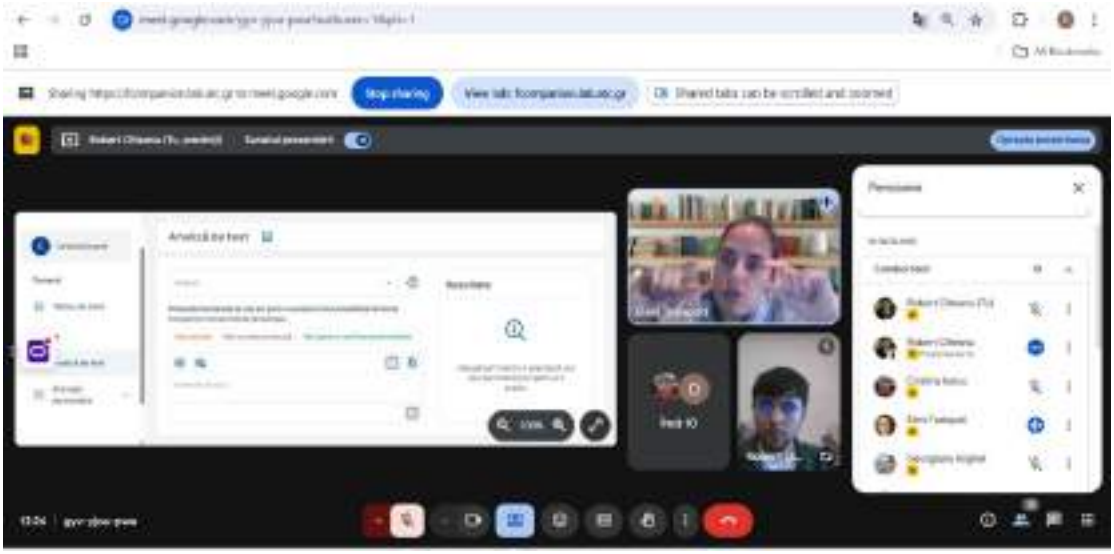


Figure 6 Workshop session conducted on 09.10.2025.

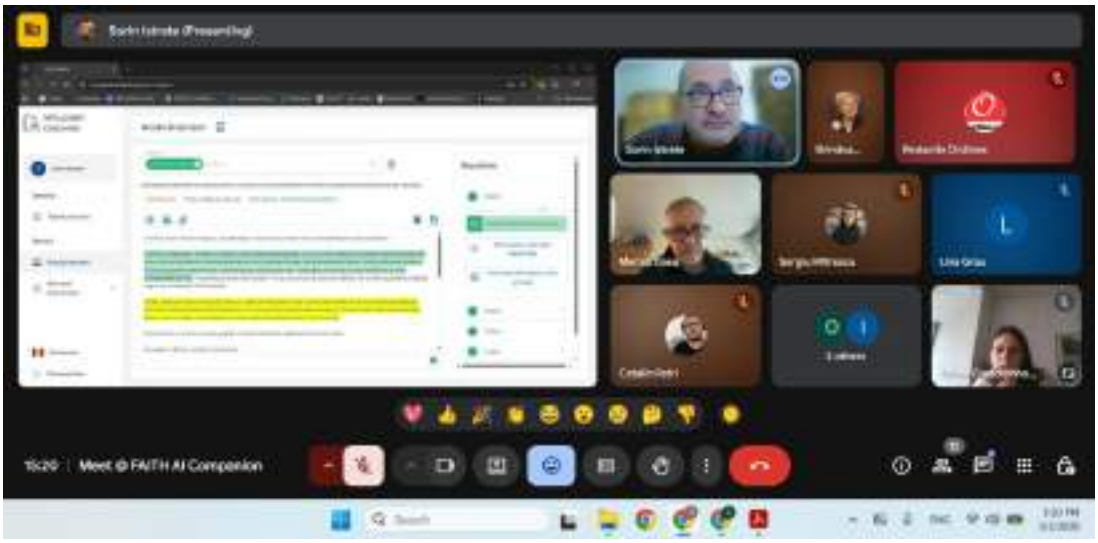


Figure 7 Workshop session conducted with journalists from the PRESShub network on 03.02.2026.

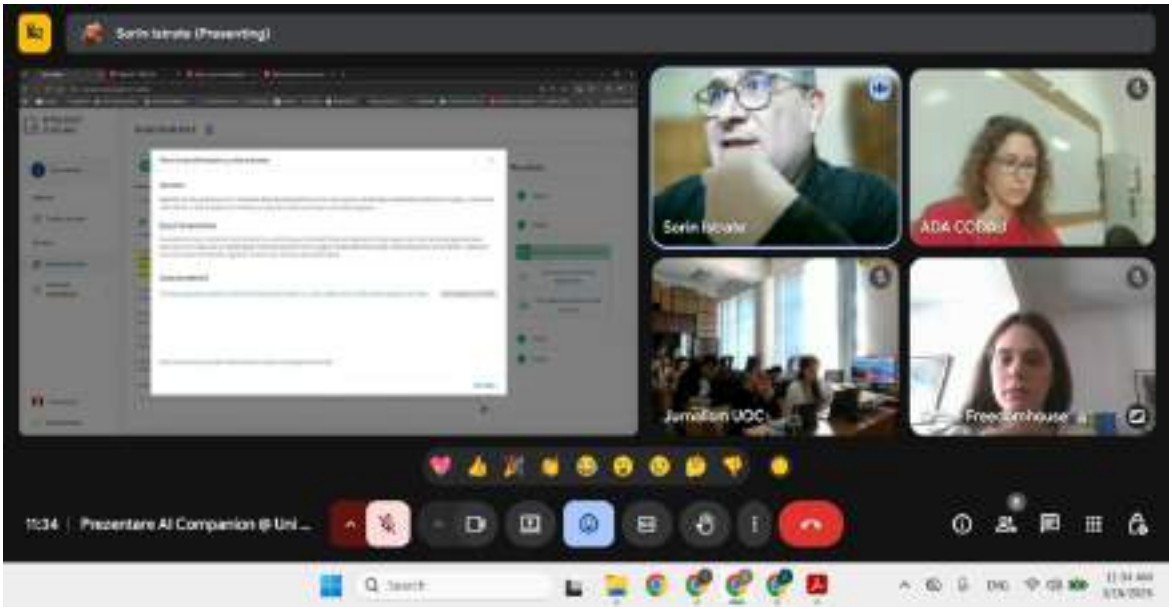


Figure 8 Workshop session held at the Journalism Department of Ovidius University of Constanța on 16.03.2026.

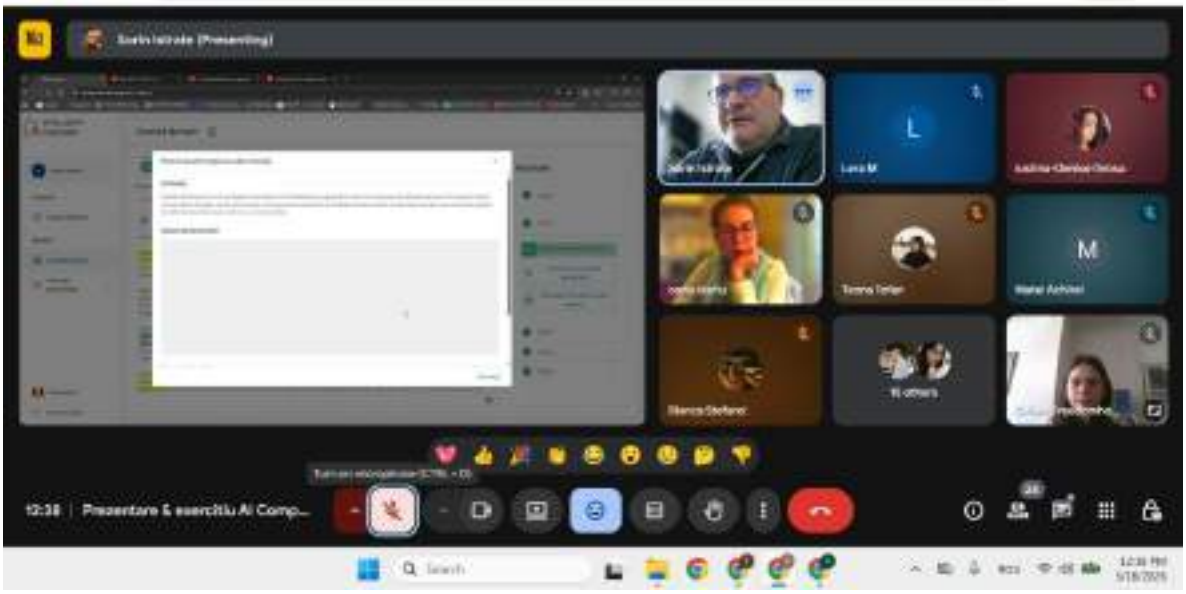


Figure 9 Workshop session held at the Media, Communication, and Journalism Department of Alexandru Ioan Cuza University of Iași on 18.03.2026.

2.2.1 Deployment Context & Domain Specifics

The first Media Large Scale Pilot (LSP) evaluated an AI Intelligent Coaching application that is a web-based platform that can be accessed using a typical web browser. Consequently, there was no need for user-level technical deployment or installation. The application was accessible remotely by participating journalists from their usual workplaces, such as newsroom offices or home offices. Throughout the pilot activities, this strategy allowed for flexible involvement and ongoing system interaction. The pilot project was initially carried out within the PRESShub media network but has since expanded to other local and regional media outlets in the country.

The professional media environment in which the participating users work is marked by a high volume of information flows, time-sensitive editorial processes, and an expectation to quickly validate and contextualize content. Thus, journalists are regularly exposed to contradictory narratives, possibly deceptive information, and situations of suspected disinformation in this setting.

The nature of journalistic work, which entails processing vast amounts of information, confirming sources and facts, spotting possible misinformation or disinformation, and making decisions under extreme time pressure, makes the media industry a pertinent setting for testing AI-based support tools. AI tools can help with information analysis and verification in this situation. The pilot's findings, however, show that although the AI system can help journalists with some tasks, human judgment and experience are still superior to automated analysis in some circumstances, underscoring the ongoing significance of human oversight in journalistic workflows.

The AI Intelligent Coaching application is largely used in the pilot as a support tool for the study of potentially problematic media content, such as articles or information that may contain aspects of disinformation and hate speech.

Based on large language model (LLM) capabilities, the system checks if the article provided by the user contains claims that have already been debunked or hate speech. While retaining complete editorial control over the interpretation and application of the produced outputs, this technique enables users to investigate various analytical viewpoints.

Journalists used the system both on their own as part of their work and during planned pilot events like webinars. Participants used the tool in the context of their own journalistic investigations or articles twice per month, recording the results and identifying problems that occurred throughout the analysis process. The pilot was able to approximate real-world conditions of AI use in the media sector thanks to the web-based platform and the dispersed participation of journalists. Additionally, it contributed to a systematic collection of feedback on the system's usability, reliability, and accuracy in supporting disinformation and hate speech analysis. These findings assist in evaluating the AI system's reliability and usefulness in the media industry.

2.2.2 Training

To ensure the effective onboarding and preparedness of all participating end-users, multiple parallel training activities were implemented to support familiarisation with the AI Intelligent Coaching application and its intended use within the pilot.

First, a comprehensive User Guide was developed by ATC and provided to FH. This guide provided clear instructions on system access, core functionalities, workflow steps, and recommended practices for using the application during the pilot phase.

In parallel, ATC prepared and delivered a recorded demonstration of the Intelligent Coaching Application, presenting its key features, operational logic, and practical use cases. This demonstration enabled participants to gain an initial understanding of the platform's capabilities and its role in supporting the identification and assessment of potential disinformation cases.

To further facilitate hands-on familiarisation, sample texts and example disinformation scenarios were provided by ATC. These materials allowed users to explore the service in a controlled environment and become acquainted with the types of inputs, generated question-and-response schemes, and expected outputs before engaging with real-world cases.

A "train-the-trainers" approach was adopted to ensure efficient knowledge transfer and scalable capacity building across the pilot network. In this framework, ATC first trained the FH project team, equipping them with both technical and methodological knowledge regarding the application and the objectives of the pilot.

Subsequently, the FH project team, together with a senior journalist from PRESShub's network, delivered dedicated training sessions to the participating journalists. These sessions were primarily organised as online webinars and live demonstrations, allowing broad participation from the regional newsrooms across Romania.

During the training sessions, participants were introduced not only to the functional features of the AI Intelligent Coaching application, but also to the broader objectives of the pilot, including the assessment of trustworthiness characteristics, responsible AI use in journalism, and the identification and mitigation of risks related to validity, reliability, bias, transparency, privacy, and accountability.

This structured training process ensured that all participating journalists were adequately prepared to use the application consistently and responsibly throughout the pilot period.

At this stage, the users were asked to assess the Claim Analysis and Hate Speech Detection features. A smaller working group of journalists who are regularly engaged by the project team were used to test the system at the beginning of the pilot. This method enabled the FH team and the journalists to have a greater focus on the platform's functionality and made it easier to spot mistakes and usability problems that occurred during testing.

During the initial training sessions, participating journalists actively engaged with the platform, providing valuable practical insights regarding operational workflows, such as refining error-reporting procedures and managing data accuracy within the system. These discussions extended to collaborative strategies for database curation and user-led content enhancement. A central theme of these exchanges was the strengthening of human-in-the-loop oversight, ensuring that the technology effectively complements editorial judgment even in complex edge cases.

To address these constructive inputs and support the testing phase, additional interactive webinars were organized. These sessions provided a forum for technical partners (ATC) to reinforce the pilot's purpose: utilizing the current prototype as a dynamic foundation for user-centered assessment and enhancement of the system trustworthiness. This approach ensured that the system's capabilities are being directly aligned with professional journalistic standards and needs through real-world user feedback.

Participants were encouraged to actively test the application, submit pertinent articles for review, and record any problems they ran into. The feedback gathered supports the system's iterative improvement and helps assess its dependability and usefulness inside journalistic workflows, especially with regard to hate speech detection and fact-checking features.

2.2.3 Users

The group of users involved in the first Media Pilot includes journalists from PRESShub's network and from other local and regional newsrooms, such as SangeorzTV, ImpactTV, Bihoreanul, as well as freelance journalists collaborating with online media platforms (e.g., Panorama, HotNews, Scoala9). Their areas of specialization are very varied, e.g. politics, geopolitics, foreign policy, and

social issues.
Participants include investigative journalists, reporters, editors, and media professionals, some of whom teach journalism and media studies at universities while also freelancing. While some participants are journalists in their early or mid-career stages, others have senior editorial positions and considerable expertise in the field. The journalists taking part in the pilot are in charge of testing the AI Intelligent Coaching application. Their responsibilities include evaluating possible instances of disinformation via the platform and offering comments on how well the system performs in terms of trustworthiness-related elements, including dependability, transparency, fairness, and usability.

Each participating journalist was required to use the platform to investigate at least two possible examples of misinformation each month as part of the initial pilot. The information acquired from these exercises helps assess the system's efficacy and possible incorporation into journalistic operations.

2.2.4 Evaluation Process

The evaluation process of the Initial Pilot was designed as an iterative, user-centred assessment procedure, aligned with the trustworthiness dimensions defined by the FAITH AI_TAF V1. Its primary objective was to collect structured feedback from participating journalists regarding the perceived trustworthiness of the Intelligent Coaching application in real-world media workflows.

At the beginning of the pilot, an initial evaluation instrument was developed and deployed as a Google Form. At the beginning of the pilot, an initial evaluation instrument was developed and deployed as a Google Form. This first version was intentionally comprehensive and closely mapped to the trustworthiness characteristics foreseen in the task description, including safety, security and resilience, explainability and interpretability, privacy, fairness and bias management, accountability and transparency, and validity and reliability.

The first form was structured to collect both quantitative and qualitative feedback. In addition to demographic and professional background questions, participants were asked open-ended questions on whether AI is useful for countering disinformation and the extent to which they would trust such a system in their journalistic practice. The form is available in this link[3].

[3] https://docs.google.com/forms/d/e/1FAIpQLSdSi8xi_NFjVFmTFJ9LiAvgf6rhkdhP_U-fgR9X13p_UZaE5Q/viewform?usp=sharing&ouid=117555197365128641367

A major part of the questionnaire focused on the trustworthiness dimensions of the FAITH AI RMF. For each dimension, the journalists were provided with a simplified explanation in media-related terms and were then asked to reflect on:

- their interpretation of the concept,
- risks stemming from the absence of the respective property,
- functionalities that may compromise that property,
- suggestions for improvement.

For example, the form explicitly addressed issues such as whether the system could amplify misinformation or hate speech (safety), whether there was functionality in the AI system that could compromise its security (security and resilience), whether its outputs could be clearly understood (explainability), and whether the tool treated all groups fairly without amplifying stereotypes (fairness and bias).

While this first version was designed to provide rich and detailed feedback, the pilot participants reported that the questionnaire was too complex, lengthy, and overly technical, making it difficult to complete alongside their regular editorial duties. In particular, several questions were perceived as using terminology more suitable for technical experts than for end-users in the media domain.

Taking these comments into consideration, ATC, with the help of FH, revised the evaluation methodology and developed a second, simplified evaluation form. This revised version-maintained alignment with the trustworthiness objectives of T4.2 while significantly improving usability and response efficiency. The second form is available in this link[4].

The second form adopted a more user-friendly and operational structure, replacing long open-ended questions with simple Likert-scale ratings (1–5) and clearer language expressed in practical journalistic terms. Instead of asking users to explicitly reason about technical trustworthiness properties, the revised form focused on direct experiential questions such as:

> *"Rate how consistent the Claim Analysis results are. For instance, did it return the same results when used on the same articles?"*

4 https://docs.google.com/forms/d/e/1FAIpQLSdZ3fwSzaiZy_wVcGYlPuvnnxn5zWLoxwmAoWeZkl5sNLDlaw/viewform?usp=sharing&ouid=117555197365128641367

"Rate how inclusive the Hate Speech Detection is (1 = lowest, 5 = highest). Have you noticed any signs of bias or stereotyping (e.g., racism, sexism, homophobia)?"

This second version proved substantially more effective for the pilot context, as it reduced respondent burden while still enabling FH to gather actionable evidence on system trustworthiness. The questions included in the first version of the evaluation form, particularly those related to the risks arising from the compromise of the various trustworthiness characteristics, were subsequently addressed through informal discussions during the workshops conducted throughout the pilot. Furthermore, to facilitate broader participation and improve accessibility for the end-users, the second version of the evaluation form was also translated into Romanian.

Both evaluation forms were conducted anonymously to encourage open, honest, and unbiased feedback, without participants feeling constrained by concerns about sharing personal information or expressing critical views. This approach supports the collection of more reliable and representative user insights.

The transition from Evaluation Form v1 to v2 reflects the adaptive and human-centric methodology of Task T4.2, where user feedback was not only collected as evaluation data but also directly informed the refinement of the pilot processes themselves. This iterative approach ensured that the assessment framework remained aligned with the needs, capacities, and working realities of journalists, while continuing to support the broader objective of preparing the AI Intelligent Coaching application for the next piloting phase (T4.3) with identified risks properly documented and mitigated.

2.3 Status of the Replication Pilot preparation/initialization

The replication pilot's preparation phase is now in progress. To ensure continuity from the first pilot, journalists have already been identified and engaged to take part in this phase. The initial focus of the engagement has been on familiarizing journalists with the AI Intelligent Coaching program and setting up the process for reporting observations, cases of disinformation, and comments and feedback on the functionality and reliability of the system.

2.4 Deviations from the original deployment plan

No substantial deviations were identified regarding the scope of the use case, the participating users, or the overall pilot objectives.

A minor change from the initial deployment strategy was made during the initial pilot by including junior journalists with working experiences from six Romanian universities to participate in this process. Participating institutions included:

- University of Oradea
- Alexandru Ioan Cuza University of Iași
- Ovidius University of Constanța
- Universitatea Lucian Blaga from Sibiu
- University of Bucharest
- West University of Timișoara

This modification was made as an addition to the input received from senior journalists and to offer more viewpoints on the AI system's usability, trustworthiness, accessibility, and usefulness. According to preliminary findings, junior journalists are often more responsive and willing to interact with AI tools, incorporating them into their work with comparatively little difficulty. They have real-world expertise with fact-checking, source verification, and other essential journalistic procedures, which allows them to assess AI outputs critically and successfully incorporate the tool into actual reporting assignments. They are also actively researching media and AI's effects, which made it possible for them to see right away how applicable and helpful the AI coaching application is to their present work and future professional endeavours. They provide insightful assessments of the system's dependability, usability, and credibility by utilizing both adaptability and essential journalistic knowledge.

2.5 Data collection throughout the 1st Pilot execution

The data collection process during the execution of the initial pilot was designed to capture both structured evaluation data and qualitative experiential feedback from the participating journalists. In line with the objectives of Task T4.2, data were collected through three complementary channels, ensuring a comprehensive assessment of the AI Intelligent Coaching application from both operational and trustworthiness perspectives.

The first and most systematic source of data collection was the set of Google Forms used for the formal evaluation of the system (detailed in Section 3.7). These forms enabled the collection of structured feedback on the application's performance across the trustworthiness characteristics defined by the FAITH AI_TAF V1, including validity and reliability, safety, privacy, fairness, explainability, accountability, and transparency. Depending on the version of the form used, data were collected either through open-ended responses or through quantitative Likert-scale ratings, allowing both measurable indicators and descriptive feedback to be gathered. This channel provided comparable evidence across participants and supported the identification of recurring strengths and weaknesses in the system's behaviour.

A second important source of data was the workshop discussions held throughout the pilot period. These discussions provided a more interactive environment in which journalists could elaborate on their experiences, reflect on the risks associated with potential system malfunction (detailed in Section 3.1.3), and discuss operational concerns that may not have been fully captured through the formal questionnaires. In particular, the workshops were used to collect insights on the possible consequences of inaccurate outputs, missed detections, biased assessments, and other issues that could affect the trustworthiness and responsible use of the application in the media context. This qualitative feedback was especially valuable for understanding the broader editorial, ethical, and societal implications of system failures.

The third source of data collection consisted of concrete examples and use cases in which the system did not function as expected. Participating journalists were encouraged to document cases where the application produced incomplete, inaccurate, inconsistent, or otherwise problematic results. These examples included instances such as false detections, failure to identify relevant claims or harmful content, inconsistencies across similar inputs, and outputs perceived as unclear or difficult to interpret. Such case-based evidence played a crucial role in identifying specific technical shortcomings and informing targeted risk mitigation actions for the subsequent piloting phase.

2.6 Integration into LSP workflows: Lessons learned

- **Stakeholder mapping and engagement**

One key takeaway regarding stakeholder engagement is that mapping and involving of journalists should always be done in a timely manner, and that they should receive all the necessary information about the project and the tool before using it, as well as a disclaimer regarding the stage of development of the tool. Journalists were contacted via FH's existing professional networks, partnerships with media outlets, and connections to journalism departments and faculties at six Romanian universities. This strategy worked well to draw people and spark their interest in the use of AI technology in journalism. Since junior journalists from university institutions demonstrated a greater willingness to experiment with AI-supported technologies in journalistic work, their inclusion increased engagement levels and offered additional viewpoints.

The experience indicated that promoting participation requires early interaction with stakeholders and clear information about the pilot's goal. Reducing scepticism and promoting positive conversations regarding the use of AI in journalism were made possible by presenting the tool as a support system rather than a substitute for editorial decision.

- **Workshops and user engagement**

The Intelligent Coaching application was successfully introduced to participants through structured webinars, which also helped them through the testing process. In addition to giving participants the chance to familiarize themselves with the system's features, these meetings created an atmosphere in which users could freely discuss their experiences, questions, and concerns they might have regarding the ethical and practical issues of the application.

Participants were also better positioned to comprehend the capabilities and limitations of the system because of the **interactive style**. In a number of instances, users stated that guided examples and hands-on exercises enhanced their capacity to use the platform autonomously in their regular journalistic work.

- **Feedback collection and evaluation process**

The gathering of systematic feedback is another important lesson. The project team was able to collect both qualitative and quantitative information about the system's usability, dependability, and perceived trustworthiness by using feedback forms combined with workshop discussions.

Regarding technical problems, analytical constraints, and user experience, participants offered helpful feedback. In order to find recurrent themes, such as issues with content loading, claim recognition coverage, and handling longer texts, it was very helpful to combine **direct user input** with **formal assessment questionnaires**. These observations are necessary for guiding additional system enhancements and improving the assessment process in next stage of the project.

- **Integration into journalistic workflows**

The AI Companion tool can be incorporated into journalistic workflows mainly as a support tool for **preliminary content screening** and structured analysis, as this pilot phase showed. In general, participants found the interface to be **user-friendly** and **accessible**, which made it possible for them to integrate the tool into their current fact-checking and content analysis procedures without encountering any major technological difficulties.

The pilot also confirmed that careful human interpretation is necessary for AI-generated outputs. Journalists stressed the value of upholding editorial control, especially when handling intricate stories, subtle political situations, or time-sensitive content. As a result, the experience reaffirmed the significance of human-AI cooperation, in which human expertise is crucial to the final evaluation while AI helps with the **first discovery of potentially problematic information.**

A second key takeaway in this regard is how human editorial judgment interacts with AI-generated results. Participants repeatedly emphasized that journalists should continue to be in

charge of assessing context, interpreting claims, and confirming sources, and that the system should serve as a **support tool** rather than a decision-making process.

Although the AI tool was helpful in reducing the time needed to initially identify potentially problematic content, human judgment was still necessary for deciphering intricate storylines and socio-political contexts. Therefore, the pilot results underlined the significance of preserving human oversight, critical review, and editorial accountability when incorporating AI technologies in journalistic processes.

- **Tool usage in operational environments**

The Intelligent Coaching application was incorporated into the everyday routines of journalists, as they used it both independently in their professional work and during organized webinars, in order to examine journalistic content for possible instances of disinformation, hate speech, and logical errors. The tool supported the **preliminary screening** and **systematic assessment** of articles by acting as a complementary element to already established editorial workflows.

Participants observed that while integration into workflows was usually intuitive, several content types, especially longer texts, foreign policy pieces written in Romanian, and time-sensitive subjects, required extra human assessment.

- **Advantages observed by the users**

Participants frequently stated that the platform is **user-friendly** and **accessible**, making it possible to integrate it without the need for extensive technical knowledge. The tool's structured analysis facilitated **preliminary fact-checking** procedures and promoted a more methodical examination of journalistic content by supporting the early identification of potentially problematic statements. Two main functions - “Claim analysis” and “Hate Speech Detection” were appreciated by users as tools for identifying information that could need further human research and intervention. Users also found that the system performed more consistently when processing content in English compared to Romanian, suggesting better performance in this linguistic environment. The tool still promoted a more methodical analysis of the material and prompted consideration of the potential risks associated with bias and misinformation.

- **Challenges and limitations observed by the users**

The pilot deployment revealed a number of analytical and technical limitations that offer useful insights for the system's future improvement. Some articles imported by URLs were incomplete, and participants experienced intermittent issues retrieving content. Additionally, some websites **prohibit automated access**, because they have protections incompatible with the software,

limiting the system's capacity to thoroughly retrieve and evaluate content. Users also had trouble finding previously saved articles, which suggests that **better archive** and **search features** are required to facilitate more effective integration into editorial workflows.

The **consistency of analytical results** is another significant finding. When evaluated on various devices, the same article occasionally produced different results. Additionally, participants reported that time-sensitive claims were often assessed using out-of-date sources. This revealed a **flaw** in the system's **temporal reasoning abilities**, especially when it comes to differentiating between events and matters that are accurate in the present vs. those that were accurate in the past.

Participants also noticed **gaps in the coverage of claim detection** - statements with numerical data or statistics are usually given priority by the system, while portions that are descriptive or interpretive are usually ignored. Longer texts and intricate pieces about foreign policy or politics written in Romanian were said to be especially difficult for the system to process efficiently. In several cases, the AI-generated analysis removed significant **contextual nuances** by simplifying or reformulating phrases.

Although participants pointed out that several remarks with potentially harmful meanings were not detected, the **hate speech recognition** module failed to identify hate speech in the examined articles. This implies that in order to increase **sensitivity to subtle expressions**, more **contextual training** may be required.

- **Lessons learned and recommendations**

Due to the system's accessibility and user-friendliness, the integration of Intelligent Coaching application into LSP processes showed a great potential for aiding **structured content analysis** and disinformation identification. Simultaneously, the pilot identified a number of areas that require additional development prior to the subsequent stage (T4.3):

- **Ongoing engagement with end-users** was essential during the pilot stage in order to guarantee that the system actually represents newsrooms requirements;
- **Clear instructions**, **easily available training materials**, and **open discussion** opportunities can all help increase user interest and engagement;
- **Structured feedback methods** should be maintained and expanded because it offers important insights into the tool’s technical performance as well as the overall user experience;

- **Enhancing communication between developers and end-users** is essential in order to guarantee that concerns are effectively addressed and the AI tool is used in the most optimal way;
- Enhancing **multilingual performance**, especially when it comes to Romanian-language content;
- Extending **the scope of claim detection beyond numerical statements**;
- Enhancing **temporal thinking** for assertions that are time-sensitive;
- Improving **public source identification** and **database integration** to aid in fact-checking procedures;
- Increasing the **sensitivity** of hate speech detection;
- Enhancing **article storage, content retrieval**, and **archive features**;
- Strict guidelines for the **storage** and **confidentiality** of journalistic content uploaded.

The lessons discovered during this stage offer insightful information for the Intelligent Coaching application tool's ongoing improvement and will help to enhance the dependability, openness, and responsible integration of AI technologies within journalistic workflows.

3 Application of FAITH Tools

3.1 Application of the FAITH AI_TAF V1 methodology

As the AI system was assessed in isolation, i.e., at the asset level, four separate risk assessments were conducted, corresponding to each of the core AI models that comprise the Intelligent Coaching application (as described in Section 2.1). Specifically, the following individual risk assessments were performed:

- Risk Assessment for Claim Detection
- Risk Assessment for Matching to Debunked Claims
- Risk Assessment for Matching to Other Articles
- Risk Assessment for Hate Speech Detection

This asset-level assessment approach enabled a granular evaluation of the risks, vulnerabilities, and trustworthiness characteristics associated with each AI component, allowing potential issues to be identified and mitigated at the model level before considering the system as an integrated whole.

Following FAITH AI_TAF V1 phases as described in D2.1, **each one** of the four risk assessments was conducted in six phases (Figure 10):



Figure 10 FAITH AI_TAF V1 Risk Assessment Phases according to D2.1

3.1.1 Phase 1: Cartography

In the cartography phase the boundaries of the assessment are clearly defined in order to establish the scope of the trustworthiness and risk analysis. This includes the:

1. identification of the AI system under assessment,
2. its intended use and purpose,
3. the relevant AI team and participants involved across the lifecycle, as well as
4. its constituent components and associated AI assets,
5. criticality of the AI system according to the AI Act.

Next, the approach followed for each type of boundary is described in detail.

3.1.1.1 Identification of the AI system under assessment

For each risk assessment we regard as "AI system" the module of AI Intelligent Coaching application to be assessed. Hence, the AI systems across the various risk assessments are determined in a straightforward way:

- Risk Assessment for Claim Detection
  - **AI System:** Claim Detection
- Risk Assessment for Matching to Debunked Claims
  - **AI System:** Matching to Debunked Claims
- Risk Assessment for Matching to Other Articles
  - **AI System:** Matching to Other Articles
- Risk Assessment for Hate Speech Detection
  - **AI System:** Hate Speech Detection

3.1.1.2 Intended Use & Purpose

The intended use of each module coincides with the intended use of the AI Intelligent Coaching application as a whole, which is to support journalists and media professionals in the identification, assessment, and analysis of potential disinformation cases within their editorial workflows.

More specifically, the application is designed to assist end-users in examining news articles, claims, and online narratives by providing AI-assisted analytical support through functionalities such as:

- claim detection, to identify verifiable factual statements within a text;

- claim matching, to compare detected claims against already debunked or verified claims stored in relevant databases;
- web search support, to facilitate rapid retrieval of external corroborating or contradicting sources;
- hate speech detection, to identify potentially harmful or inflammatory language.

The overarching purpose of the application is not to autonomously determine whether a piece of information is true or false, nor to replace the professional judgement of journalists. Instead, its purpose is to function as an intelligent decision-support and coaching tool, enabling users to more efficiently detect signals of disinformation, evaluate claims, and critically assess the credibility of information sources. A key aspect of the intended use is that all outputs remain advisory and subject to mandatory human verification, with final responsibility for interpretation and publication resting entirely with the journalist.

3.1.1.3 AI Team & Participants

According to D2.1, the AI participants relevant for maturity assessment for handling AI trustworthiness are the AI Domain Users, since the AI system is in the "Deployment" lifecycle stage. Also, following the guidelines of D2.1, rather than individual participants, the AI Domain Users team is assessed as a whole. In this case, the AI Domain Users are the same across all four risk assessments, and these are the journalists from the FH network. To assess their maturity, FH organized an online webinar with journalists from PRESShub network, where the participants discussed in detail each question provided by TrustSense[5].

3.1.1.4 AI Assets

At this stage of the project, the Intelligent Coaching application is in the deployment phase; therefore, the selected lifecycle stage for the present assessment is "Deployment."

Following the in-isolation assessment approach, the AI assets considered within the scope of this evaluation include the individual AI models that constitute the Intelligent Coaching application and the associated datasets.

The AI assets across the various risk assessments are determined in a straightforward way:

- Risk Assessment for Claim Detection
    - **AI Assets:**

[5] https://trustsense-xu4xd.ondigitalocean.app/Index.html

        - Model: Claim Detection Model (multi-stage pipeline of LLMs)
- Risk Assessment for Matching to Debunked Claims
    - **AI Assets:**
        - Model: Matching to Debunked Claims Model (RAG-based AI system using LLM as a judge)
- Risk Assessment for Matching to Other Articles
    - **AI Assets:**
        - Model: Matching to Other Articles Model (Q&A functionality based on agentic RAG)
- Risk Assessment for Hate Speech Detection
    - **AI Assets:**
        - Model: Hate Speech Detection Model (two-stage, zero-shot learning LLM)
        - Dataset: Hate Speech Validation Dataset (used only for validation and performance assessment)

As indicated above, among the abovementioned components, only the Hate Speech Detection model is currently associated with a dedicated dataset, namely the Hate Speech Validation Dataset, which is used for validation and performance assessment purposes. The remaining models have been designed and configured primarily through the use of zero-shot learning techniques, complemented by iterative user testing and validation during pilot activities.

3.1.1.5 Criticality assessment

The criticality assessment of the AI Intelligent Coaching application was conducted in collaboration with partner KUL and is presented in Deliverable D4.1. For the sake of completeness, a brief overview is also provided in this section.

According to D2.1 and FAITH AI_TAF V1, the criticality of the AI system depends upon:

⇒ if 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-risk systems listed in AI Act Annex III

⇒ If the AI system poses a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially 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 essential service as defined in Annex I of NIS 2 Directive or an important service as defined in Annex 2 of NIS2.

⇒ If the AI system is processing personal data (where GDPR needs to apply).

Although the application supports journalists in identifying potential disinformation cases and may provide AI-assisted analyses, its outputs are strictly advisory in nature and do not autonomously determine editorial outcomes. In all cases, journalistic practice requires the end-users to independently verify, validate, and cross-check the information produced by the system before any editorial decision or publication takes place.

This mandatory human verification process constitutes meaningful human oversight, substantially reducing the likelihood that potential system malfunctions could directly lead to harm to health, safety, or the fundamental rights of natural persons.

Furthermore, the application does not fall within the high-risk categories of Article 6 and Annex III of the AI Act (such as biometric ID, critical infrastructure, education, or law enforcement). While "Democratic processes" are mentioned in Annex III (Point 8), the focus is on systems intended to *influence* elections or voters directly. Since the AI Intelligent Coaching application is an assistant for professional journalists with clear "Human-in-the-loop" safeguards, it avoids the High-Risk designation.

Also, the application is not used for the provision of essential or important services under the NIS2 framework as most media organizations do not qualify as "Operators of Essential Services" under NIS2 (which focuses on energy, transport, banking, health, etc.). Even if used within a large digital infrastructure, fact-checking is generally not considered an "Essential" or "Important" service as defined in Annex I or II of NIS2.

Potential risks, such as false detections, missed claims, or biased outputs, are mitigated through professional editorial review and fact-checking procedures. As a result, the system is considered to pose low risk of harm, primarily limited to workflow inefficiencies or the need for additional manual validation.

3.1.2 Phase 2: Threat Analysis

According to FAITH AI_TAF V1, documented in D2.1, for the threat assessment we performed desk analysis to comprehensively examine and instantiate the diverse threats of all AI assets during each phase of the AI lifecycle that could undermine system trustworthiness. These include technical threats like malware, data poisoning as well as social threats such as phishing, social engineering (depending on the trustworthy AI maturity level of the organization's teams). The potential threats investigated within the FAITH AI_TAF V1 are derived from the FAITH Golden Table, as introduced in Deliverable D2.1. This threat taxonomy is continuously updated based on insights from the scientific literature (e.g., [ENISA, 2021], [Poretschkin et al., 2023]), in line with recognized risk management frameworks, including ISO/IEC 23894:2023 (Artificial Intelligence — Risk Management), the NIST AI Risk Management Framework (AI RMF 1.0) and the LSPs.

In this phase we estimate the occurrence of the threat, ranging from “Never” to “Twice a year”. Factors influencing the occurrence of a threat include:

- historical data: previous occurrences of similar threats can provide insight into future probabilities;
- environmental factors: Conditions in a specific area or sector that the AI system is being used (e.g., natural disasters, political/economic/ business stability);
- stability & trends: geopolitical situations (e.g., economic crisis, disasters, wars, pandemia), technological trends (e.g., AI attack systems, current cybersecurity landscape) can indicate rising threats.
- criticality of the AI system

As discussed in Section 3.1.1, the Intelligent Coaching application is considered low risk. Non-critical AI systems are generally less likely to suffer from severe cybersecurity threats compared to critical infrastructure AI systems, mainly because the consequences of compromise are less significant and, hence, less attractive targets for adversaries [Anderson, 2001]. However, in the cybersecurity landscape of 2026, the risk to AI systems stems from automated agentic bots that scan the entire internet for vulnerabilities, regardless of the target's "importance"[Li et al., 2025]. Hence, this factor was also taken into account.

Threats not applicable to the Media LSP context or not relevant to the specific implementation of the AI Intelligent Coaching application (e.g., the threat “Bias in Personalized Learning Recommendations” applies to LSP3 but not to LSP1-Media, as AI Intelligent Coaching application does not provide any recommendations to the user) are assigned a likelihood of “Never”, as they are considered outside the operational scope of the system or not feasible under the current deployment conditions.

The rest of the threats, their descriptions and the likelihoods along with a justification—derived from practical experience, common reasoning or from the literature—are presented in Table 2. The threats from [ENISA, 2021], [Poretschkin et al., 2023], and the Media LSP are distinguished respectively by blue, green and purple rows.

Threat	Description	Likelihood
Label modification	An attack in which the attacker corrupts the labels of training data. This sub-threat is specific to Supervised Learning.	**Never.** There is no training data in *any* of the assessed models.
Evasion	A type of attack in which the attacker works on the ML algorithm's inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors). It is as if the attacker created an optical illusion for the algorithm. Such modified inputs are often called adversarial examples. Example: the projection of images on a house could lead the algorithm of an autonomous car to take the decision to suddenly make it brake	**Claim detection: Once a year.** Being a multistage pipeline system, evasions can happen in the summarization phase (summarization hijacking [Leite et al., 2026]) or the ranking phase (ranking manipulation) [Leite et al., 2026] **Match to DB: Never.** This module performs a similarity check against a local, vetted database. Unlike a generative process or a search of the open web, the matching is a comparison between the user's claim and a static, pre-existing set of entries. **Match to Other Articles: Once a year**. As the system consumes external content via Search APIs (Google/Tavily). An attacker can perform 'Data Source Impersonation' by publishing web pages containing Indirect Prompt Injections (IPIs). When the API retrieves this content, the LLM may follow the hidden commands to misclassify the resource as 'supporting' even if the content is false [Kim et al., 2026] **Hate speech detection: Once a year**. Research shows that "Lexicon Evasion"—using non-toxic metaphors (e.g., cooking terms) or specific dialects—is highly effective at bypassing binary "Hate vs. No Hate" filters [Barakat, 2025].

Use of adversarial examples crafted in white or grey box conditions	In some cases, the attacker has access to information (model, model parameters, etc.) that can allow him to directly build adversarial examples. One example is to directly use the model's gradient to find the best perturbation to add to the input data to evade the model.	**Once every 2 years for all functionalities.** By using Llama models through Groq, the AI Intelligent Coaching application exists in a unique "Grey-Box" state. While Groq acts as a black-box API (preventing direct access to gradients), the Llama models used are open-weight, hence an attacker can download the exact same model weights from Meta/HuggingFace, craft a White-Box attack on their own machine using gradients, and then send that finished "adversarial payload" to the Groq-hosted system. Because the models are identical, the attack has a near-perfect Transferability rate. However, this type of attack is targeted and requires high technical skill.
Oracle	A type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be previous steps to more harmful types, evasion or poisoning for example. It is as if the attacker made the model talk to then better compromise it or to obtain information about it (e.g., model extraction) or its training data (e.g., membership inference attacks and Inversion attacks). Example: an attacker studies the set of input-output pairs and uses the results to retrieve training data.	**Claim Detection: Once every 2 years**. Attackers can use "Sensitivity Analysis" on the final output to reverse-engineer the hidden summarizer's "fact-filtering" rules [Jawad et al., 2025]. **Match to DB**: **Never.** The model is already using well-known embedding techniques. **Match to Other Articles: Once every 5 years**. While possible to probe search-keyword logic, the effort-to-reward ratio is low for a non-critical tool, keeping this rare. **Hate Speech Detection: More than once a year.** Attackers constantly "test the fence" with 2026's automated tools (like AutoDAN) to find the exact phrasing that slips past the binary gate [Liu et al., 2024].

Failure or malfunction of ML application	This threat refers to ML application failure (e.g. denial of service due to bad input, unavailability due to a handling error). Example: the service level of the support infrastructure of the ML application hosted by a third party is too low compared to the business needs, the application is regularly unavailable. Note that this threat does not consider failure of business use cases (for example, the algorithm fails because it is not accurate enough to handle all real-life situations it is exposed to).	**Claim Detection: More than once a year.** Large-scale text summarization creates high Key-Value (KV) Cache demands. In shared hosting environments, "bad inputs" (e.g., extremely long, repetitive, or non-tokenizable strings) can trigger memory overflows or worker-node crashes, leading to temporary service unavailability for the specific pipeline stage [Yu et al. , 2026]. **Match to DB: Once every 5 years.** As a locally deployed component, this module processes structured data from a controlled environment rather than external streams. System failure is unlikely because the data does not cross third-party cloud boundaries. A malfunction would only occur in the event of a local logic error within the LangGraph orchestration layer or a service timeout from the local database engine. Because the database URLs are hardcoded, the risk of external redirection or connection failure is minimized [OWASP, 2025]. **Match to Other Articles: More than once a year.** This module is highly dependent on synchronous external API calls (Google, Tavily). The complexity of managing Agentic Concurrency often leads to 'Cascading Failures' (ASI08). If a search API hits a rate-limit (RPM/TPM) or suffers a latency spike, it can cause the entire LangGraph orchestration thread to stall. Without robust circuit breakers, these delays propagate through the pipeline, resulting in a Denial-of-Service for the end-user [OWASP, 2026]. **Hate Speech Detection: More than once a year.** Automated bots can saturate the Groq free-tier limits (e.g., 30 RPM), rendering the detection service unavailable for legitimate users and causing a total application hang.

Denial of service due to inconsistent data or a sponge example	ML algorithms usually consider input data in a defined format to make their predictions. Thus, a denial of service could be caused by input data whose format is inappropriate. It may also happen that a malicious user of the model constructs an input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service	**Claim Detection: More than once a year.** As the primary "ingestion" point for large documents, this module is highly vulnerable to Sponge Examples. Attackers provide high-entropy, repetitive, or "infinite-loop" text prompts that force the model to maximize its output token generation, saturating the KV cache and causing worker-node timeouts [Liu et al., 2024]. **Match to DB: Once every 5 years.** Inconsistent formats (e.g., non-UTF-8 characters) are typically caught by the LangGraph orchestration or the local database schema before reaching the model. Furthermore, since the system uses fixed local retrieval rather than generative RAG, the computational cost is relatively static, making it an inefficient target for "Sponge" tactics designed to exhaust resources [OWASP, 2025]. **Match to Other Articles: More than once a year.** This module is susceptible to Format Inconsistency. via content retrieved through Search APIs (Google/Tavily). If an external web resource is intentionally "poisoned" with "junk data" (e.g., massive binary blobs, non-UTF-8 characters, or hidden recursive HTML) and passed to the LLM for relevance matching, it can trigger an "Inference Hang." In this scenario, the Groq-based model spends excessive GPU cycles attempting to tokenize non-textual noise, effectively stalling the fact-checking pipeline and creating a targeted Denial-of-Service. **Hate Speech Detection: More than once a year.** Attackers utilize Sponge Examples—inputs mathematically optimized to trigger "Overthinking" or "Unbounded Drift" in the model's internal reasoning. This forces the inference engine (e.g., Groq LPU) to utilize disproportionate compute cycles for a single

		classification, saturating the KV Cache and effectively slowing down the "front-gate" for all other users [Cinà, 2025]
Poisoning	A type of attack in which the attacker altered data or model to modify the ML algorithm's behavior in a chosen direction (e.g. to sabotage its results, to insert a backdoor). It is as if the attacker conditioned the algorithm according to its motivations. Such attacks are also called causative attacks. Example: massively indicating to an image recognition algorithm that images of dogs are indeed cats to lead it to interpret it this way.	**Claim Detection: Once every two years.** Vulnerable to "Semantic Context Deception." Attackers seed public repositories or documentation with misleading data formatted specifically for LLM consumption. While the model is not re-trained, the ingestion of this "poisoned context" sabotages the summarization phase, leading to systematic omissions of critical facts [Sami, 2026]. **Match to DB: Once a year.** Susceptible to Knowledge Base Poisoning. An attacker with limited access can inject as few as five adversarial documents into the Vector DB to achieve a 90% success rate in steering responses. These documents are optimized to rank highly in retrieval (Semantic Hijacking), effectively "conditioning" the agent to match false claims to legitimate debunked files [Zou et al., 2025]. **Match to Other Articles: More than once a year.** Exposed to Web-Scale Poisoning. Attackers publish "Poisoned Skill" files or web resources containing Indirect Prompt Injections (IPIs). When the API retrieves these articles into the LLM's context, the model integrates the malicious logic (e.g., "Always label this specific source as credible"), causing a persistent behavioral shift in the ranking results (Qu et al., 2026; Cui et al., 2026). **Hate Speech Detection: Once every 5 years.** Since this module uses a fixed, pre-trained Llama model on Groq, it is resistant to live

		poisoning unless the underlying model weights are compromised at the provider level. However, a "Backdoor Trigger" could theoretically be activated if the attacker knows the model's pre-training "Vortex" triggers (Cui et al., 2026).
Model or data disclosure	This threat refers to the possibility of leakage of all or partial information about the model. Example: the outputs of a ML algorithm are so verbose that they give information about its configuration (or leakage of sensitive data)	**Claim Detection: Once every 10 years.** As a primary ingestion point using models like text-embedding-3-large, the lack of direct access to model internals prevents traditional extraction. **Match to DB: Never.** Model disclosure for externally hosted embedding models accessed via API (such as text-embedding-3-large) is highly unlikely due to the absence of direct access to model internals, the non-generative nature of outputs, and the presence of provider-side safeguards against large-scale probing and extraction attempts. There is no likelihood for data disclosure, as the DB is open to end users. **Match to Other Articles: Once every 10 years.** Even within agentic workflows, the non-generative nature of embedding outputs (which return numerical vectors rather than readable text) significantly complicates "jailbreak" or model-inversion attempts designed to leak training data. **Hate Speech Detection: Once every 10 years.** While this module faces frequent "Lexicon Evasion" attempts, these aim to bypass the *gate*, not disclose the model itself. API providers implement strict rate limits and monitoring to prevent the high-volume probing required for model extraction.

Data disclosure	This threat refers to a leak of data manipulated by ML algorithms. This data leakage can be explained by inadequate access control, a handling error of the project team or simply because sometimes the entity that owns the model and the entity that owns the data are distinct. To train the model, it is often necessary for the data to be accessed by the model provider. This involves sharing the data and thus share sensitive data with a third party.	**Claim Detection: Once every 5 years.** Large volumes of text (e.g., entire documents) are transmitted to the third-party provider (Groq). Groq implements reasonable and appropriate data encryption and security designed to protect customer data.[6] **Match to DB: Never.** Because the database is open to end users and the matching is a deterministic property of the stored embeddings, there is no proprietary logic to "disclose" in this component. **Match to Other Articles: Once a year.** This module involves the transmission of proprietary search queries and the receipt of external data via Search APIs (Google/Tavily). A leak occurs if the API request/response metadata—which contains the user's specific fact-checking queries—is stored insecurely in centralized logging systems or cloud-based observability platforms. Furthermore, since scrapers are not used, the risk shifts to Identity-Based Access Controls: if an attacker compromises a developer's identity, they can gain access to the API keys or the "search thread" history stored within the local orchestration logs, leading to a breach of search privacy and project intellectual property. **Hate Speech detection. Once every 5 years.** For the same reason as in the claim detection.
Model disclosure	This threat refers to a leak of the internals (i.e. parameter values) of the ML model. This model leakage could occur because of human error or contraction with a third party with a too low security level.	**Claim Detection: Once every 2 years.** Vulnerable to Direct Context Extraction (DCE). The final outputs are often verbose enough to allow an attacker to "probe" the system prompts. Automated tools can reverse-engineer the hidden summarization

[6] https://trust.groq.com/faq?search=What+security+measures+are+in+place+to+protect+customer+data%3F

		rules by analyzing how input variations change the final claim list [Li & Gao, 2025]. **Match to DB: Once every 10 years.** This module uses standard RAG logic; therefore, the "model" itself is not a high-value target for disclosure. The disclosure risk here is primarily linked to Data Disclosure (the Vector DB) rather than the model's proprietary architecture (OWASP LLM02: 2025). **Match to Other Articles: Once every 2 years.** The System Prompts and Orchestration Logic (LangGraph nodes) are high-value targets for disclosure. Under OWASP LLM02:2025, this module is susceptible to Inference-Based Exfiltration. Adversaries can use the tool as an "Oracle" to extract snippets of the local Vector DB or reconstruct the proprietary matching logic through iterative probing. The risk is not "Model Theft" in the traditional sense, but Contextual Data Disclosure where the model unintentionally reveals the content of the "debunked" database it was meant to protect [OWASP LLM08, 2025]. **Hate Speech Detection: More than once a year.** This module is the most targeted for Safety Boundary Mapping. Attackers treat the binary gate as an Oracle to disclose the "Refusal Boundaries" of the system. By identifying exactly what triggers a refusal, they disclose the model's hidden moderation logic and safety configuration.
Compromise of ML application components	This threat refers to the compromise of a component or developing tool of the ML application. Example: compromise of one of the open-source libraries used by the	**Claim Detection: Once every 5 years.** While the model is remote, a compromise of the local SDK or the environment where it resides could lead to credential harvesting. However, as a "lean" component, it remains a difficult target for non-targeted attacks.

	developers to implement the ML algorithm.	**Match to DB: Once every two years.** Because this is locally deployed, the "component" is the database engine (e.g., FAISS/Chroma) and the local embedding model. An attacker would target a vulnerability in the local DB binary or a Poisoned Model File to gain local execution. **Match to Other Articles: Once every 5 years.** Using Google/Tavily APIs through hardcoded endpoints minimizes the "Man-in-the-Middle" or "Redirection" risk. The primary remaining component risk is the JSON parsing library used to handle the search results, which could be exploited by a "Malformed Response" from a poisoned site **Hate Speech Detection: Once every 5 years.** Like Claim Detection, the risk is limited to the local environment's security. Without third-party RAG or scrapers involved in this step, the component surface is restricted to the API communication logic and local secret management.
Human error	The different stakeholders of the model can make mistakes that result in a failure or malfunction of ML application. For example, due to lack of documentation, they may use the application in use-cases not initially foreseen.	**Claim Detection: Once a year.** Developers may update the summarization prompt without adequate regression testing. A minor wording change in the system instructions can cause the model to miss subtle claims or change output formats (e.g., switching from JSON to Markdown), breaking the downstream LangGraph nodes. **Match to DB: More than once a year.** Since this is **locally deployed**, human error in database indexing, schema updates, or "dirty" data ingestion (e.g., uploading duplicates or incorrectly labeled debunked claims) will lead to high false-negative rates that the system cannot automatically self-correct.

		Match to Other Articles: More than once a year. Misunderstanding the "Search Top-K" or "Temperature" settings in the Tavily/Google API calls. If a stakeholder increases the search breadth too far without adjusting the context window logic, it can lead to "Context Overflow" or "Hallucinated Consensus"[7]. **Hate Speech Detection: More than once a year.** A lack of documentation may lead stakeholders to use this module for "Sentiment Analysis" or "Sarcasm Detection," use-cases it was not tuned for. This results in "Misclassification Drift," where users lose trust in the tool because it is being applied to linguistic nuances it cannot handle
Cybersecurity incident not reported to incident response teams	This threat refers to the possibility that a project team may not report security incidents to dedicated teams, while a policy of mandatory incident reporting has been defined.	**Claim Detection: Once every 2 years.** High-pressure environments where "production speed" is prioritized often lead to staff dismissing minor anomalies (e.g., unusual extraction errors) as technical glitches rather than potential security breaches. **Match to DB: Never.** Since this module is largely deterministic and internal, the "incidents" are typically system failures rather than security breaches, leading to a higher natural rate of reporting to IT support rather than security teams. **Match to Other Articles: Once every 5 years.** Staff may fail to report suspicious search behaviors or "Indirect Prompt Injections" if they do not perceive the search-driven results as an active security incident. **Hate Speech Detection: More than once a year** Research shows that staff frequently fail to report "filter bypasses" or "jailbreaks" if

[7] https://www.vectra.ai/topics/genai-security

		they believe the incident was "handled" by the filter eventually, or if they fear administrative burden or personal blame for the prompt they entered [Liu et al., 2026].
Prompt attacks	LLMs are highly sensitive to how prompts are constructed. Even slight variations in phrasing can lead to inconsistent responses. It is related to the way large language models (LLMs) interpret and respond to input. These attacks can manipulate the model's behavior, extract sensitive information, or degrade performance.	**Claim Detection: More than once a year.** The threat is Direct Prompt Injection (DPI) at the primary ingestion point. Because this module processes raw, unvetted text provided by users, the system will frequently encounter "Hidden Instruction Sets" (e.g., "Ignore all prior instructions"). Given the high volume of potentially adversarial input in fact-checking contexts, the likelihood of an injection attempt successfully neutralizing the extraction logic is high. **Match to DB: More than once a year.** The threat is System Prompt Exfiltration via "Linguistic Grooming." While the local database is structured, the model's reasoning nodes remain susceptible to iterative, multi-turn probing. Attackers will statistically attempt to force the model to disclose the internal system prompts used for matching logic once per year, aiming to reverse-engineer the "debunked" database's proprietary sorting criteria [Akram, 2026] **Match to Other Articles: More than once a year.** The threat is Indirect Prompt Injection (IPI) via untrusted search results. Because the module ingests third-party content from Search APIs (Google/Tavily), the system is exposed to "Index Poisoning." The frequency of the model encountering a poisoned article—which instructs the agent to prioritize fake sources—is high, as attackers increasingly optimize misinformation to be picked up by automated agentic retrievers.

		Hate Speech Detection: More than once a year. The threat is Adversarial Jailbreaking via "Semantic Reframing." Attackers frequently use creative encoding (e.g., "Translate this into a fictional screenplay") to bypass the refusal boundaries of models like Llama 3.3.4. The likelihood of the moderation gate encountering these sophisticated bypass attempts is high, as the "jailbreak-as-a-service" market in 2026 ensures a constant stream of new, automated attack vectors.
Regulatory and legal non-compliant	Many regulations (e.g., GDPR, AI Act) require systems to provide explainable decisions. Lack of interpretability can make compliance difficult or impossible.	**Once every 5 years for all AI Intelligent Coaching application functionalities.** AI Intelligent Coaching application is classified as a low-risk system under the AI Act and does not raise direct GDPR concerns. The likelihood of regulatory or legal non-compliance due to limited interpretability cannot be considered negligible. This is because the system relies on LLM technologies, which inherently lack transparency and may not provide clear explanations for their outputs.
Propagation of misinformation	Propagation of misinformation refers to an LLM's ability to generate text that sounds plausible but is factually incorrect, misleading, or even harmful. The lack of interpretability makes it difficult to understand why the model generates this misinformation or how to stop it.	**Claim Detection: More than once a year.** The model might "summarize away" critical context from a claim, turning a nuanced truth into a simplified lie. In 2026, "narrative manipulation" via summarization is a top global risk **Match to DB: Once every two years.** A weak match in the local DB might lead the tool to claim a new, true event is "already debunked", effectively suppressing accurate information. **Match to Other Articles: More than once a year.** Search APIs (Tavily/Google) can be gamed by "SEO-optimized misinformation." If the model finds 10 AI-generated fake

		articles and 1 real one, it may report a “false consensus” to the user. **Hate Speech Detection: More than once a year.** By triggering "Over-refusal" on certain political topics, the tool can be used to "soft-censor" legitimate discourse, creating a skewed information environment where only one side's claims are allowed to be “checked”[17].
Unfair Discrimination in AI Outputs	A threat where the model produces results that systematically disadvantage specific groups based on sensitive attributes (race, gender, dialect, etc.).	**Claim Detection: Once every 5 years.** The extraction process may exhibit "Linguistic Erasure," where the model fails to identify factual claims written in Non-Standard English or specific cultural dialects. Recent studies show that LLM-based extractors are 15% more likely to ignore valid claims from marginalized communities [Vaswani et al., 2025]. However, from interactions with media professionals within this pilot, it was resulted that it is highly unlikely that such resources would be searched. **Match to DB: Never.** This is a deterministic similarity check against a pre-existing, vetted database. The "fairness" here is a property of the database content, not the matching algorithm itself. **Match to Other Articles: Once every 2 years.** Bias may occur in the "relevance ranking" phase. **Hate Speech Detection: Monthly.** This is the highest risk area. Research into "Subgroup Over-sensitization" shows that moderation models frequently flag neutral discussions of identity (e.g., "I am a trans woman") as "Hate Speech" simply because those keywords are over-represented in toxic training data.

Discriminatory Behavior Learned by AI	A threat where the AI replicates and scales societal inequities found in its massive training datasets. *(Interpretation of the threat in the LLM domain)*	**Claim Detection: Every 5 years.** The model is prompted with a very specific definition the notion of claim. It is very unlikely that will have discriminatory behaviour, **Match to DB: Never.** The module uses frozen embeddings; it does not "learn" or adapt during its lifecycle. **Match to Other Articles: Every 2 years.** The model may exhibit "Source Preference Bias," where it has learned to implicitly judge as more relevant certain domains over others, leading to a skewed representation of "verified" facts. **Hate Speech Detection: Once a year.** Even with safety filters, models can exhibit "Bias Drift." As the underlying API is updated by the provider, the model may learn new, subtle discriminatory patterns from contemporary internet slang used in its updated training sets [Singh et al., 2026].
Misrepresentation of Sensitive Characteristics	A threat where the system incorrectly categorizes or mislabels a user's identity, intent, or speech style. *(Interpretation of the threat in the LLM domain)*	**Claim Detection: Every 2 years.** Occurs if the model misinterprets the intent behind a claim (e.g., mistaking sarcasm for a factual statement in specific cultural contexts). **Match to DB: Never.** The system matches text-to-text and does not categorize the user's identity. **Match to Other Articles: Once a year.** An international study released in October 2025 by the European Broadcasting Union (EBU) and the BBC indicated that leading AI assistants, such as ChatGPT, Copilot, Gemini, and Perplexity, misrepresent news content 45% of the time. This issue is consistent across different languages, territories, and platformsThe model may mislabel because it

		misinterprets the specific cultural or religious terminology used in the article[8]. **Hate Speech Detection: More than once a year.** LLMs in tasks like hate speech have lower performance in low-resource languages, like Romanian [Muminovic et al., 2025].
Changes in External Conditions	A threat where shifting societal norms, new regulations, or geopolitical events render the model's current logic "unfair" or "non-compliant." *(Interpretation of the threat in the LLM domain)*	**Claim Detection: Once a year.** New transparency laws (like the updated EU AI Act of 2026) may require deeper justifications for why a claim was extracted, making current "black-box" extraction non-compliant. **Match to DB: Never.** Even if the database is not updated to reflect new societal "truths" (e.g., a change in scientific consensus), the user is presented with the fact checks and their dates to ensure that the assessment is up to date. **Match to Other Articles: Monthly.** External "Data Poisoning" of the web sources. Coordinated campaigns can flood search engines with biased articles, shifting the "consensus" the LLM sees via the Search API. **Hate Speech Detection: Once a year.** Legal definitions of hate speech change. A model calibrated for 2025 standards may fail to detect new forms of "coordinated harassment" or "coded hate" that emerge in 2026.

[8] https://www.ebu.ch/news/2025/10/ai-s-systemic-distortion-of-news-is-consistent-across-languages-and-territories-international-study-by-public-service-broadcaste

Dynamic User Behavior Impacts	A threat where user interactions (prompts) push the model into biased or unfair outputs. *(Interpretation of the threat in the LLM domain)*	**Claim Detection: Once a year.** Users may "bait" the extractor with highly biased leading questions, forcing the model to extract claims that reinforce the user's existing prejudices (Confirmation Bias loops). **Match to DB: Never.** The semantic matching is against a static table; user behavior cannot change the database content. **Match to Other Articles: Once a year.** "Prompt Injection" via the search query. A user can craft a query that "forces" the LLM to only find and rank search results that support a specific biased viewpoint [Geng, 2026]. **Hate Speech Detection: Monthly.** Users constantly innovate "Adversarial Slang" or use "Persona Adoption" (pretending to be a victim) to bypass filters or trick the model into flagging innocent users [Ahmad, 2025].

Table 2. Threats' frequencies relevant to the Media LSP

A key observation from the Threat Assessment is the need to update and refine the definitions of the identified threats to ensure their applicability to Generative AI (GenAI) models.

3.1.3 Phase 3: Impact Assessment

The purpose of this assessment is to evaluate the potential consequence of each threat to the various dimensions of trustworthiness (technical characteristics such as accuracy and robustness, transparency, fairness, explainability, accountability, privacy). The impact level of each threat is further refined based on the potential consequences—technological, legal, financial, and others—that organizations may face when these trustworthiness dimensions are compromised.

For the risk assessment of the AI Intelligent Coaching application, the TrustGuard tool was employed. TrustGuard enables the automated calculation of the impact of identified threats and vulnerabilities on the various trustworthiness dimensions, leveraging the FAITH Golden Table. For instance, the vulnerability "lack of training based on adversarial attacks" is mapped to impacts on the trustworthiness dimensions of "Safety" and "Security and Resilience". For this

reason, in the context of this phase of the pilot, the analysis focused specifically on the organisational consequences arising from the potential compromise of these trustworthiness dimensions, rather than on technical or individual-level impacts, which, as explained, are automatically calculated through TrustGuard.

The impact assessment was conducted by FH based on information gathered from the workshops with participating journalists, an internal examination of the system's possible operational and ethical ramifications. The results of the user feedback forms that participants filled out, which recorded opinions about the system's usability, transparency, dependability, and trustworthiness, supported these insights.

What is the business consequence level if the safe operation of the AI system is compromised?

- **Moderate**

When assessing journalistic content, the tool might produce erroneous or deceptive evaluations if the AI system's safe operation were compromised. In reality, this might have an impact on the detection of false information, misleading assertions, or other content-related problems in examined publications. Unreliable outputs could sometimes affect how content is interpreted and lower the effectiveness of journalistic processes, even though the system is intended as a support tool and does not replace the position of editorial decision-making. Considering that human monitoring is still a crucial component of the process, the possible business consequence is therefore seen as moderate.

What is the business consequence level if the security or the resilience of the AI system to adverse events is compromised?

- **Moderate**

Service disruptions, unauthorized access, or manipulation of system outputs could arise from a breach in the security or resilience of the system. The availability or dependability of the instrument inside journalistic workflows may be momentarily impacted by such occurrences, which could also interfere with the analytical process. Even though the AI system has no direct control over publication choices, any security-related issue could erode user confidence and make journalists less inclined to use the system for content analysis. As a result, the business effect of these risks is rated as moderate.

What is the business consequence level if the AI system is not interpretable and explainable?

- **Moderate**

For people to comprehend how AI-generated outputs are created, interpretability and explainability are crucial components. Journalists may find it challenging to evaluate the validity of the insights produced if the system does not offer adequately lucid explanations of its analytical findings. This ambiguity may limit the system's use in work settings where editorial decision justification and openness are crucial. Therefore, a moderate amount of possible business consequence could result from inadequate explainability, which could lower user confidence and impede the system's successful integration into newsroom procedures.

What is the business consequence level caused by a privacy compromise of the AI system?

- **Very High**

In the context of journalism, a privacy compromise poses an especially serious risk because articles and written content that may contain sensitive investigative data, unpublished reports, or confidential documents are processed by the AI system. Unauthorized access to this kind of information could compromise journalistic privacy, harm one's reputation as a professional, and deter users from using the platform. The potential commercial repercussions of a privacy breach are thought to be extremely severe, given the significance of safeguarding editorial materials and journalistic sources. This evaluation is in line with the EU AI Act's strong emphasis on data protection and privacy safeguards, which emphasizes the necessity for AI systems to guarantee safe data handling, personal information protection, and suitable governance mechanisms throughout the system lifecycle. Additionally, it is consistent with the European Commission's more general guidelines for trustworthy AI, which emphasize the significance of privacy and responsibility as essential conditions for the ethical development and application of AI technologies.

What is the business consequence level if the AI system is not fair?

- **High**

A crucial prerequisite for AI systems that analyze journalistic content is fairness. The interpretation of particular stories or subjects during the analytical process may be impacted if the system generates unfair or biased results. Such bias could damage the tool's trustworthiness in a media setting and create moral questions about how to understand delicate subjects. Even though journalists still have the ultimate authority over editorial decisions, chronic bias in AI results might undermine user acceptability and erode system trust. These factors make the potential business impact of inadequate fairness considerably high.

What is the business consequence level if the AI system is not transparent?

- **Moderate**

Transparency refers to the degree to which users can comprehend how the system functions, how data is processed, and how analytical results are produced. Users may find it more challenging to evaluate the system's strengths and weaknesses if openness is restricted. Inadequate transparency could undermine user confidence and make it more difficult to successfully integrate the AI tool into current editorial workflows, even though the technology is meant to support human decision-making rather than replace it. As a result, the possible business impact of a lack of openness is rated as moderate.

What is the business consequence level if the AI system is not accurate and robust?

- **High**

The precision and resilience of the AI system's analytical results are critical to its efficacy. The system's ability to spot false information, identify misleading assertions, or flag dangerous content would be greatly diminished if it generated erroneous, inconsistent, or unstable results. Journalists may have to spend more time confirming results in these circumstances, which could reduce the tool's anticipated productivity gains. Additionally, persistent errors may erode user confidence and lower the possibility of long-term adoption. As a result, inadequate accuracy and robustness are thought to have a significant business impact.

3.1.4 Phase 4: Vulnerability Assessment

This phase focuses on identifying weaknesses in the AI system, including both technical vulnerabilities—such as software flaws, insufficient data governance practices, and network weaknesses—that could be exploited by malicious actors, as well as human-related vulnerabilities, such as a low level of trustworthy AI maturity within the development and operational teams.

The vulnerabilities assessed in Task T4.2 are derived from the FAITH Golden Table, which, like the taxonomy of threats, is continuously updated based on insights from both the scientific literature (e.g., ENISA, 2021; Poretschkin et al., 2023) and findings emerging from the Large-Scale Pilots.

Within the FAITH Golden Table, each identified vulnerability is mapped to a set of control questions, which serve as a structured mechanism for its assessment. For example, the vulnerability “lack of detection of abnormal inputs” is associated with questions such as: “Have tools or techniques been implemented to detect adversarial inputs or abnormal data during model execution?” and “Is the ML application integrated into the organisation’s security incident

detection and response procedures?" Through this mapping, system vulnerabilities are identified and evaluated in a consistent and traceable manner.

The vulnerability assessment was implemented by using TrustGuard which leverages the FAITH Golden Table, and by requesting from the owners (which are also the developers) of each module (Claim Detection, Matching to DB, Matching to Other Articles, Hate Speech Detection) to answer the various control questions.

An overview of the vulnerability assessment of each module is presented next.

3.1.4.1 Claim Detection Component

In summary, the Claim Detection component incorporates a combination of technical, organisational, and privacy-preserving controls aimed at enhancing the security, transparency, and robustness of the system. These measures include awareness of AI-specific attack vectors, restricted exposure of model internals, lifecycle security management practices, integration within organisational asset management processes, and support for explainability and traceability. In parallel, privacy-preserving and data minimisation practices have been adopted to reduce unnecessary data exposure and support secure operation in diverse deployment environments.

However, the majority of relevant controls are not implemented, particularly those related to security, robustness, monitoring, and governance. There are no adversarial defence mechanisms in place, such as adversarial training, input sanitisation, or detection of malicious inputs. Similarly, the system lacks continuous monitoring capabilities, including performance tracking, drift detection, and anomaly detection, which are critical for identifying failures or degradation in production. Organisational controls are also missing: the ML application is not integrated into incident response processes or security-by-design practices. Furthermore, no formal risk analysis, compliance assessment, or documentation standards have been applied, and there are no implemented measures for bias detection, fairness auditing, or regulatory alignment.

A substantial portion of controls are marked as not applicable (N/A), primarily due to the system's dependence on pre-trained models developed and hosted by third parties (e.g., Meta and Groq). Controls related to model training—such as adversarial training, dataset curation, poisoning detection, differential privacy during training, and federated learning—are considered outside the system's scope. Similarly, controls involving access to model architecture, training data, embeddings, or internal parameters are not applicable, as these elements are managed by the model providers and not accessible to the system developers. Fairness interventions at the model level, including embedding adjustments or bias mitigation during training, also fall into this category.

Overall, the control landscape reveals an imbalance: a small set of implemented controls focused on limiting output exposure, a large number of unimplemented but applicable controls related to operational security and governance, and many controls deemed not applicable due to architectural dependency on third-party models.

In conclusion, the vulnerability assessment demonstrates that the Claim Detection system operates with minimal internal safeguards, relying heavily on external model providers for core AI-related controls while leaving many application-level and organisational controls unimplemented. To improve its security and trustworthiness posture, priority will be given to implementing controls that remain within scope—such as input validation, monitoring, incident response integration, and compliance processes—while also establishing stronger assurance mechanisms for third-party dependencies.

3.1.4.2 Match to Debunked Claims

The vulnerability assessment of the "Match to Debunked Claims" component follows the same structural pattern observed in the other modules, with controls distributed across implemented, non-implemented, and not applicable (N/A) categories. This distribution is again largely driven by the architectural decision to use pre-trained large language models within an AI workflow, combined with reliance on third-party providers for model training, hosting, and infrastructure. As a result, many core AI controls are externalised, while several application-level and organisational safeguards remain unimplemented.

A small number of controls are implemented, mainly at the level of output restriction and workflow design. The system ensures that outputs are limited to textual responses and do not expose internal model details such as probabilities, confidence scores, or reasoning traces, thereby reducing the risk of adversarial probing or reverse engineering. Additionally, error handling is implemented in a secure manner, with only generic internal server errors returned to users, preventing leakage of internal prompts or workflow logic. The component also constrains model behaviour by operating through a predefined sequence of steps, with no autonomous agent capabilities. Importantly, a form of retrieval grounding is implemented: the system queries a verified database of debunked claims (e.g., Elections24), which provides a more controlled and curated knowledge source compared to open web retrieval, thereby partially mitigating misinformation risks. Some of the organisation controls have been implemented: the component is integrated on the local asset management system and organization established processes have been implemented for documentation. Controlled practices have been developed for accessing the AI models.

Despite these implemented elements, the majority of relevant controls are not implemented, particularly in the areas of security, robustness, monitoring, and governance. There are no input sanitisation or adversarial detection mechanisms, leaving the system vulnerable to prompt injection, malicious inputs, and manipulation of matching results. There is also no continuous monitoring of model performance, meaning that degradation, unexpected behaviour, or matching inaccuracies may go undetected. Some of the organisational controls are similarly absent: the component is not integrated into incident response processes and security-by-design frameworks. Furthermore, there are no established processes for auditing, or lifecycle risk management. Fairness-related controls, including bias detection, fairness audits, and mitigation strategies, are also not implemented, even though biases in the underlying models or in the debunked claims database could affect outputs.

A significant number of controls are classified as not applicable (N/A), primarily due to the use of pre-trained, open-source models and the absence of a model training lifecycle within the system. Controls related to adversarial training, dataset curation, poisoning detection, differential privacy, federated learning, and fairness-aware training are all outside the system's scope. Similarly, controls requiring access to model internals (e.g., embeddings, parameters, or training data) are not applicable, as these are managed entirely by external providers. The same applies to lifecycle controls related to retraining, fine-tuning, or adaptive learning, which are not performed within this component. In multiple instances, the assessment explicitly notes that such measures are described—if at all—in model cards provided by the model developers, rather than being implemented or verified within the system.

The assessment comments make it clear that many N/A classifications reflect a transfer of responsibility rather than a true absence of risk. While the system relies on assurances from model providers, it does not independently validate these safeguards or implement compensating controls. Additionally, risk assessment results indicate that key mitigation actions—such as reducing model information exposure and assessing component vulnerabilities—have been identified but not implemented ("no action taken"), highlighting a gap between risk identification and mitigation.

Overall, the control framework of the "Match to Debunked Claims" component is characterised by a limited set of implemented safeguards focused on output restriction and controlled data sources, a large number of unimplemented but applicable controls related to security, monitoring, and governance, and a broad set of controls marked as N/A due to dependence on third-party models. This creates a system that benefits from a more curated knowledge base

compared to open web search, but still lacks sufficient protections against adversarial manipulation, operational failures, and governance risks.

3.1.4.3 Match to Other Articles

The vulnerability assessment of the "Math to Other Articles" component reveals a control landscape structured around three categories—implemented controls, non-implemented controls, and controls deemed not applicable (N/A)—with the distribution heavily influenced by the system's reliance on pre-trained large language models and third-party infrastructure. As with the Claim Detection module, many core AI-related controls are externalised to providers (e.g., Meta and Groq), while application-level and organisational controls remain largely unaddressed.

A limited set of controls is implemented, focusing on output restriction, workflow design, and partial retrieval safeguards. The system ensures that only textual outputs are returned through the API, without exposing internal model details such as confidence scores, probabilities, or reasoning traces, thereby reducing the risk of model probing and reverse engineering. Error handling is also implemented in a secure manner, with only generic internal server errors exposed to users, preventing leakage of internal workflow logic. Additionally, the workflow design constrains the model's behaviour by avoiding autonomous agency—LLMs do not independently select tools or actions, but instead operate within a predefined sequence of steps. Importantly, retrieval-augmented generation (RAG) is implemented, allowing the system to ground outputs in web-based sources; users are given some control to include or exclude domains, although this is optional rather than enforced.

In contrast, a large number of relevant controls are not implemented, particularly those related to security, robustness, monitoring, and governance. There are no input sanitisation or adversarial detection mechanisms in place, leaving the system exposed to prompt injection, malicious queries, and manipulation of retrieval results. Similarly, no continuous monitoring is performed to detect model drift, performance degradation, or anomalous system behaviour. Organisational controls are also absent: the component is not integrated into incident response processes, asset management systems, or security-by-design frameworks. Furthermore, there is no evidence of compliance assessments (e.g., GDPR or AI Act), nor are there implemented processes for documentation, auditing, or lifecycle risk management. Fairness-related controls—such as bias detection, fairness audits, and mitigation strategies—are also not implemented, despite the potential for biased or misleading retrieval results.

A substantial portion of controls are classified as not applicable (N/A), primarily due to the architectural reliance on pre-trained, open-source models and third-party infrastructure. Controls related to model training—such as adversarial training, dataset curation, poisoning mitigation, differential privacy, and federated learning—are outside the system's scope, as training is performed entirely by model providers. Similarly, controls requiring access to model internals (e.g., embeddings, parameters, training data) or ML pipelines are not applicable, since these components are not accessible. The same applies to fairness interventions at the model level and to lifecycle controls tied to training, fine-tuning, or retraining. In several cases, comments explicitly state that such responsibilities lie with the model publishers and are only documented in model cards, not enforced within the system itself.

The comments across the assessment highlight that many N/A classifications stem from a transfer of responsibility rather than true irrelevance. While the system does not perform training or manage infrastructure, it also does not validate or audit the safeguards claimed by external providers. For example, while RAG is implemented, the definition of "authoritative sources" is left largely to user discretion, and by default the system may retrieve content from any web resource, increasing the risk of misinformation propagation. Similarly, although the system assumes that training data consists of publicly available text, it does not independently verify data provenance or associated risks.

Overall, the Web Search component exhibits an imbalance in its control framework: a small number of implemented controls focused on limiting information exposure and structuring outputs, a broad set of unimplemented but applicable controls related to security, monitoring, and governance, and a large number of controls marked as N/A due to dependence on third-party models. This results in a system that is functionally capable—particularly through its use of RAG—but insufficiently safeguarded against adversarial manipulation, biased retrieval, and operational risks.

In conclusion, the vulnerability assessment demonstrates that while the Web Search component incorporates some baseline safeguards (e.g., output restriction and controlled workflow execution), it lacks critical application-level and organisational controls, and relies heavily on external providers for core AI assurances. Strengthening the system would require implementing missing controls within its scope—such as input validation, monitoring, compliance processes, and source verification—while also establishing stronger assurance and validation mechanisms for third-party dependencies.

3.1.4.4 Hate Speech Detection

The vulnerability assessment highlights a system with a broad exposure to risks across multiple trustworthiness dimensions, largely driven by its architectural reliance on pre-trained, third-party language models and the absence of implemented mitigation controls. While risks such as adversarial attacks, misinformation propagation, bias, lack of explainability, and model/data disclosure are clearly identified, the assessment repeatedly indicates that no concrete mitigation actions have been taken. This creates a situation where critical threats—including poisoning, prompt attacks, evasion, and component compromise—remain unaddressed, elevating the risk level of key trustworthiness characteristics such as safety, security, fairness, and reliability to high or very high. The lack of transparency and accountability mechanisms, including monitoring, documentation, and auditability, further compounds these vulnerabilities.

At the same time, the assessment reveals that many controls are either not implemented or considered outside the system's scope due to dependence on external model providers. As a result, safeguards related to training data governance, adversarial robustness, fairness mitigation, and privacy preservation are effectively delegated without independent verification. Within the system's own scope, essential controls—such as input validation, continuous monitoring, incident response integration, and compliance processes—are largely absent. This imbalance between identified risks and implemented safeguards leads to a weak overall security and governance posture, where vulnerabilities persist not because they are unknown, but because they remain untreated or externalised, leaving the system exposed to both technical and operational threats.

3.1.5 Phase 5: Risk Analysis

Following the FAITH AI_TAF V1 methodology, the risks of all AI assets of the AI system per trustworthiness dimension is estimated in terms of:

- Threat Level: The frequency of occurrence of a threat (Phase 2).
- Vulnerability Level: System's susceptibility based on control measures (Phase 4).
- Consequence Level: The potential consequences of the threat to all dimensions of trustworthiness (automatically calculated through TrustGuard) along with the business consequences from the potential compromise of these trustworthiness dimensions (Phase 3).
- Trustworthy AI maturity of teams and trustworthiness of AI participants (Phase 1).

For each trustworthiness dimension, the corresponding risk levels were automatically calculated per identified threat using the TrustGuard tool, in accordance with Table 3.

For example, within the "Accountability and Transparency" dimension, the risk level associated with the threat "model or data disclosure" is assessed as very high. This results from the combination of a very high threat level and its dependency on a single control—"conduct a risk analysis of the ML application"—which has not yet been implemented for any of the AI models within the Intelligent Coaching application.

Threat Level	Vulnerability Level	Consequence Level	Risk Level
Very High	Very High	Very High	Critical
Very High	High	High	Severe
Very High	Medium	Medium	High
Very High	Low	Low	Medium
High	Very High	High	Severe
High	High	Medium	High
High	Medium	Low	Medium
High	Low	Low	Low
Medium	Very High	Medium	High
Medium	High	Medium	Medium
Medium	Medium	Low	Low
Medium	Low	Low	Minimal
Low	Very High	Low	Medium
Low	High	Low	Low
Low	Medium	Minimal	Minimal
Low	Low	Minimal	Minimal

Table 3. Risk Calculator Matrix (D2.1)

To calculate the final risk level the trustworthiness of the organizational AI maturity level (tAIP) estimated in Phase 1 has also been taken into account by utilizing Table 4:

Risk Level (R)	tAIP	Final Risk Level (fR)
Very High	Very High	High
Very High	High	High
Very High	Medium	High
Very High	Low	Very High
High	Very High	Medium
High	High	Medium
High	Medium	Medium
High	Low	Very High
Medium	Very High	Low
Medium	High	Low
Medium	Medium	Low
Medium	Low	High
Low	Very High	Very Low
Low	High	Very Low
Low	Medium	Very Low
Low	Low	Medium

Table 4. Final Risk estimation-2 (D2.1)

3.1.6 Phase 6: Countermeasures

This phase includes an initial review of the output of phases 1-5 for human-centred validation of findings and initial planning of needed mitigations. On this basis, the risk management activities

concern devising technical solutions like implementing robust data governance practices, access control mechanism, continuous model testing, encryption protocols, firewalls, and intrusion detection systems; It also involves human oversight measures such as training programs, co-creation workshops, policies, and procedures to enhance user awareness, trustworthy behaviour and technical skills.

At this phase of the project, the suggested mitigation actions are mainly technical and stem from the FAITH Golden Table.

The methodology used to evaluate and respond to the various mitigation actions follows a context-aware gap analysis that balances technical feasibility against organizational constraints. Each suggested control was filtered through three primary lenses: Architectural Scope, Developmental Agency, and Organizational Integration. By categorizing the controls this way, the development team could realistically distinguish between risks they can directly manage and those inherent to the use of third-party, pre-trained Large Language Models (LLMs).

Under this methodology, any mitigation requiring internal model adjustments—such as adversarial training, embedding invariance testing, or fine-tuning—was classified as "Ignored". This is a pragmatic decision since the system utilizes external APIs (Meta/Groq), the team lacks the "access to internal model parameters" necessary to implement feature-level fairness or structural robustness. Similarly, many governance-related mitigations were also marked as "Ignored" because they depend on Organizational Integration, requiring "senior architects" or "company-wide restructuring" that falls outside the immediate technical scope of the individual project modules.

Conversely, the actions marked as "Treated" are the actions focused on Post-Processing and Deployment-Side Controls. This approach prioritizes “Control Data Used by the ML Model" and "Detect patterns of probing or prompt-revealing questions" which are active measures that can be wrapped around a third-party model. By focusing on sanitizing free-form text and monitoring API performance indicators, the team implements a defense-in-depth strategy that protects the application without needing to modify the underlying AI architecture. This methodology ensures that the risk treatment plan is not a "wishlist," but a functional roadmap tailored to the realities of an agentic pipeline built on top of external foundational models. The full list of the countermeasures is presented in Section 3.5.

3.2 Tool Validation Planning & Training Processes

This section provides an overview of the training and validation activities carried out for the various FAITH tools, together with the implementation plan for the upcoming phases of the project.

3.2.1 Dry-run

Based on the FAITH AI_TAF V1, a structured questionnaire was developed by ATC in Excel format, operationalising the key phases of the framework, namely Cartography (Phase 1), Threat Assessment (Phase 2), Impact Assessment (Phase 3), and Vulnerability Assessment (Phase 4). This questionnaire was distributed to the Large-Scale Pilot (LSP) partners to evaluate the quality, relevance, completeness, and answerability—in relation to end-users' expertise—of the questions posed through the TrustGuard tool. For brevity, this process is hereafter referred to as "dry-run".

For the dry-run, we focused on the assessment of the Hate Speech Detection model. The various phases were allocated among partners according to their expertise. The FH partners completed the Impact Assessment (Phase 3), while ATC was responsible for the Cartography (Phase 1), Threat Assessment (Phase 2), and Vulnerability Assessment (Phase 4). More specifically, the Cartography and the Threat Assessment was carried out by the Risk Manager, who is at the same time AI expert, whereas the Vulnerability Assessment (for the Design lifecycle stage) was completed by the owner (AI-developer) of the asset and two AI experts and one DevOps who had good technical knowledge of the tool. The spreadsheet can be found here[9]. A screenshot of the dry-run is presented in Figure 11.

9 https://docs.google.com/spreadsheets/d/1pI6l_RND_0ORxkIQmhqYsxXR4UMnNiQ5/edit?usp=sharing&ouid=117555197365128641367&rtpof=true&sd=true



Figure 11 Dry-run exercise for the Hate Speech detection tool.

Overall, out of the 25 questions included in the dry-run questionnaire, 9 were considered too vague, while 6 were deemed non-applicable to LSP1.

3.2.1.1 Vague Questions

A number of questions were reported as difficult to answer due to their lack of clarity and specificity. Indicative examples include: *"Have you implemented processes to maintain security levels of Hate_Speech_Detector components over time?"*, *"In the context of Hate_Speech_Detector, have model specificities been integrated into existing security policies?"*, or *"In the context of Hate_Speech_Detector, has it been ensured that the model follows the global process for integrating security?"*.

Engineers reported difficulties in interpreting key terms within these questions. For instance, concepts such as *"model specificities"* and *"global processes"* were not clearly defined, making it challenging to provide consistent and meaningful responses. This ambiguity limited the effectiveness of the questionnaire as an assessment tool.

3.2.1.2 Non-applicable Questions

At the time of its distribution to the LSP partners (11 April 2025), the FAITH Golden Table included only threats and vulnerabilities derived from ENISA (2021), which primarily focuses on machine learning (ML) algorithms. As a result, the dry-run questionnaire contained questions tailored to traditional ML systems.

However, the AI Intelligent Coaching application is based on pretrained large language models (LLMs) and relies heavily on prompt engineering techniques, rather than model training or fine-tuning within the project scope. Consequently, several questions were not applicable to LSP1.

For example, the question *"Have you enlarged the training dataset by adding diverse data to increase its robustness to poisoning attacks?"* is not relevant in this context, as no training dataset is managed within the application. Similarly, questions such as *"Does the Hate_Speech_Detector model respect differential privacy?"* or *"In the context of CONAN, have you applied an RBAC model respecting the least privilege principle?"* were considered non-applicable, given the system architecture (e.g., use of externally hosted models or publicly accessible endpoints).

3.2.1.3 Missing Questions

At the same time, the dry-run highlighted that LLM-specific threats and vulnerabilities were not sufficiently covered. In particular, aspects such as prompt injection, hallucinations, context manipulation, and risks associated with third-party model dependencies were not adequately reflected in the questionnaire.

This indicates a clear need to extend and adapt the FAITH Golden Table to better capture the risk landscape of LLM-based systems, ensuring that future assessments are both technically relevant and context-aware for applications such as the AI Intelligent Coaching tool.

As a result of the dry-run, the FAITH Golden Table was updated to incorporate threats and vulnerabilities identified across the different LSPs. This update process remains ongoing and will continue throughout the project to ensure that the framework reflects emerging insights and evolving risks.

After this dry-run process, Media LSP added the following threats, vulnerabilities and mitigation actions to the FAITH Golden Table:

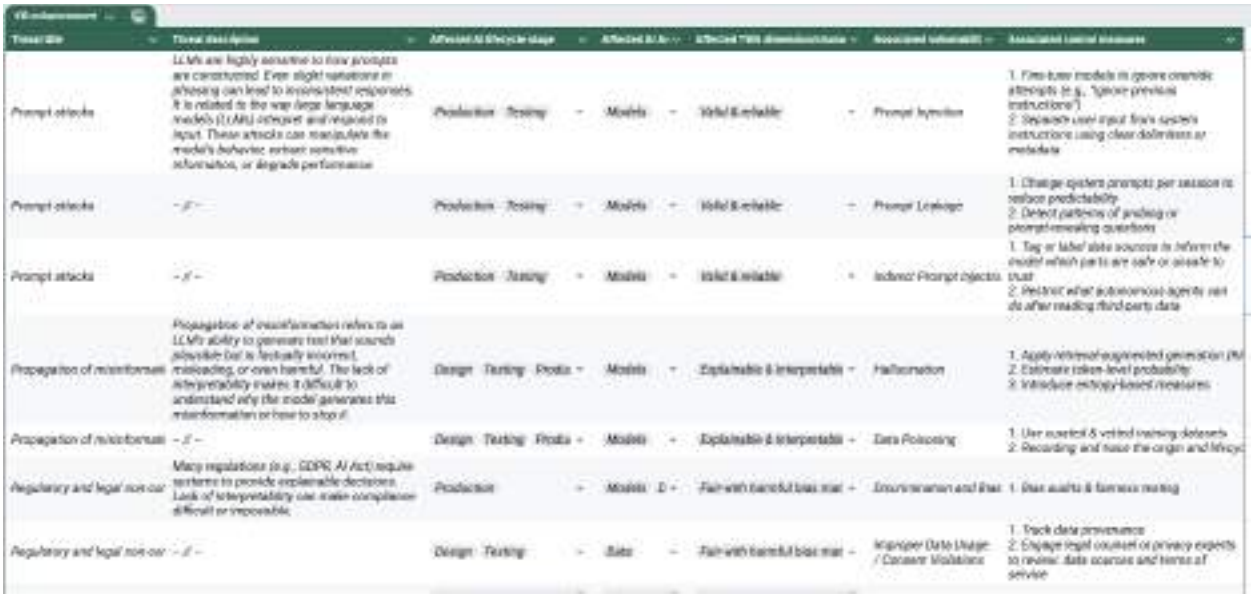


Figure 12 Overview of threats, vulnerabilities, and mitigation actions identified during the Media LSP dry-run

3.2.2 Plan development

In order to ensure effective coordination and alignment between the various LSPs and the activities of WP2, a common set of activities defined by WP2 leader was established during the project plenary meeting in January 2026 in Oslo with the aim of facilitating a structured and harmonised implementation across pilots. Each LSP defined their own timeplan based on project deadlines and taking into account the specific particularities of their pilot.

The Media LSP partners agreed on the following timeplan:

Training and Support phase

Targeted training activities on key FAITH components:

- Training of ATC to AI Model Passport by M26,
- Training of ATC to AI Data Passport by M26,
- Training of FH to TrustGuard by M26
- Training of ATC & FH to TrustSense by M23,
- Training of ATC to SSM by M33 and
- Training of ATC & FH to FAITH STM Platform M33.

All training activities by M26 have been successfully completed.

FAITH Tools Implementation and Iteration

This phase includes the deployment of the system under test (SUT) which was already completed within T4.1, the use of tools such as TrustSense for team maturity assessment (M27), AI Model Hub (M26) and TrustGuard for in-isolation assessment of the AI model and data assets (M26) and feedback collection activities for these tools (M27). Additionally, it includes the application of one or more controls suggested TrustGuard and the reassessment of the Intelligent Coaching application. All these activities have been successfully completed.

Future activities involve the application of multiple controls suggested by TrustGuard to mitigate as much as possible the relevant risks, continuous enhancement of the FAITH Golden Table with vulnerabilities, threats and mitigation actions relevant to the Media LSP needs. Also, future activities involve reassessment of the next iteration of the services AI Model Hub v2, TrustGuard and TrustSense (M30–M36), the use and refinement of the FAITH STM Platform (M34–M36), and the development and application of SSM models (M33–M35) for in-context assessment, followed by their continuous improvement (from M36 onwards).

Finally, the result of all these thorough assessment and refinement processes will culminate in the implementation of the AI Intelligent Coaching service in the Replication Pilot during the period M36–M38, marking the transition from the initial pilot phase to broader validation and scaling activities.

3.2.3 Tool Training

3.2.3.1 AI Model Hub

The training on the AI Model Hub tool, provided a comprehensive introduction to managing machine learning assets in a centralized platform. Participants, consisting mostly of development team members trained by FORTH on the 2nd of February, 2026, learned how to upload model definitions, organize datasets and attach detailed metadata to improve discoverability and reproducibility. The session combined guided instruction with hands-on examples, allowing users to practice versioning AI assets (models & datasets). Not much emphasis was placed on the process of automatic metadata uploading given the advanced implementation status of the Persona Companion AI models/components. It was also clarified how to properly tag the defined AI assets in order to be transparently ingested into the TrustGuard application.

3.2.3.2 TrustGuard

As ATC is the technical developer of TrustGuard and therefore already highly familiar with the tool, training for LSP1 was provided exclusively to FH during one of the WP4 meetings. During this session, ATC presented the process for accessing and using the system, clarifying which

components of the assessment are intended for AI users (i.e., non-developers). Particular emphasis was placed on the Consequence Assessment, with ATC providing a detailed explanation of the relevant concepts and terminology to ensure a consistent understanding among participants.

3.2.3.3 TrustSense

Training for TrustSense was provided in order to get all LSPs acquainted with the features and results used for evaluating the team's level of maturity in trustworthy AI. In order to facilitate ongoing reflection and development, the application offers role-specific questions and visual trust profiles. It also generates trust scores and suggests mitigation techniques. The tool's risk assessor flow, which assesses teams along seven major dimensions, was shown to the participants. These include:

- **Proactivity & threat awareness**: early identification and mitigation of potential AI-related human risks.
- **Responsibility & ethics**: ensuring fairness and adherence to ethical norms.
- **Innovation & adaptability**: learning from feedback and improving practices.
- **Resilience**: effective responses to failures or incidents linked to human vulnerabilities.
- **Collaboration & knowledge sharing**: supporting team communication and coordinated mitigation actions.
- **Technical proficiency**: maintaining an understanding of AI challenges and limitations.
- **Policy Adherence**: ensuring compliance with relevant regulations and governance structures.

The implementation of mitigation techniques, the analysis of risk profiles, and the interpretation of trust scores were all given special attention during the course.

3.2.3.4 2.3. FAITH Tool Validation Activities

3.2.4 Data & AI model Passport in the FAITH AI Model Hub (ATC)

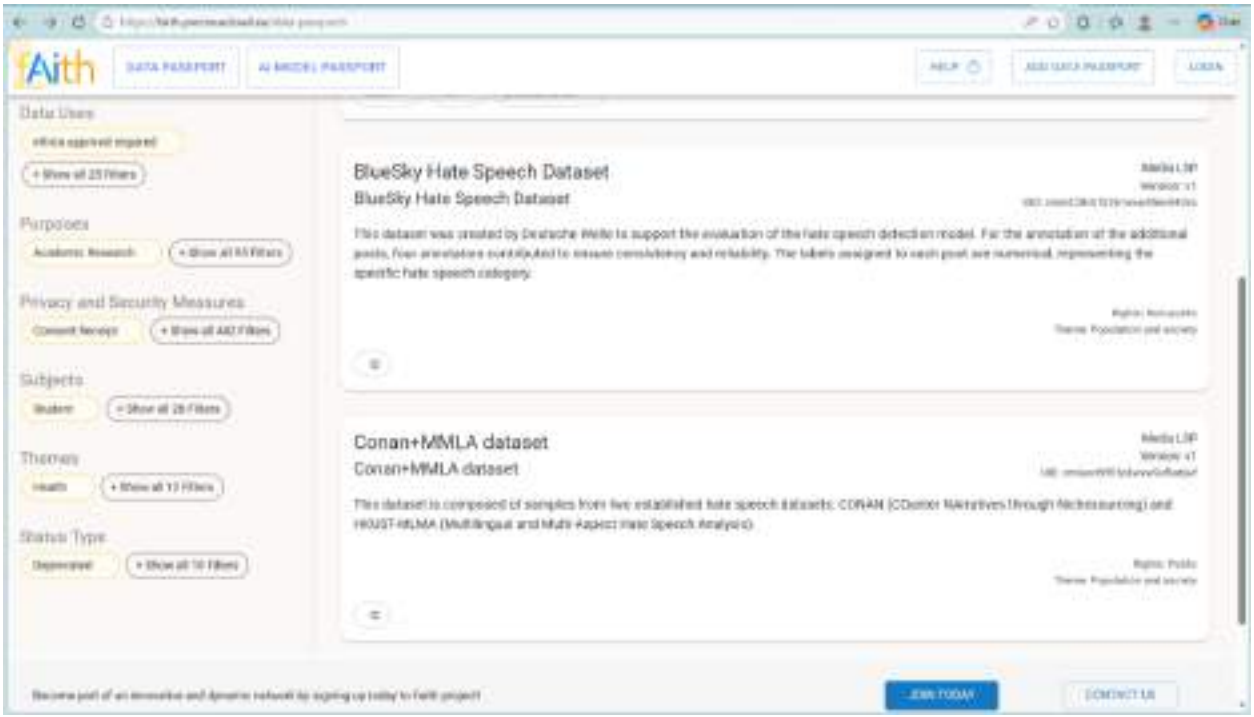


Figure 13 Datasets used for the Hate Speech Detection Model in AI Data Passport.

The ATC developer (and owner) in charge of the Hate Speech Detection tool documented their workflow on the FAITH AI Model Hub. The tool was accessed on February 21st, 2026. Because this tool functions as an AI workflow relying on external calls to pre-trained LLMs, the engineer utilized the Data Hub specifically to register the evaluation datasets used to assess the tool's performance (Figure 12). Since no local model training was performed, only these testing datasets were documented.

For the Model Passport, the relevant sections were completed to reflect this architecture. In the Experiment and Details sections, the developer provided the necessary context for the Media LSP pilot, along with a risk assessment and ethical evaluation concerning the workflow's internal structure and its reliance on external inference (i.e., Groq and the Llama family of open-source models). Finally, the Pipeline, Algorithm, and Configuration sections were not applicable and were left blank, as the hate speech tool utilizes pre-trained models rather than being developed or trained from scratch by ATC.

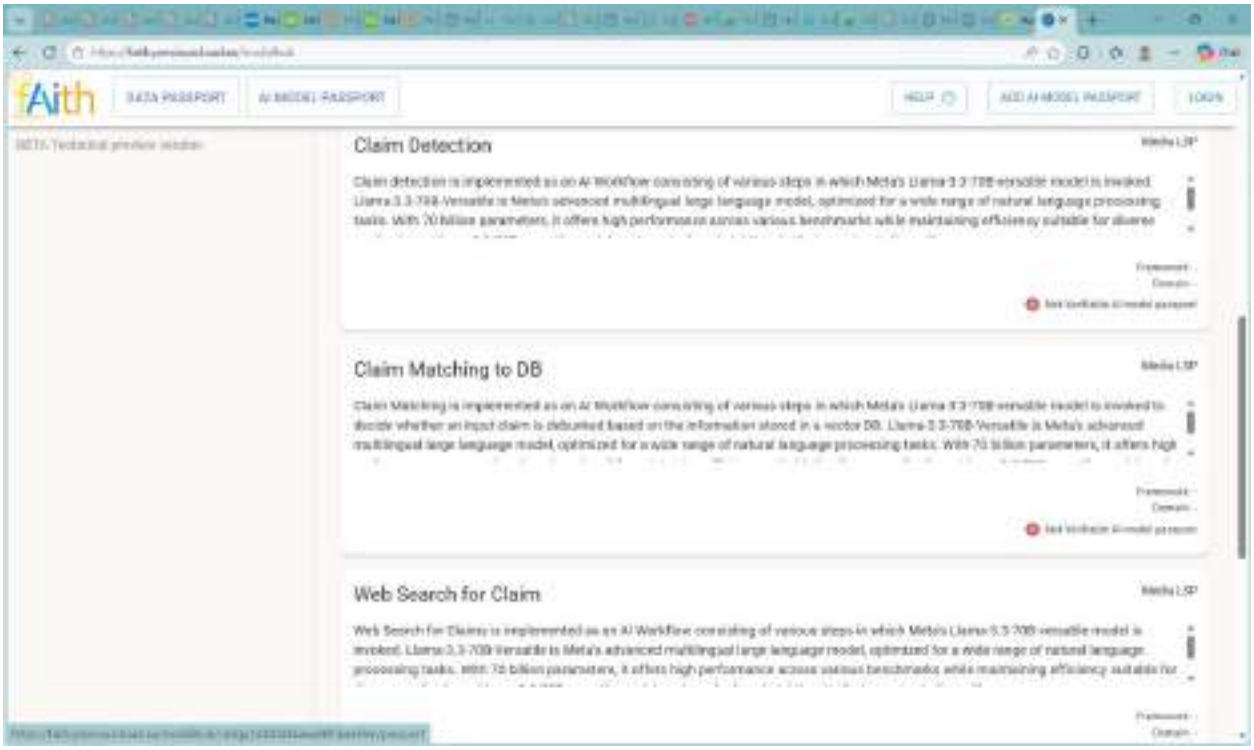


Figure 14 Models of Claim Analysis functionality in AI Model Passport.

The FAITH AI Model Hub tool (https://faith.preciouscloud.eu) was also used by ATC's AI engineer in charge of developing the claim analysis tools. The tool was accessed on Feb 17th, 2026. During this session ATC's engineer supplied the model information as determined by the model hub's forms, for three functionalities developed in the Media LSP: claim detection, web search for claims, and matching claims to DB. Note, that for the tools related to claim analysis, the data hub tool does not apply, given that these tools are AI workflows relying on external calls to pre-trained LLMs.

For all three functionalities the relevant model passport sections were submitted, where applicable (Figure 13). In particular, in the Experiment and Details sections, information was provided on which pilot these AI workflows were developed for, and a risk assessment and ethical concerns estimation was provided based on the workflow's internal structure and external dependencies, i.e., the LLM inference provided by Groq on pre-trained open source models used (Llama family of LLMs) The Pipeline, Algorithm, and Configuration sections did not apply, as the models used by these workflows are pre-trained, i.e., not developed/trained by ATC.

3.2.5 TrustGuard

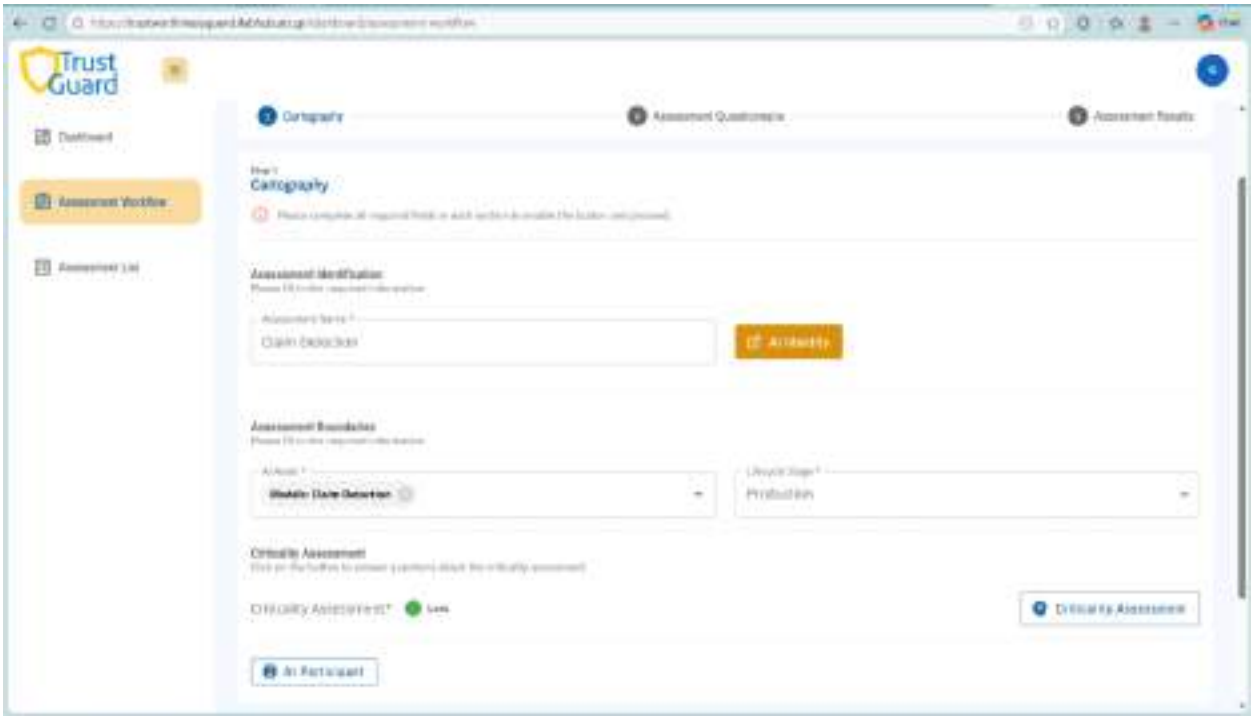


Figure 15 Cartography phase in TrustGuard for the Claim Detection Model.

During March 2026, TrustGuard was used collaboratevelly by the owners of each component of the AI Intelligent Coaching application, AI end-users and the risk assessor. Specifically, an account for LSP1 was created by the Risk Assessor (ATC) and shared among the partners. As discussed in Section 3.1.1, as in this stage of the project, the risk assessment is conducted in-isolation, the 4 separate risk assessments were conducted, one for each AI model (Claim Detection, Matching to Debunked Claims, Matching to Other Articles, Hate Speech Detection). The Risk Assessor completed the Cartography Phase (Figure 14) for each risk assessment, by providing the models (and the data for the hate speech detection case) for assessment, the lifecycle stage under assessment (Production) and by completing also the Criticality Assessment as discussed in Section 3.1.1. Additionally, the Risk Assessor, being at the same time AI-expert, completed the Threat Assessment by following the methodology described in detail in Section 3.1.2.

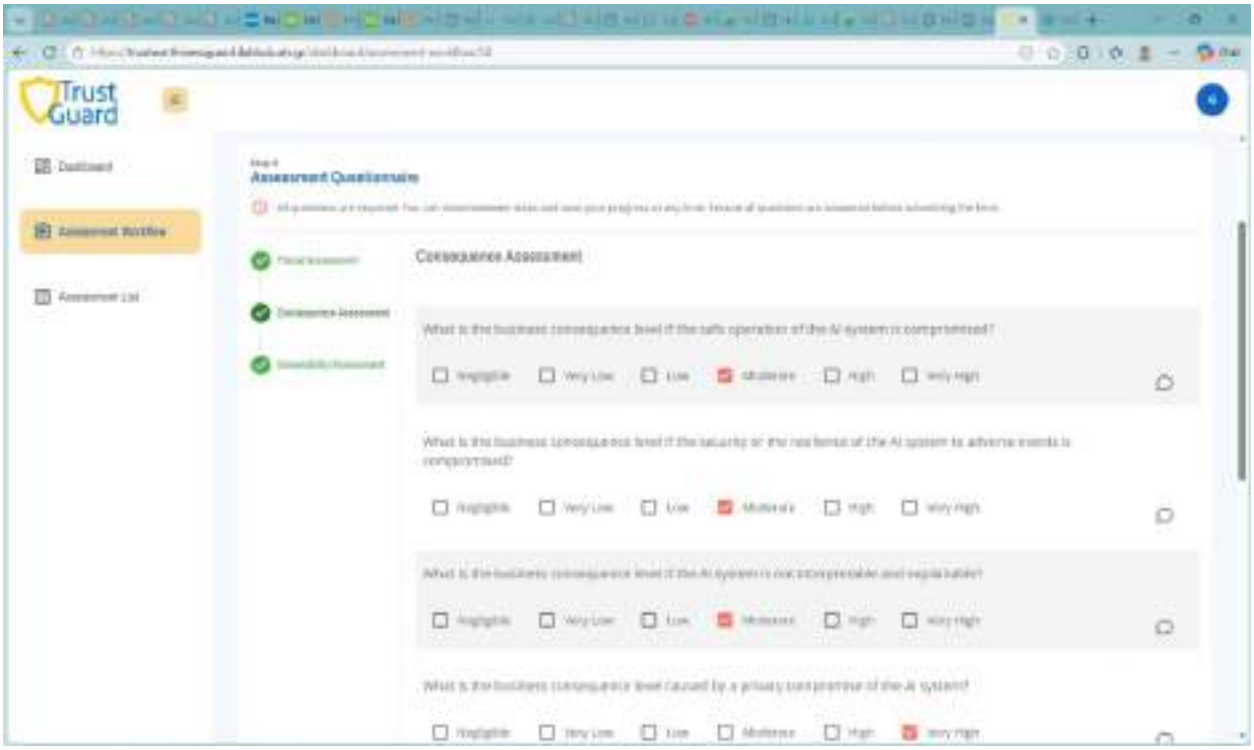


Figure 16 Consequence Assessment in TrustGuard in Claim Detection filled in by FH.

FH, representing the AI end-users, inserted the outputs of TrustSense (presented in Section 3.3.3) with respect to the trustworthy AI maturity of teams into TrustGuard. Additionally, FH completed the Consequence Assessment following the methodology described in Section 3.1.3, for each risk assessment (Figure 15).

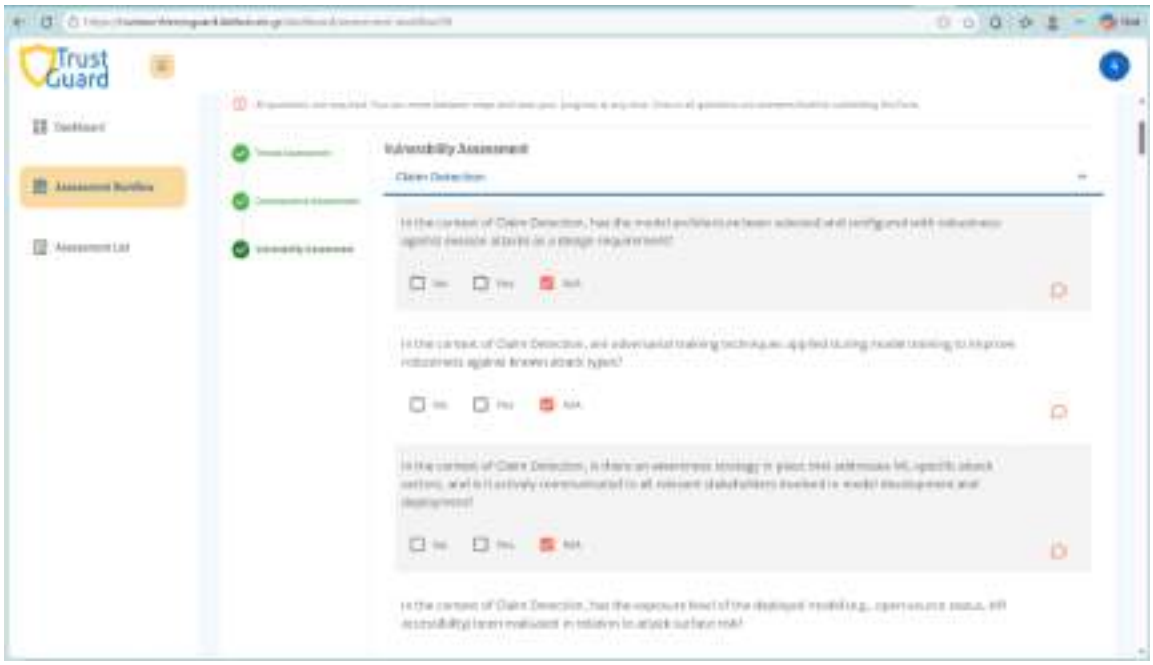


Figure 17 Vulnerability Assessment for the Claim Detection model.

Finally, the owner (ATC) of each AI model and dataset completed the various Vulnerability Assessments (Figure 17) for each risk assessment and reviewed the respective countermeasures suggested by TrustGuardFigure 18).

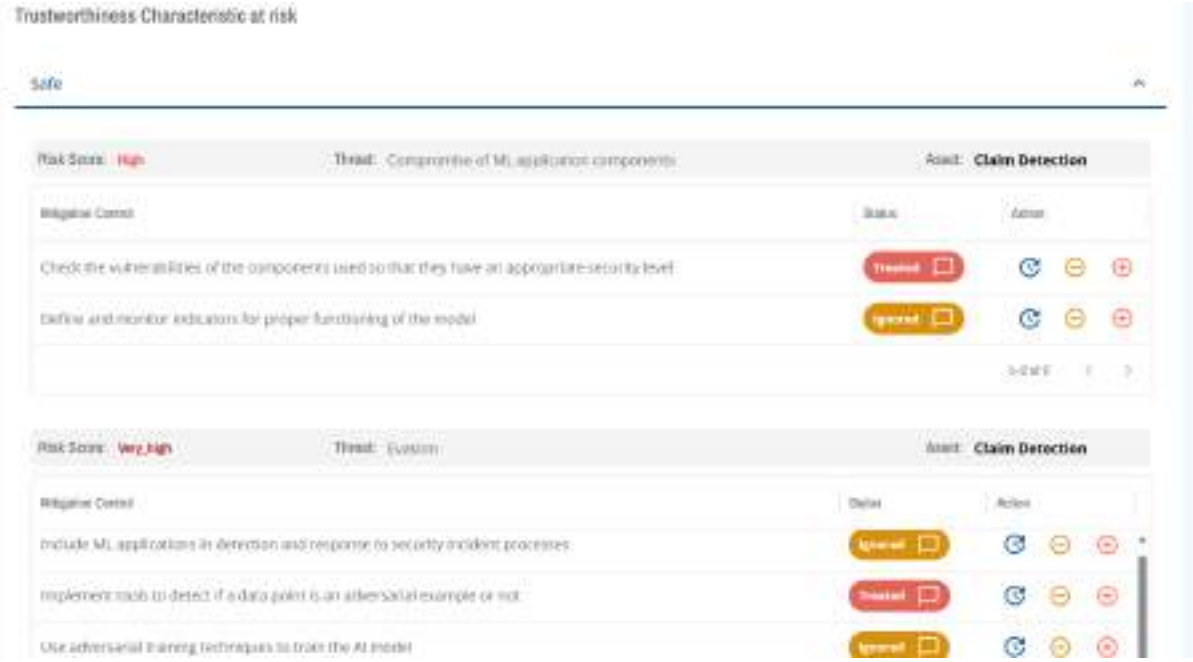


Figure 18 Risk Assessment and Mitigation Actions suggested by TrustGuard.

3.2.6 TrustSense

TrustSense evaluates teams' level of maturity with respect to reliable AI practices. Some of the aspects of trustworthy AI that are assessed are proactivity and threat awareness, accountability and ethics, creativity and adaptation, resilience, cooperation and information sharing, technical competency, and policy adherence are

On February 17, 2026, the FH team conducted the TrustSense assessment during an online webinar with journalists from PRESShub network. The activity's goal was to assess participants' knowledge of and readiness for reliable AI practices when using AI-supported technologies created as part of the FAITH project.

Two members of the FH team gave an overview of the assessment's goal and its methodology at the start of the meeting. After that, during the webinar, everyone answered the questionnaire. Every question was examined in detail, and when needed, technical jargon were explained. Throughout the discussion, the questions were also translated in real time from English into Romanian to help everyone understand. The results of the assessment, including the scores and corresponding maturity levels for each TrustSense dimension, are presented below in **Table 5**. However, it is important to note that the TrustSense assessment provides a high-level, indicative overview of participants' maturity with respect to trustworthy AI practices. Also, the results are based on a single assisted session and show aggregated scores across dimensions. Therefore, the results should be viewed as a preliminary diagnostic snapshot that can be used to determine general strengths and areas that need work.

TrustSense Dimension	***Score***	***Level***
Proactivity & Threat Awareness	*3.67*	*Moderate*
Responsibility & Ethics	*5.00*	*High*
Resilience	*2.50*	*Moderate*
Collaboration	*5.00*	*High*
Integrity	*5.00*	*High*
Technical Proficiency	*5.00*	*High*
Problem Solving	*5.00*	*High*
Resource Accessibility	*2.00*	*Low*
Policy Adherence	*3.00*	*Moderate*
Motivation & Commitment	*5.00*	*High*
Privacy & Compliance	*3.00*	*Moderate*

Table 5. TrustSense dimensions with corresponding levels and scores

A high degree of credibility within the AI User Team area was indicated by the assessment's trustworthiness score of 4.40 (see **Figure 17**). This score shows that the team mostly complies with trustworthy AI norms, with only a few minor areas needing improvement, according to the TrustSense interpretation framework. Continuing focused efforts to improve trustworthiness across particular dimensions, guaranteeing access to up-to-date tools and resources, and promoting knowledge sharing with high-performing teams are all suggested courses of action.

The findings show that participants had a generally high degree of trustworthiness awareness, especially in areas like motivation, technical skill, responsibility and ethics, and teamwork. Opportunities to improve risk awareness and governance-related knowledge are indicated by moderate ratings in areas including proactivity, policy adherence, and privacy awareness. Resource accessibility had the lowest score, suggesting that users dealing with AI-based systems require better access to tools, documentation, and supporting resources. **Figures 18–22** present the mitigation actions per category, together with their corresponding scores and levels, across all TrustSense dimensions, as also summarized in Table 5.

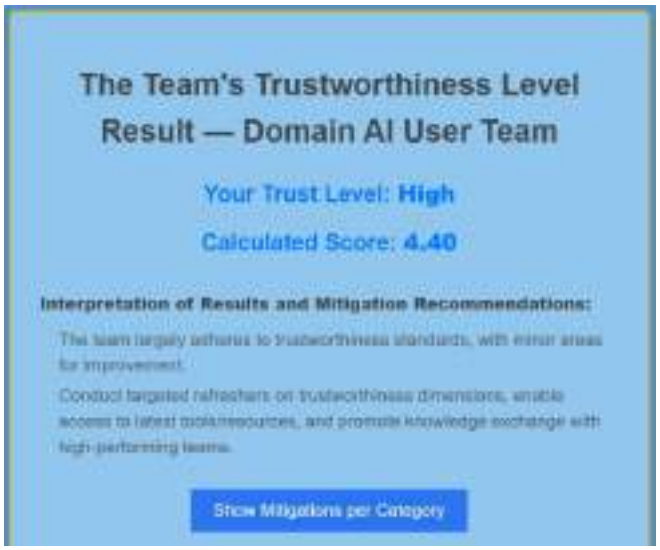


Figure 19 TrustSense assessment results for the AI User Team domain (FH).



Figure 20 Mitigation actions per category, showing scores and levels for Proactivity, Threat Awareness, and Responsibility and Ethics.



Figure 21 Mitigation actions per category, showing scores and levels for Proactivity, Threat Awareness, and Responsibility and Ethics.



Figure 22 Mitigation actions per category, showing scores and levels for Resilience, Collaboration, Integrity, and Technical Proficiency



Figure 23 Mitigation actions per category, showing scores and levels for Problem Solving, Resource Accessibility, Policy Adherence

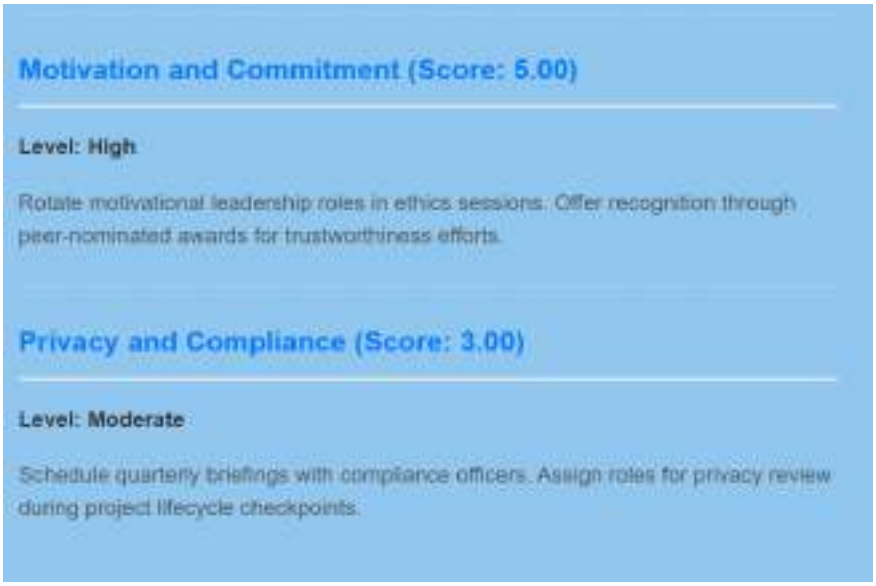


Figure 24 Mitigation actions per category, showing scores and levels for Motivation and Commitment, and Privacy and Compliance.

3.3 Threat Analysis

Table 6. Threats relevant to the four functionalities of the AI Intelligent Coaching service.

Threat	Claim Detection	Matching to DB	Matching to Other Articles	Hate Speech Detection
Compromise of ML application components	X	X	X	
Data disclosure	X		X	
Denial of service due to inconsistent data or a sponge example	X	X	X	
Evasion	X		X	X
Failure or malfunction of ML application	X	X	X	X
Human error	X	X	X	X
Model disclosure	X		X	
Model Drift Leading to Inaccurate Assessments				
Model or data disclosure	X		X	X
Oracle	X		X	X
Poisoning	X	X	X	
Prompt attacks	X	X	X	X
Propagation of misinformation	X	X	X	X
Regulatory and legal non-compliance				
Use of adversarial examples crafted in white or grey box conditions				

According to TrustGuard, the main list of threats that the AI Intelligent Coaching application is under is presented in Table 6. Some significant threats include:

- Three out of the four AI models are under the threat of **Compromise of ML application components** due to the lack of formal organizational integration (specifically the failure to integrate ML specificities into organizational security policies and asset management).
- The assets Claim Detection, Claim Matching to DB and Claim Matching to Other Articles are under the threat of **Denial of Service** (Sponge Examples), the models use third-party services, hence it is impossible to control the data used by the model—across training, validation, and inference.
- The assets Claim Detection and the Match to Other Articles models are under the threat of **Model or Data Disclosure** due to the Lack of Specialized Risk Analysis vulnerability, as these modules have not undergone red-teaming to protect their API interactions and organizational security policies need to be investigated.
- All models are under the threat of **Prompt Attack** (specifically listed as "Adversarial Examples/Input") due to the Third-Party Data Trust Vulnerability, where the model may be forced to follow malicious instructions embedded in retrieved web content.

- All models are under the threat **Propagation of Misinformation**, mainly due to the use of third-party models, hence training, fine-tuning or conditioning are not performed.
- The Claim Detection, Match to Other Articles and Hate Speech Detection are subject to Evasion attacks

3.4 Initial Risk Assessment results

The risk assessment across the four core functionalities—Hate Speech Detection, Match to Other Articles, Match to Debunked Claims, and Claim Detection—reveals a significant concentration of Very High risks primarily within the “Accountable and Transparent”, “Robust and Accurate”, and “Safe” trustworthiness characteristics. The "Accountable and Transparent" property is consistently compromised across all modules due to the threat of Compromise of ML application components. This is largely an organizational vulnerability, as the application currently lacks integration into security-by-design processes and global cyber-resilience strategies. In the Web Search and Hate Speech modules, this risk is further compounded by the threat of “Model or data disclosure”, where a lack of specialized red-teaming and risk analysis leaves the system's external API interactions unvetted against sophisticated extraction attacks.

Regarding the technical resilience of the application, the “Robustness and Accuracy” and “Safety” trustworthiness characteristics are under severe threat from Denial of Service (DoS) attacks. Specifically, the system is vulnerable to "Sponge Examples"—adversarial inputs designed to cause excessive computational resource consumption—and inconsistent data that can trigger logic failures. The Claim Matching to DB and Claim Detection modules are particularly susceptible to the threat “Changes in external conditions”, where the lack of active performance monitoring and the inability to retrain third-party models could lead to a silent degradation of accuracy. These technical risks are exacerbated by the current status of mitigation controls, many of which are marked as "Ignored" due to the team’s reliance on pre-trained LLMs, which limits their ability to implement internal architectural fixes or adaptive learning techniques.

Finally, the Fairness property is identified as being at “Very High” risk, particularly in the Hate Speech Detection module. As the system relies on external embeddings provided via third-party APIs, the developers lack the necessary access to internal model parameters to test for embedding invariance or to apply adversarial training that could suppress group identifiers. This creates a governance challenge where the application may inadvertently produce biased moderation outcomes without a technical mechanism for the developers to intervene at the feature level.

3.5 Countermeasures

Based on the vulnerability assessments for all four modules—Claim Detection, Match to DB, Match to Other Articles, and Hate Speech Detection—the following mitigation actions are known to be in place (Treated) by the owners (AI-developers) of the respective AI Models:

Table 7. Countermeasures denoted as "Treated" in TrustGuard.

Mitigation Action	Comment	Trustworthiness Dimension	Asset
Conduct a risk analysis of the ML application	Risk analysis from the Risk Assessor in colab with the asset owners	Accountable and transparent	Claim Detection, Matching to DB,
Include ML applications into asset management processes	Handled by company processes regarding software component ownerships on repos etc.	Accountable and transparent, Secure and Resilient, Valid and reliable, Valid and Reliable	Claim Detection, Matching to DB, Mathcing to Other Articles
Define and monitor indicators for proper functioning of the model	Typical indicators of model performance would be or high classification accuracy. When the DB gets populated with additional claims we can re-run our accuracy evaluations and adapt the system	Accountable and transparent, Valid and Reliable, Safe, Valid and Reliable	Claim Detection, Matching to DB, Mathcing to Other Articles, Hate Speech Detector
Reduce the information given by the model	Experiment with prompts	Accountable and transparent, Valid and Reliable, Safe, Valid and Reliable	Hate Speech Detector
Reduce the amount of output information that the model gives	Prompting will restrict output as much as possible	Safe, Secure and Resilient	Hate Speech Detector
Apply documentation requirements to AI projects	Repositories within the company come with their README including any documentation.	Explainable and interpretable	Claim Detection, Matching to DB, Mathcing to Other Articles
Assess the regulations and laws the ML application must comply with	Monitoring of regulations with respect to use of models in the EU and examine alternative models should any restrictions apply.	Fair with harmful bias managed	Claim Detection, Matching to DB,
Implement tools to detect if a data point is an adversarial example or not	In the context of LLMs an adversarial data point would be an input prompt aiming to manipulate the model into unsafe behaviour. Addition of a guardrail layer examining inputs for safety (e.g., using a guardrail model to examine input / final output)	Safe, Secure and Resilient	Claim Detection, Matching to DB, Mathcing to Other Articles, Hate Speech Detector
Apply modifications to inputs	Wrapoing of the article text in a fixed and strict system prompt to ensure the model treats all input as data rather than new instructions	Safe, Secure and Resilient	Hate Speech Detector
Control all data used by the ML model	Implementation of input data checks: input should be free form text, excluding special tokens/strings that denote system/user message	Secure and Resilient	Claim Detection, Matching to DB, Mathcing to Other Articles

	role formats (model deployment phase applicable)		
Adopt assessment tools to detect possible malfunctioning of the system	It could be possible to periodically run claim detection against some texts and examine measures such as AlignScore to ensure that extracted claims keep aligned with the meaning of the text they were extracted from.	Secure and Resilient	Claim Detection, Matching to DB,
Build explainable models	Include traces regarding individual steps in the Claim Matching workflow, showing the user the step by step decisions taken by the workflow.	Valid and Reliable	Claim Detection, Matching to DB,
Restrict what autonomous agents can do after reading third-party data	This is handled since our workflows are deterministic and do not allow for LLM-based decisions on third party tools	Valid and Reliable	Claim Detection, Matching to DB,
Detect patterns of probing or prompt-revealing questions.	Options exploration to examine input for requests to reveal system prompts.	Valid and Reliable	Claim Detection, Matching to DB,
Separate user input from system instructions using clear delimiters or metadata	Revisit prompt structures and separate system instructions from user data.	Valid and Reliable	Claim Detection, Matching to DB,
Control all data used by the ML model	Implementation of input data checks: input should be free-form text, excluding special tokens/strings that denote system/user message role formats (model deployment phase applicable)	Valid and Reliable	Matching to DB,
Ensure that models are unbiased	Already done with validation datasets, will explore the need for further testings	Fair with harmful bias managed	Hate Speech Dataset
Bias audits & fairness testing	Already done with validation datasets, will explore the need for further testings	Fair with harmful bias managed	Hate Speech Dataset

1. Governance and Asset Management

- Asset Management Integration: ML applications have been included in formal asset management processes, handled through company protocols for software component ownership on repositories.
- Documentation Compliance: Documentation requirements have been applied to AI projects, with repositories including mandatory README files and detailed documentation.

The following mitigation actions will be treated by the owners (AI-developers) of the respective functionalities, outlining the specific technical and operational steps the development team will implement:

1. Governance and Asset Management

Across the Claim Detection, Claim Matching to DB, and Web Search modules, the following organizational controls will be treated:

- Regulatory Monitoring (by M36): The team will treat the failure of the ML application by actively monitoring EU regulations regarding model usage and examining alternative models should restrictions apply.

2. Hate Speech Detection (Module-Specific)

The Hate Speech module includes several unique treated actions aimed at fairness and security:

- Fairness and Bias Management (by M36): The threat of unbiased models will be treated through the use of validation datasets, with further checks applied as needed.
- Audit and Quality Control (by M36): Bias audits and fairness testing will be performed via validation dataset quality checks. Additionally, the threat of poisoning is addressed by ensuring reliable sources through quality checks of these validation datasets.
- Output Restriction (by M32): To prevent data disclosure, prompting will be used to restrict the model's output as much as possible.

3. Enhanced Monitoring and Performance

- Performance Indicators (Web Search & DB - periodically starting from M31): For the Web Search module, accuracy evaluations will re-run for specific steps like text snippet classification. For Claim Matching to DB, accuracy evaluations will re-run and the system will be adapted whenever the database is populated with new claims.
- Hate Speech API Monitoring (by M36): The team will implement active monitoring for API response times and classification success rates to ensure proper functioning.

4. Technical Data Controls

- Input Data Sanitization (by M36): For Claim Detection, Claim Matching to DB, and Hate Speech, the team will implement input data checks to ensure data is free-form text and excludes special tokens that could denote system/user message roles.

- Restricted Agency (by M36): In the Web Search module, the team will treat potential hijacking by ensuring workflows are deterministic and do not allow the LLM to make autonomous decisions on third-party tools.

3.6 Formal Tool Feedback

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

On April 6th, 2026, ATC's AI engineer responsible for developing the Claim Detection, Claim Matching to DB, and Claim Web Search functionalities (all Claim Analysis functionalities) accessed the AI model hub evaluation feedback form[10] and submitted the questionnaire provided.

The system was found to be useful for documenting all model training metadata in traditional ML/DL training and deployment pipelines. However, for workflows relying on pre-trained LLMs (e.g., AI workflows) the latter sections of the model hub do not apply and could be adapted. Overall the interface was found to be intuitive and easy to use, however requiring some overall technical knowledge, given that the metadata recorded in itself requires familiarity with ML/DL concepts.

On the same date, the ATC engineer in charge of the Hate Speech Detection tool completed the questionnaire for the AI Model Hub. While the hub is clearly more suited for traditional machine learning and deep learning pipelines, a few specific questions were relevant to an LLM-based workflow. These particular sections were helpful for documenting the tool's specific configuration and evaluation setup. However, many other fields did not apply to this use case. The interface was found to be intuitive, provided the user has the necessary technical background in ML/DL concepts.

3.6.2 TrustGuard

On April 6th, 2026, ATC's AI engineers responsible for developing the Claim Detection, Claim Matching to DB, and Claim Web Search functionalities (all Claim Analysis functionalities) and the engineer responsible for the Hate Speech Detection functionality accessed the TrustGuard evaluation feedback form[11] and submitted the questionnaire provided. In terms of overall ease of use, intuitiveness, and soundness of results, the tool was found to be satisfactory. Overall, while the system was found to be useful in the context of risk assessment during development, training, and deployment of traditional ML/DL systems, the various assessments seemed to lack

[10] https://southampton.qualtrics.com/jfe/form/SV_1HzOpXFhWeo7302

[11] https://southampton.qualtrics.com/jfe/form/SV_9XBxpgbbfRTljRc

questions that could reveal guidelines about how to produce trustworthy systems utilizing recent technologies such as AI workflows with LLMs, Agentic AI systems, etc. The following comments target mostly the FAITH Golden Table, from where TrustGuard retrieves the knowledge from, rather than TrustGuard itself.

3.6.2.1 Threat Assessment

Providing contextualised definitions of the identified threats would support users in more accurately assessing their likelihood within their specific domain. In particular, tailoring threat descriptions to the characteristics of each LSP would improve both clarity and applicability.

For instance, in the context of LSP1, the threat *"Evasion"* could be more effectively interpreted if accompanied by examples specific to LLM-based systems, such as prompt injection attacks, adversarial phrasing designed to bypass safeguards, or manipulation of inputs to avoid detection of harmful or misleading content. Such contextualisation would enable end-users to better understand how abstract threat categories manifest in practice, thereby improving the overall quality and consistency of the risk assessment.

Additionally, certain questions were identified as irrelevant within the context of the specific LSP. This issue arises because, beyond the generic questions derived from the literature, the FAITH Golden Table also incorporates questions contributed by all LSPs, which are then presented uniformly to TrustGuard users regardless of their domain. As a result, users may be required to respond to questions that are not applicable to their use case. For example, the Risk Assessor in the Media LSP was asked to evaluate the likelihood of the threat "Personal video data acquisition", which is primarily relevant to LSP5 and not to the media-focused application.

This observation highlights the need to differentiate and contextualise the question sets across LSPs, identifying both commonalities and domain-specific aspects. Accordingly, the FAITH Golden Table should be further refined to ensure that only relevant, context-aware questions are presented to users, thereby improving the efficiency, clarity, and accuracy of the assessment process.

3.6.2.2 Vulnerability Assessment

It was observed that a large number of vulnerability questions were irrelevant, as the tool is primarily geared toward the risk assessment of traditional ML/DL models. This observation emphasized further the need for the continuous processing and maintenance of the FAITH Golden Table. Similarly to the Threat Assessment, the presence of questions originating from other LSPs within the Vulnerability Assessment created confusion for end-users.

3.6.2.3 Risk Assessment & Mitigation Actions

Another important observation concerns the lack of contextualisation of the suggested mitigation actions. In several cases, mitigation measures are presented in a generic manner, without clearly linking them to the specific vulnerabilities they are intended to address.

For example, the mitigation action "Define and monitor indicators for proper functioning of the model", according to ENISA (2021), may be suggested to address vulnerabilities such as "excessive false positive alerts due to uncalibrated thresholds", "semantic drift in cluster definitions", or "undefined indicators of proper functioning, making compromise detection difficult". However, without an explicit mapping between the control and the specific vulnerability, as well as proper contextualisation of how the control should be implemented in each case, users may struggle to understand its relevance and practical application.

This lack of clarity can reduce the effectiveness of the risk mitigation process, as end-users may be uncertain about which actions to prioritise, how to implement them, and how they address the identified risks. Therefore, providing context-aware and vulnerability-specific guidance for mitigation actions is essential to support meaningful and actionable risk management.

At the same time, the user experience of TrustGuard could be further improved by avoiding the presentation of mitigation actions for vulnerabilities that users have marked as "not applicable (N/A)". Instead, these controls could be aggregated and presented separately at the end of the assessment. This would reduce unnecessary cognitive load and help users focus on relevant and actionable mitigation measures within their specific pilot context and obtain a more realistic understanding for the risks associated with their AI system.

3.6.3 TrustSense

As part of the formal feedback process, the questionnaire for the TrustSense tool was completed on April 27, 2026, by a member of the FH team. The evaluation focused on the tool's usability, clarity, and contribution to the risk assessment process. Overall, TrustSense was considered a useful tool that supports a more structured approach to assessing trustworthiness aspects, contributing to the overall quality and consistency of the risk assessment.

According to the feedback provided, the tool is easy to learn and generally accessible, with the interaction perceived as relatively clear and not overly complex. However, restrictions also draw attention to some restrictions. While the questionnaire provides a structured set of questions, some of them were not sufficiently clear, which affected the perceived efficiency of completing tasks compared to usual methods. Similarly, although the interface was considered simple, there is room for improvement in making it more intuitive and user-friendly. To enhance usability and broader adoption, it is recommended that the purpose and intended use of the tool be more

clearly stated at the beginning of the questionnaire, for example through a short introductory disclaimer. In addition, simplifying the language and reducing technical complexity would make the tool more accessible to users without a strong technical background.

4 AI System Evaluation

A data-centric evaluation of the system under test was conducted within Task T4.1 and is documented in Deliverable D4.1. The following section complements this analysis by presenting the user-centric evaluation of the system across the trustworthiness dimensions, as performed by the FH journalistic network.

4.1 Detailed user feedback

The user feedback gathered from offers a thorough understanding of the AI Intelligent Coaching application's perceived performance, dependability, and trustworthiness across several parameters. The average ratings for **reliability**, **security, data privacy**, **inclusivity (fairness)**, **clarity**, **accountability**, and **output understandability (interpretability)** range from **3.31** to **3.93**, indicating a **moderate-to-high** level of trust and satisfaction overall.

The findings presented in this section are based on a total of 107 responses from 75 participants, corresponding to the valid answers collected for each question and used for the calculation of percentages and average ratings. While the evaluation process provided valuable insights into user perceptions and experiences, certain aspects of the data collection process should be considered when interpreting the results. Participation was actively monitored throughout the pilot, although the possibility of repeated submissions cannot be entirely excluded. In addition, the feedback form was initially configured to collect responses anonymously; however, this setting was updated on February 10, 2026, following the inclusion of an open-ended “Observations and Feedback” field to facilitate more detailed qualitative input from participants. Consequently, the results should be interpreted primarily as indicative of user impressions and experiences rather than as statistically representative measurements. Nevertheless, the collected feedback offers meaningful evidence regarding the perceived usability, trustworthiness, and overall reception of the AI Intelligent Coaching application.

In general, participants assessed the system as accessible, and helpful for journalistic workflows, especially when it came to identifying potentially dangerous content, assessing claims, and spotting disinformation. The system's AI character was well known (average rating 3.93), and most users acknowledged that human judgment is still crucial, with journalists ultimately bearing responsibility (average rating 3.55). While overall performance and consistency were rated positively, several areas for improvement were identified. These include refining temporal reasoning, increasing the coverage of claim and hate speech detection, boosting language capabilities, and making data storage and privacy regulations clearer. In order to ensure that AI-generated outcomes assist informed decision-making without taking the place of human oversight, users also underlined the necessity for outputs to be easily understood and contextualized.

4.1.1 TRUSTWORTHINESS PROPERTY: SECURITY

Participants reported a **moderate level of trust** in the system's security, with an average rating of **3.45**. Most responses were neutral or positive: **31.8% rated 3**, **27.1% rated 4**, and **21.5% rated 5**. Lower trust levels were less common, with **14% rating 2** and **5.6% rating 1**. Overall, the results suggest a **balanced perception**, where many users show some confidence in the system's security, but a portion remain cautious about using it for **highly sensitive or confidential content**.

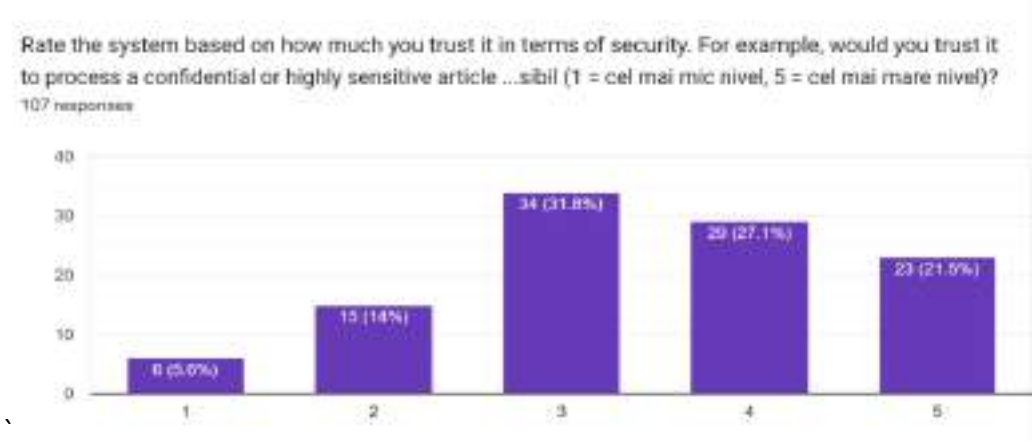


Figure 25 Aggregated user ratings of trust in the security of the AI Intelligent Coaching application

4.1.2 TRUSTWORTHINESS PROPERTY: VALIDITY & RELIABILITY (ACCURACY)

4.1.2.1 Claim Analysis component

Participants were asked to assess the Claim Analysis component's performance, particularly whether it accurately recognized all refuted assertions that were either fact-checked online or existing in the database.

The results indicate an **average rating of 3.43**, reflecting a **moderate-to-good perception of the component's effectiveness**. The distribution of responses shows that **25.2% of participants selected the neutral score of 3**, while **28% rated it at 4** and **22.4% at the highest level of 5**. Lower ratings were given by **18.7% of respondents for 2** and **5.6% for 1**.

Despite some discrepancies or overlooked claims, these results indicate that participants usually view the Claim Analysis component as helpful and fairly accurate.

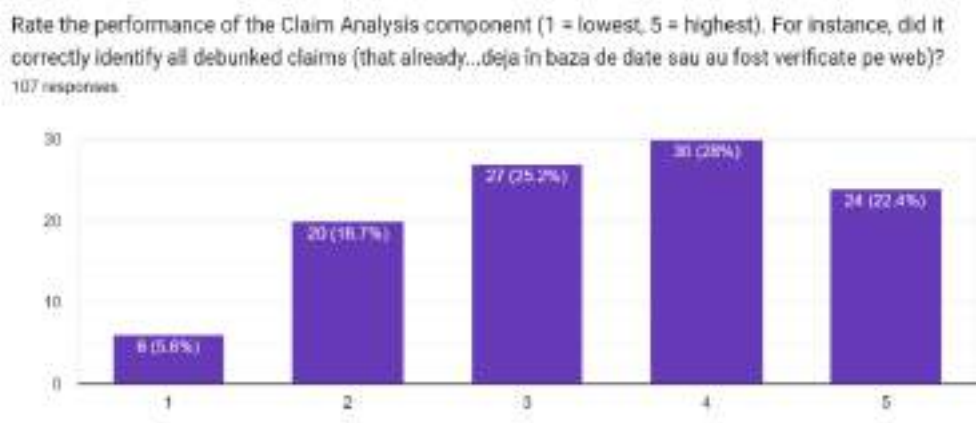


Figure 26 Aggregated user ratings of the AI system's Claim Analysis performance

4.1.2.2 Hate Speech Component

Participants were asked to assess how well the Hate Speech component performed, particularly if it properly recognized every incidence of hate speech in the given articles. A **reasonable degree of satisfaction** with the component's performance is indicated by the findings, which show an **average rating of 3.33**. Some words with subtle or context-dependent connotations were not consistently identified, despite the fact that many participants regarded the tool useful in screening for potentially harmful information.

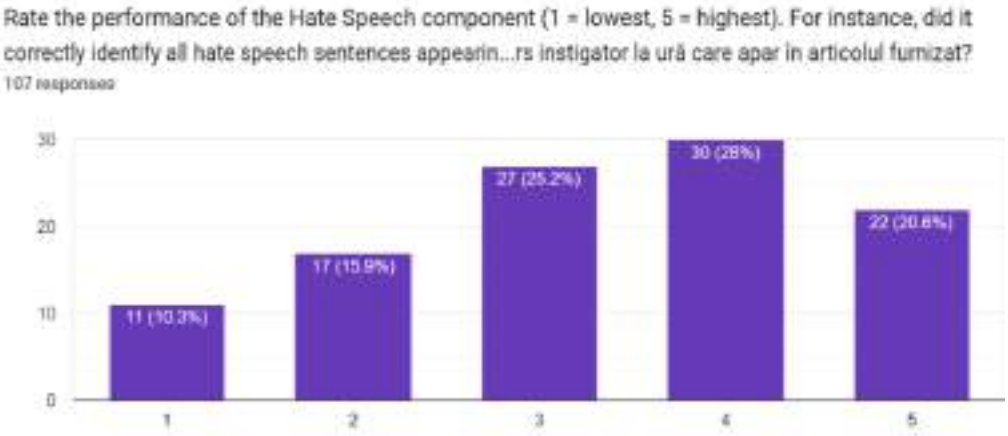


Figure 27. Aggregated user ratings of the AI system's Hate Speech Detection performance

4.1.3 TRUSTWORTHINESS PROPERTY: VALIDITY & RELIABILITY (ROBUSTNESS)

4.1.3.1 Claim Analysis Results

Participants were asked to assess the Claim Analysis component's consistency, specifically whether it produced the same outcomes when used on the same articles. A moderate to good perception of result consistency is suggested by the findings, which show an **average rating of 3.46**. When the same article was analyzed on multiple devices or under different situations, some

participants noticed differences in the findings. These findings emphasize how important it is to continue improving the component in order to guarantee consistent and dependable results, which is necessary for fostering user confidence and bolstering faith in the tool during regular journalistic workflows.

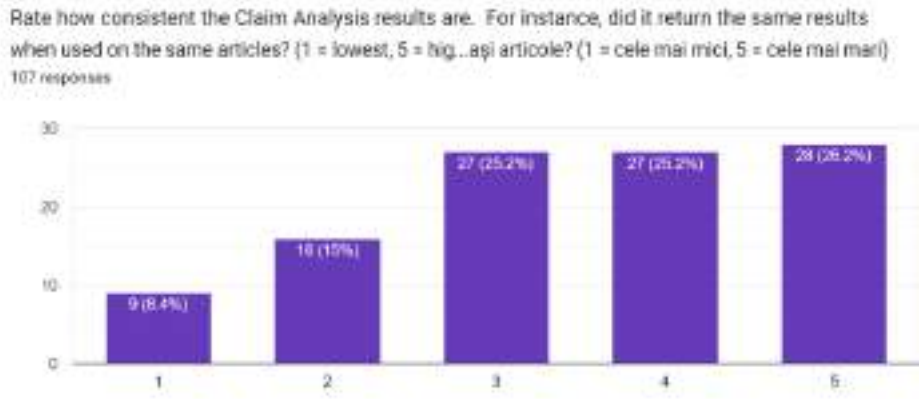


Figure 28 Aggregated user ratings from the feedback forms regarding the consistency of the AI system's Claim Analysis results

4.1.3.2 Hate Speech Results

A moderate perception of consistency is indicated by the results, which provide an **average rating of 3.43**. The outputs were generally considered stable by the participants; nevertheless, differences were noted when comparing similar articles or running tests again, indicating that the component's outcomes were not always completely repeatable.

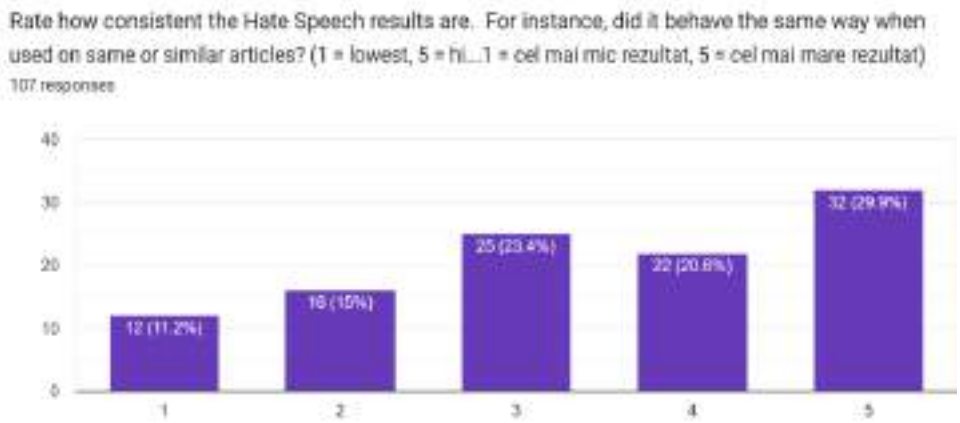


Figure 29. Aggregated user ratings from the feedback forms regarding the perceived consistency of the Hate Speech Detection component.

4.1.4 TRUSTWORTHINESS PROPERTY: PRIVACY

Participants were asked to score how much they trusted the AI Intelligent Coaching application to protect their privacy. The findings show an average rating of **3.52,** indicating a **moderate-to-high degree of confidence** in how the system handles sensitive and private data.

Overall, these findings imply that the majority of participants believe the system handles personal information in an ethical manner. To further increase trust in the application, it is crucial to communicate clearly about data protection procedures, storage guidelines, and confidentiality measures, as some users are still wary. For instance, **several participants specifically recommended clarifying the policy on storing uploaded texts to better protect confidentiality.**

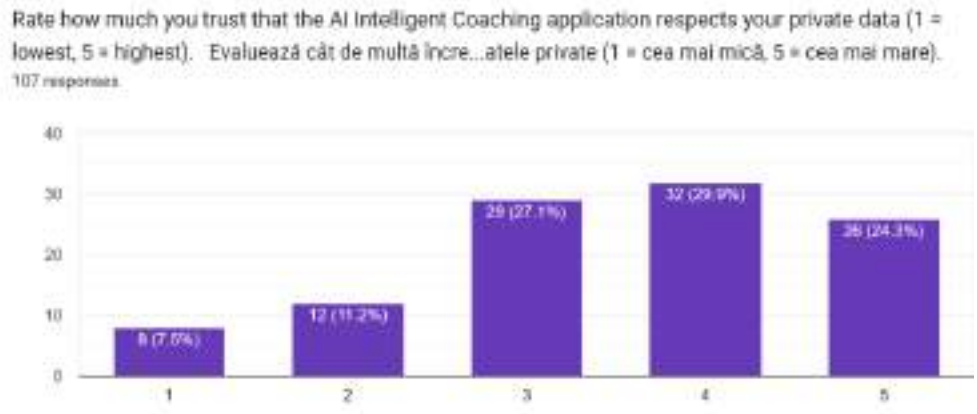


Figure 30. Aggregated user ratings from the feedback forms regarding their trust in the AI system's handling of data privacy.

4.1.5 TRUSTWORTHINESS PROPERTY: FAIRNESS WITH HARMFUL BIAS MANAGED

4.1.5.1 Claim Analysis Component

Participants were asked to evaluate the inclusiveness of the Claim Analysis component, specifically whether they noticed any signs of bias or stereotyping, such as racism, sexism, or homophobia.

A moderate view of inclusivity is shown by the results, which show an **average rating of 3.49.** The rating shows that some users were still wary of the possibility of subtle or context-dependent stereotyping in the AI's outputs, even though many participants did not notice significant biases.

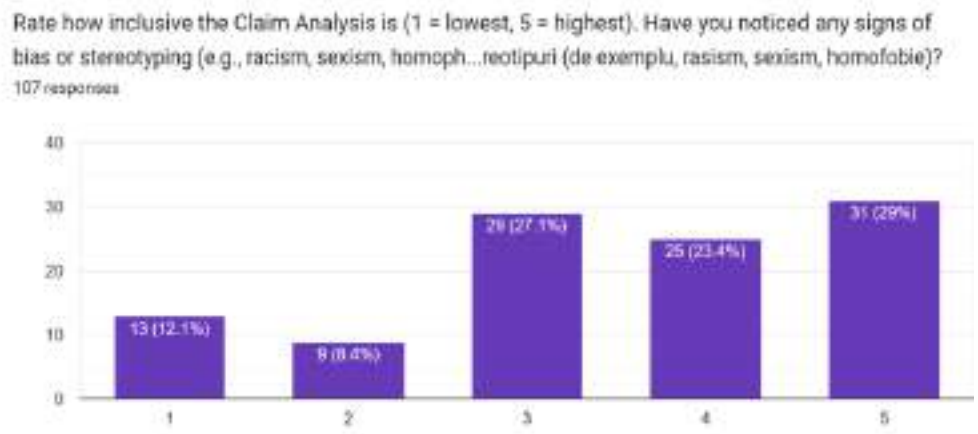


Figure 31. Aggregated user ratings on the inclusiveness of the Claim Analysis component.

4.1.5.2 Hate Speech Component

Participants were asked to evaluate the Hate Speech Detection component's inclusivity, particularly if they observed any instances of prejudice or stereotyping.

The findings show a moderate view of inclusivity, with an **average value of 3.31.** The score indicates that some users are still wary of any subtle or context-dependent stereotyping in the system's results, even though participants generally did not report widespread bias. These findings support the reliable and ethical use of the technology in journalistic workflows by highlighting the necessity of continuous monitoring and improvement to guarantee fairness and minimize any unintentional bias.

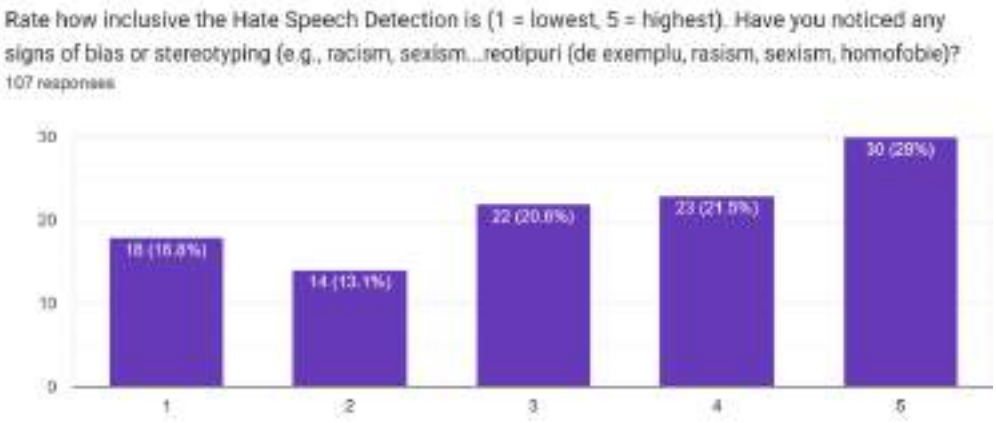


Figure 32. Aggregated user ratings on the inclusiveness of the Hate Speech Detection component.

4.1.6 TRUSTWORTHINESS PROPERTY: TRANSPARENCY

Participants were asked to rate how clearly the AI Intelligent Coaching application communicates that it is an AI system. The findings show an average rating of **3.93**, indicating a comparatively high degree of user awareness regarding the tool's artificial intelligence. This shows that most participants are aware that an AI system, not a human expert, is producing the outputs and suggestions. These results highlight the need to uphold transparency about AI involvement, which is essential for building user confidence, assisting with well-informed decision-making, and guaranteeing responsible use in journalistic workflows.



Figure 33. Aggregated user ratings on the clarity of the AI nature of the system.

4.1.7 TRUSTWORTHINESS PROPERTY: ACCOUNTABILITY

Participants were asked to assess how well the AI Intelligent Coaching application conveys who is in charge of the results it produces. A moderate-to-high impression of accountability clarity is reflected in the results, which show an average rating of **3.55.** The rating indicates that some users would benefit from more precise instructions on the limits of accountability, even though many participants were aware that human users, not the AI system, are ultimately responsible.

These results emphasize how crucial it is to clearly convey user **duties** and the **boundaries** of AI decision-making in order to build **trust**, guarantee **responsible use,** and support moral behaviour in journalistic workflows.



Figure 34. Aggregated user ratings regarding the clarity of the responsibility for the AI system's outputs

4.1.8 TRUSTWORTHINESS PROPERTY: INTERPRETABILITY

4.1.8.1 Claim Analysis Results

Participants were asked to score their level of trust in this component, based on the outputs' comprehensibility and clarity. The results show an average rating of **3.46**, indicating a moderate degree of confidence in the results' clarity. Although many participants thought the outcomes were generally accessible, several pointed out that some analyses would benefit from more context or explanations to help with understanding.



Figure 35. Aggregated user ratings on the intelligibility of the Claim Analysis results.

4.1.8.2 Hate Speech Results

Participants were asked to rate their level of trust in the Hate Speech Detection component based on the clarity and understandability of its outputs. The results show an average value of **3.36,**

indicating a **modest degree of confidence** in the results' interpretability. Although most participants thought the outcomes were quite clear, several pointed out that in order to assist appropriate interpretation, some analyses needed further **context** or explanations.

These findings emphasize how crucial it is to improve the openness and clarity of AI-generated outputs so that journalists may successfully comprehend and act upon the outcomes while maintaining their confidence in the system's dependability inside professional workflows.

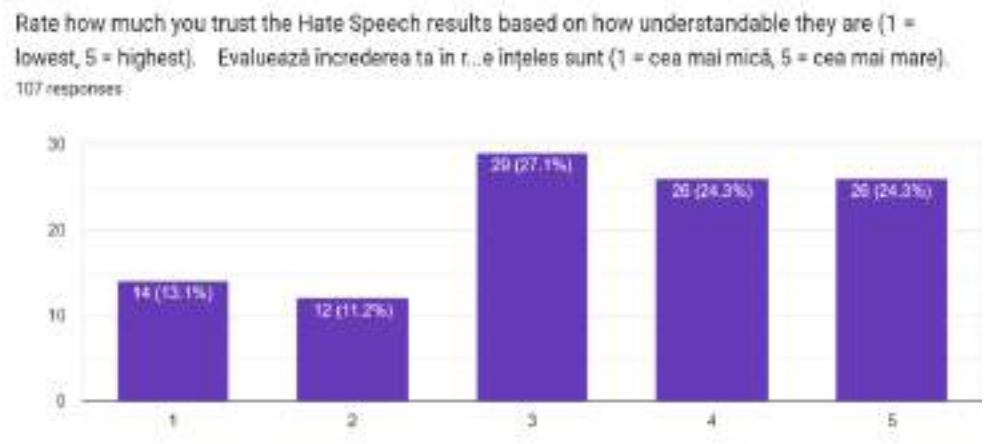


Figure 36. Aggregated user ratings on the intelligibility of the Hate Speech Detection.

4.2 Trustworthiness Assessment

In this trustworthiness assessment, we focus on the trustworthiness dimensions most crucial to the end-users, i.e., the ones mentioned in Section 3.1.3.

4.2.1 Validation & Reliability

Across all modules, Validation & Reliability is currently rated at High or Very High Risk, largely due to the "black-box" nature of the pre-trained models and a lack of localized performance monitoring. For each functionality Table 8 illustrates the risk level of this trustworthiness dimension, ranging from very high ("VH") to low ("L"), due to each associated threat. For instance, the validity and reliability of the Claim Detection model is in very high risk as the model is susceptible to the threat Poisoning.

Table 8. Risk levels of the trustworthiness dimension "Validation & Reliability" per associated threat, for each functionality of the system.

Threat	Claim Detection	Match to DB	Match to Other Articles	Hate Speech Detection
Propagation of misinformation	VH	VH	VH	VH
Compromise of ML application components	VH	VH	VH	L

Denial of service due to inconsistent data or a sponge example	VH	VH	VH	L
Failure or malfunction of ML application	VH	VH	VH	L
Poisoning	VH	VH	VH	H
Prompt attacks	VH	VH	VH	H
Lack of generalizability in production settings	VH	L	L	L
Use of adversarial examples crafted in white or grey box conditions	VH	VH	VH	L
Lack of effective fair data representations	H	L	L	VH
Model Drift Leading to Inaccurate Assessments	L	L	L	L
Changes in external conditions	L	L	L	VH

4.2.2 Fairness

Based on the assessment results from TrustGuard, the performance of the four core functionalities against the "Fair with harmful bias managed" dimension is from Low to Very High Risk due to the limitations of using third-party, pre-trained models. The various risk levels per threat and functionality are presented in Table 9.

Table 9. Risk levels of the trustworthiness dimension "Fairness" per associated threat, for each functionality of the system.

Threat	Claim Detection	Match to DB	Match to Other Articles	Hate Speech Detection
Failure or malfunction of ML application	VH	VH	VH	L
Poisoning	VH	VH	VH	VH
Regulatory and legal non-compliance	VH	VH	VH	L
Bias in Personalized Learning Recommendations	L	L	L	L
Changes in external conditions	L	L	L	VH
Insufficient safeguards against bias	L	VH	L	L
Lack of effective fair data representations	L	L	L	VH

4.2.3 Privacy

Based on the assessment documents for all four modules, the performance against the **Privacy** trustworthiness property is primarily influenced by the "black-box" nature of using third-party models and the lack of internal data access. The various risk levels per threat and functionality are presented in Table 10.

Table 10. Risk levels of the trustworthiness dimension "Privacy" per associated threat, for each functionality of the system.

Threat	Claim Detection	Match to DB	Match to Other Articles	Hate Speech Detection
Data disclosure	VH	L	VH	L
Model disclosure	VH	L	VH	L
Oracle	VH	L	VH	VH
Poisoning	VH	VH	VH	VH
Use of adversarial examples crafted in white or grey box conditions	VH	VH	VH	L
Model or data disclosure	H	L	H	VH
Need of external computation resources to infer segmentation	L	L	L	L

Important note: The first version of TrustGuard considers controls that have been denoted by the user in the Vulnerability Assessment with "not applicable" (NA), as not implemented controls. This has as a result to increase the level of the associated risks and to suggest to the user mitigation actions that are not applicable.

The second version of TrustGuard will consider the NA controls properly, and at the same time the FAITH Golden Table will have been updated to include controls more relevant to the specific use case. This is expected to change both the risk levels and the associated mitigation actions.

4.3 Preliminary Impact Evaluation

Within the Media Large Scale Pilot (LSP), the initial evaluation of the Intelligent Coaching application program offers insights into its usability, efficacy, and reliability in actual journalistic workflows. The tool is intended to help journalists find disinformation, spot logical fallacies, assess statements, and enable preliminary fact-checking. This assessment highlights the tool's present performance as well as opportunities for development based on more than 100 inputs from the feedback form from journalists and discussions held with them during webinars.

Positive impact observed

The application's user-friendly interface makes it very simple to integrate into daily routines, according to participants, who also highlighted that it is accessible and straightforward. It was commended for facilitating more precise and moral reporting, encouraging organized fact-checking, and expediting the comparison of information sources. Feedback from participants indicate that the tool assists in identifying possibly dangerous content and evaluating claims while protecting privacy without sharing any personal information. Users valued how the platform's AI features support journalistic work without taking the place of human judgment. Overall, journalists noted the tool's ability to lessen effort while offering organized analytical support and found it helpful, especially for early detection of misinformation.

Challenges and negative feedback observed

Participants noted a number of technical and analytical constraints despite its potential. Descriptive or interpretive passages were frequently ignored, and claim detection was restricted to statements with figures or statistics. Occasionally, hate speech recognition, particularly in longer texts, was unable to recognize statements with possibly harmful meanings, indicating that validity and reliability should be further improved. Similarly, users saw inconsistent outcomes across devices and time-sensitive claims assessed using out-of-date sources. Comprehensive fact-checking was hampered by the absence of a strong internal database and the restricted retrieval of public sources, and the inability to find saved articles indicated the need for better archive and search features. To guarantee secrecy and boost user confidence in the AI Intelligent Coaching application, it was also suggested that data storage regulations be clarified.

Potential future impacts

According to participant feedback, once present issues are resolved, the AI Intelligent Coaching application has significant potential to revolutionize journalistic workflows. More **thorough identification** of false information, logical contradictions, and potentially dangerous content, including subtle or context-dependent claims that are now overlooked, would be possible with the expansion of the claim and hate speech detection capabilities.

Participants pointed out that junior journalists, who are typically more open to AI technology, stand to gain the most from these advancements because they are actively involved in both academic research on media and AI and practical reporting. Such level of transparency raises the possibility that an improved the AI Intelligent Coaching application may become a common support tool in newsrooms, speeding up investigative research, improving editorial accuracy, and freeing up journalists' time for contextual analysis and interpretation. Long-term, such a tool might boost journalistic output, enhance ethical and fact-based reporting

Concluding assessment

The pilot shows that the AI Intelligent Coaching application has significant potential to enhance journalistic processes, especially in spotting misinformation and supporting organized analysis. Although its efficacy is currently limited by technological, analytical, and database-related issues, user feedback offers a clear path for improvement. In order to ensure credibility, dependability, and usefulness in professional newsrooms, human monitoring must continue in this process, while advancements in multilingual performance, temporal reasoning, claim detection, and data management will be crucial elements.

5 Conclusions and Roadmap for the Final Pilot Phase

5.1 Summary of achievements in the second period

Significant progress was made in planning and carrying out the first Media Large Scale Pilot (LSP) during this period. The AI Intelligent Coaching program was actively used by overall 75 journalists to evaluate its ability to detect hate speech and debunked claims. Technical improvements and human-AI engagement methods were driven by feedback responses, which highlighted the tool's potential to enhance organized fact-checking while emphasizing the ongoing significance of human oversight. These exercises gave a strong basis for the replication stage and offered practical insights for enhancing the use of AI in journalistic environments.

During this phase, FAITH AI Trustworthiness Assessment Framework (FAITH AI_TAF V1) was successfully instantiated, while pilot partners actively utilized and formally evaluated the FAITH toolset:

· **TrustSense** confirmed a high trustworthiness maturity level AI-users.

· **Data and AI Model Passports** for the AI Intelligent Coaching application functionalities were effectively added on the FAITH AI Model Hub, enabling full traceability and transparency across both data provenance and model architecture.

· **TrustGuard** was deployed to systematically identify and manage threats and risks across the trustworthiness properties. Overall, although AI Intelligent Coaching application is a system of low-criticality, its risk profile is assessed as high, mainly due to the ML-focused questions for the assessment. It was found that many of the control questions and mitigation actions from the FAITH Golden Table are not applicable for LLMs or cannot be implemented in AI Intelligent Coaching application. This stresses the need for a curated KB with more up-to-date threats, vulnerabilities and controls for Gen-AI systems, which is planned for the next period.

5.2 Plan for Replication Phase

The Replication Phase started in M20. The purpose of this pilot is to be expanded to involve journalists nationwide, going beyond local newsrooms and junior reporters. Operational testing of the AI system in more extensive newsroom workflows, ongoing gathering of qualitative and quantitative input, and incorporation of enhancements found during the initial pilot will be the main activities. In order to achieve successful and responsible adoption, the replication phase aims to confirm the tool's scalability, usability, and reliability while strengthening user training and guidance.

Additionally, during the Replication Pilot, the AI Intelligent Coaching application will undergo an in-context trustworthiness evaluation following the FAITH_AI_TAF_v2 framework and leveraging the corresponding FAITH AI_TAF_v2 tools. This phase constitutes a crucial progression from the already completed in-isolation assessment, enabling the evaluation of the system within real journalistic workflows and operational conditions. At the same time, it is expected to contribute significantly to the further refinement and maturation of the FAITH ecosystem tools.

The Replication Pilot will not only assess the AI Intelligent Coaching application under realistic deployment scenarios, but will also support the identification of context-specific threats, vulnerabilities, and mitigation actions emerging from the media domain. Furthermore, it is expected to provide valuable guidance for the media sector regarding the risks, challenges, and responsible deployment of LLM-based technologies in the context of combating disinformation.

6 References

Anderson, R. (2001, December). Why information security is hard-an economic perspective. In Seventeenth annual computer security applications conference (pp. 358-365). IEEE.

Ahmad, A., Vallès, Y., & Idaghdour, Y. (2025). Bias in AI systems: integrating formal and socio-technical approaches. Frontiers in Big Data, 8, 1686452.

Akram, M., Ahmad, N., Delshadi, A., Sultana, M., & Iqbal, M. W. Securing Large Language Models Against Prompt Injection and Data Leakage Attacks.

Barakat, B., & Jaf, S. (2025). Beyond Traditional Classifiers: Evaluating Large Language Models for Robust Hate Speech Detection. Computation, 13(8), 196.

Chhabra, A., Datta, S., Nahin, S. K., & Mohapatra, P. (2026). Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges. IEEE Access.

Cinà, A. E., Demontis, A., Biggio, B., Roli, F., & Pelillo, M. (2025). Energy-latency attacks via sponge poisoning. Information Sciences, 702, 121905.

Cui, Y., Pan, S., Liu, Y., Zhang, H., & Zuo, C. (2026). VortexPIA: Indirect prompt injection attack against LLMs for efficient extraction of user privacy. In Findings of the Association for Computational Linguistics: EACL 2026 (pp. 587–609).

ENISA, 2021. Securing Machine Learning Algorithms. Heraklion: European Union Agency for Cybersecurity

Geng, Tongcheng, et al. "Prompt injection attacks on large language models: A survey of attack methods, root causes, and defense strategies." Computers, Materials, & Continua 87.1 (2026).

Jawad, H., & Brunel, N. (2025). PSM: Prompt Sensitivity Minimization via LLM-Guided Black-Box Optimization. arXiv preprint arXiv:2511.16209.

Kim, J., Liu, X., Wang, Z., Qiu, S., Li, B., Guo, W., & Song, D. (2026). The Attack and Defense Landscape of Agentic AI: A Comprehensive Survey. arXiv preprint arXiv:2603.11088.

ŁAZARSKI, M. (2026). From Detection to Trust: Managing Cybersecurity Risks of Generative AI Systems in Organizations

Leite, J. A., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2026). LLM-Based Adversarial Persuasion Attacks on Fact-Checking Systems. arXiv preprint arXiv:2601.16890.

Li, T., & Zhu, Q. (2025). Agentic AI for Cyber Resilience: A New Security Paradigm and Its System-Theoretic Foundations. arXiv preprint arXiv:2512.22883.

Li, X., & Gao, X. (2025). Toward understanding security issues in the model context protocol ecosystem. arXiv preprint arXiv:2510.16558.

Liu, X., Li, P., Suh, E., Vorobeychik, Y., Mao, Z., Jha, S., ... & Xiao, C. (2024). Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. arXiv preprint arXiv:2410.05295.

Muminovic, A., & Muminovic, A. K. (2025, September). Large Language Models for Toxic Language Detection in Low-Resource Balkan Languages. In 2025 International Conference on Cybersecurity and AI-Based Systems (Cyber-AI) (pp. 14-21). IEEE.

OWASP Foundation. (2025). LLM08:2025 vector and embedding weaknesses. OWASP GenAI Security Project. https://genai.owasp.org/llmrisk/llm082025-vector-and-embedding-weaknesses/

OWASP Gen AI Security Project (2025) LLM02:2025 Sensitive Information Disclosure https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/

OWASP Gen AI Security Project (2025). LLM05:2025 Improper output handling. https://genai.owasp.org/llmrisk/llm052025-improper-output-handling/

Park, S., Nan, X. Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. AI & Soc 41, 1501–1515 (2026). https://doi.org/10.1007/s00146-025-02620-3

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., ... & Wrobel, S. (2023). Guideline for Designing Trustworthy Artificial Intelligence.

Qu, Y., et al. (2026). Supply-chain poisoning attacks against LLM coding agent skill ecosystems. arXiv preprint. https://arxiv.org/html/2604.03081v1

Sami, A. A. (2026). Code poisoning through misleading comments: Jailbreaking large language models via contextual deception.

Singh, A., & Singh, K. K. (2026). Ethics, Bias, and Responsible Generative AI. In Generative AI in R: Transforming Data Science with Synthetic Data and Advanced Modeling Techniques (pp. 425-473). Berkeley, CA: Apress.

Wang, Z., Pang, Y., & Lin, Y. (2024). Smart Expert System: Large Language Models as Text Classifiers. arXiv-2405.

Yu, G., Wang, Z., Huang, Y., Zhong, R., Zhong, Y., Wang, Y., & Lyu, M. R. (2026). Why Does the LLM Stop Computing: An Empirical Study of User-Reported Failures in Open-Source LLMs. arXiv preprint arXiv:2601.13655.

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In 34th USENIX Security Symposium (USENIX Security 25) (pp. 3827-3844).Vane, S., et al. (2026, January 26). Beyond data privacy: New privacy risks for large language models. arXiv preprint.