



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437



AI services for COntinuous trust
in emerging Digital Environments

D3.1 – User-centred requirements definition

Grant Agreement no.:	101135437
Project Title:	AI-CODE - AI services for COntinuous trust in emerging Digital Environments
Contractual Due date:	30.11.2024
Actual submission date:	29.11.2024
Revised version:	04.06.2025
Lead beneficiary:	EIPCM

Deliverable number	D3.1
Deliverable Title:	User-centred requirements definition
Description:	Requirements definition for the defined user groups and use case scenarios based on results of the analysis in T3.1, T3.2 and T3.4.
Type/Nature	☒ R - Document, report ☐ DEM - <i>Demonstrator, pilot, prototype</i> ☐ DEC - <i>Websites, patent filings, videos, etc.</i> ☐ OTHER ☐ ETHICS – <i>Ethics requirement</i> ☐ ORDP – <i>Open Research Data Pilot</i>
Dissemination Level	☒ Public – fully open ☐ Sensitive — limited under the conditions of the Grant Agreement ☐ EU classified — RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level.
Work Package number	3
Work Package leader	EIPCM
Author(s) <i>in alphabetical order</i>	Barrio, Santiago Barbero-Álvarez, Miguel Antonio Becker, Mathias Cecchine, Randi Gallotti, Riccardo Guerini, Marco Gürtler, Paula Kanellos, Ilias La Cava, Lucio Lazaridis, Iosif Leoni, Riccardo Maljur, Tina Menini, Stefano Moreno, Francisco Novak, Jasminko Schinas, Manos Spinozzi, Giulio Tagarelli, Andrea Tang, Annie Tsalapati, Eleni
Internal Reviewers <i>in alphabetical order</i>	Nucci, Francesco Wilmont, Ilona

Authors

Name	Partner	e-mail
Mathias Becker	EIPCM	m.becker@eipcm.org
Annie Tang	EIPCM	a.tang@eipcm.org
Tina Maljur	EIPCM	t.maljur@eipcm.org
Jasminko Novak	EIPCM	j.novak@eipcm.org
Randi Cecchine	RU	randi.cecchine@ru.nl
Paula Gürtler	CEPS	paula.gurtler@ceps.eu
Giulio Spinozzi	DST	g.spinozzi@dstech.it
Riccardo Leoni	DST	riccardo.leoni@dstech.it
Marco Guerini	FBK	m.guerini@fbk.eu
Riccardo Gallotti	FBK	rgallotti@fbk.eu
Stefano Menini	FBK	menini@fbk.eu
Santiago Barrio	UPM	s.barrio@upm.es
Miguel Antonio Barbero-Álvarez	UPM	m.balvarez@upm.es
Francisco Moreno	UPM	fran.moreno@upm.es
Eleni Tsalapati	ATC	e.tsalapati@atc.gr
Ilias Kanellos	ATC	i.kanellos@atc.gr
Iosif Lazaridis	ATC	i.lazaridis@atc.gr
Manos Schinas	CERTH	manosetro@iti.gr
Panagiotis Galopoulos	CERTH	gpan@iti.gr
Giorgos Nikoulis	CERTH	gnikoul@iti.gr
Lucio La Cava	UNICAL	lucio.lacava@dimes.unical.it
Andrea Tagarelli	UNICAL	andrea.tagarelli@unical.it

History of changes

Version	Date	Changes
0.1	23.9.2024	First draft
0.2	15.10.2024	Revised draft
0.3	31.10.2024	Completed integration of chapters 5-10
0.4	07.11.2024	Added chapters 1, 11
0.5	08.11.2024	Revised draft
0.6	22.11.2024	Revised draft based on all partners feedback
1.0	25.11.2024	First final version submitted to Quality Control
1.1	04.06.2025	Revision addressing expert recommendations

		from the 1st project review. Text added or changed are highlighted in green: - Extended Executive Summary and section 4.1 (Outline of the Approach) highlighting number of media professionals involved - Additional target groups identified for each service (section 4.2) - User feedback analysis describing the main findings from the co-creation sessions (section 4.5.2) - Added/extended for each service: - o How they support critical thinking (sections 5.1.2, 6.1.2, 7.1.2, 8.1.2, 9.1.2 & 10.1.2) - o Additional target groups and their expected benefits (sections 5.2, 6.2, 7.2, 8.2, 9.2 & 10.2) - o How the early functional/non-functional/ethical requirements will be addressed in the service design. An additional early non-functional requirement has been added for all services to foresee differentiated support for different user cohorts within their target user group (sections 5.5/5.6, 6.5/6.6, 7.5/7.6, 8.5/8.6, 9.5/9.6 and 10.5/10.6)
2.0	04.06.2025	Final revised version

Table of contents

1	Executive Summary	8
Part I: Overview and Process		10
2	Introduction	10
2.1	Role of Requirements Analysis in the Project	10
2.2	Outline of this Deliverable	10
3	Overview of AI Code Services	11
3.1	AI-CODE Services in a Nutshell	11
3.1.1	Disinformation Detection in the Next-Generation Social Media	12
3.1.2	Trustability and Credibility Assessment	12
3.1.3	Media Asset Annotation and Management (MAAM)	13
3.1.4	Generative AI Interactive Coaching Service & Dynamic Simulator	13
3.1.5	Transparency Service for AI Model Cards	14
3.1.6	Personal Companion for Understanding Disinformation	14
3.2	AI-CODE Objectives covered by the Services	14
4	User-Centred Requirements and Design Process	16
4.1	Outline of the Approach	16
4.2	Initial Definition of Target User Groups	17
4.3	Initial Analysis of User Needs	19
4.3.1	Exploratory User Interviews	19
4.3.2	Summary of main findings from related user-centred literature	24
4.3.3	Summary of ethical frameworks that informed the ethical requirements for the services	31
4.4	Development of Early Mock-ups and Prototypes	33
4.5	Co-Creation Workshops	34
4.5.1	Workshop methodology and approach	34
4.5.2	Summary of main findings from end-user feedback	36
4.6	Definition of scenarios of use and user stories	39
Part II: Results per service		40
5	Disinformation Detection in the Next-Generation Social Media	40
5.1	Service Description	40
5.1.1	Scope and Objectives	40
5.1.2	Supporting critical thinking	41
5.2	Target Users and Expected Benefits	41
5.2.1	Fediverse Network Moderation Tool	41
5.2.2	Mastodon Instances Inspector	41
5.3	Scenarios of Use	42
5.3.1	Fediverse Network Moderation Tool	42

5.3.2	Mastodon Instances Inspector	44
5.4	User Stories.....	45
5.4.1	Fediverse Network Moderation Tool	45
5.4.2	Mastodon Instances Inspector	46
5.5	Early Requirements	47
5.6	Ethical Requirements.....	49
5.7	Early Success Criteria.....	49
5.8	Early Mock-ups of the Service	50
6	Trustability and credibility assessment	51
6.1	Service Description.....	51
6.1.1	Scope and objectives	51
6.1.2	Supporting critical thinking.....	51
6.2	Target Users and Expected Benefits.....	51
6.3	Scenarios of Use	52
6.4	User Stories.....	52
6.5	Early Requirements	53
6.6	Ethical Requirements.....	55
6.7	Early Success Criteria.....	55
6.8	Early Mock-ups of the Service	56
7	Media Asset Annotation and Management	59
7.1	Service Description.....	59
7.1.1	Scope and objectives	59
7.1.2	Supporting critical thinking.....	59
7.2	Target Users and Expected Benefits.....	59
7.3	Scenarios of Use	61
7.4	User Stories.....	63
7.5	Early Requirements	64
7.6	Ethical Requirements.....	67
7.7	Early Success Criteria.....	68
7.8	Early Mock-ups of the Service	69
8	Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals	71
8.1	Service Description.....	71
8.1.1	Scope and objectives	71
8.1.2	Supporting critical thinking.....	71
8.2	Target Users and Expected Benefits.....	72
8.2.1	Interactive Coaching and Simulation Module.....	72
8.2.2	Counter-narrative Recommendation Module	73
8.3	Scenarios of Use	73

8.3.1	Interactive Coaching and Simulation Module	73
8.3.2	Counter-narrative Recommendation Module	76
8.4	User Stories.....	77
8.4.1	Interactive Coaching and Simulation Module	77
8.4.2	Counter-narrative Recommendation Module	79
8.5	Early Requirements	80
8.6	Ethical Requirements.....	82
8.7	Early Success Criteria.....	83
8.8	Early Mock-ups of the Service	83
8.8.1	Interactive Coaching and Simulation Module	83
8.8.2	Counter-narrative Recommendation Module	90
9	Transparency services for AI-model cards.....	91
9.1	Service Description.....	91
9.1.1	Scope and objectives	91
9.1.2	Supporting critical thinking.....	91
9.2	Target Users and Expected Benefits.....	91
9.3	Scenarios of Use	93
9.4	User Stories.....	95
9.5	Early Requirements	96
9.6	Ethical Requirements.....	98
9.7	Early Success Criteria.....	99
9.8	Early Mock-ups of the Service	99
10	Personal Companion for Understanding Disinformation for Media Professionals.....	102
10.1	Service Description.....	102
10.1.1	Scope and objectives	102
10.1.2	Supporting critical thinking	102
10.2	Target Users and Expected Benefits.....	102
10.3	Scenarios of Use	103
10.4	User Stories.....	104
10.5	Early Requirements	107
10.6	Ethical Requirements.....	109
10.7	Early Success Criteria.....	110
10.8	Early Mock-ups of the Service	110
11	Conclusions and Next Steps.....	117
12	References.....	118
13	Appendix.....	120

1 Executive Summary

The main goal of the interdisciplinary AI-CODE project is to develop a novel ecosystem of six different services that will proactively support media professionals in identifying and countering disinformation and support the production of trustworthy content. These services include:

1. Disinformation Detection in the Next-Generation Social Media
2. Trustability and Credibility Assessment
3. Media Asset Annotation and Management (MAAM)
4. Generative AI Interactive Coaching Service & Dynamic Simulator
5. Transparency Service for AI Model Cards
6. Personal Companion for Understanding Disinformation

These services will assist journalists, fact checkers and other media professionals how to work in emerging digital environments and how to utilise generative AI effectively and credibly, to detect new forms of content manipulation, as well as to assess the reputation and credibility of sources and their content.

This deliverable describes the process and early outcomes of the user-centred requirements analysis for these AI-CODE services based on the work performed in tasks T3.1, T3.2 and T3.4. and informed by the literature review carried out in WP2 (see D2.1 and D2.2). This includes initial scenarios of use, user stories, and early functional, non-functional, and ethical requirements as well as early success criteria for each AI-CODE service.

As the initial phase of the iterative user-centred design approach, the first part of the requirements analysis process involved defining target user groups for each service, conducting exploratory interviews with 18 media professionals, reviewing relevant literature on user-centred requirements for media professionals and fact-checkers, and analyzing attitudes towards AI use in media and journalism. The review included a survey of AI in media organisations with 105 news and media organisations from 46 different countries (Beckett at al., 2023), as well as a study of AI in news organisations (134 interview participants from 35 news organizations and 36 international experts from industry, academia, technology, and policy) (Simon, 2024).

Subsequently, based on obtained insights first user stories and user-centered requirements were developed and mock-ups and prototypes were created for each service. This was followed by co-creation workshops with end-users to obtain user feedback and potential suggestions for revising and further developing the user stories, requirements and mock-ups. These workshops involved 121 participants including 59 external to AI-CODE (from 12 different professional roles in the media field). Accordingly, the initial scenarios of use, user stories, early functional, non-functional, and ethical requirements as well as early success criteria for each AI-CODE service were completed and are presented in this deliverable. Ethical considerations were informed by the Ethics Guidelines for Trustworthy AI, published by the European High Level Expert Group on AI, and tailored to the specific context of the AI-CODE project.

The results presented in this deliverable set the foundation for a more detailed specification to be provided in the upcoming deliverable D3.2 "Functional specifications and technical requirements" (M16). This deliverable thereby serves as a bridge between the user-centred requirements gathering process and the technical development phase, ensuring that the final AI-CODE solutions are both technologically advanced and closely aligned with the needs of media professionals and their professional context. The user-centred approach ensures that the AI-CODE services will address the real needs of media professionals while incorporating state-of-the-art technological innovations for fighting disinformation and supporting trustworthy use of Generative AI.

In order to address the expert recommendations from the 1st project review, the following additions to the original version of this deliverable have been made (highlighted in green for easier traceability):

- Additional target groups identified for each service (section 4.2)
- User feedback analysis describing the main findings from the co-creation sessions (section 4.5.2)
- Added/extended for each service:
 - o Extended Executive Summary and section 4.1 (Outline of the Approach) highlighting number of

D3.1 User-centred requirements definition

- media professionals involved
- How they support critical thinking (sections 5.1.2, 6.1.2, 7.1.2, 8.1.2, 9.1.2 & 10.1.2)
- Additional target groups and their expected benefits (sections 5.2, 6.2, 7.2, 8.2, 9.2 & 10.2)
- How the early functional/non-functional/ethical requirements will be addressed in the service design. An additional early non-functional requirement has been added for all services to foresee differentiated support for different user cohorts within their target user group (sections 5.5/5.6, 6.5/6.6, 7.5/7.6, 8.5/8.6, 9.5/9.6 and 10.5/10.6)

Part I: Overview and Process

2 Introduction

2.1 Role of Requirements Analysis in the Project

The setting of the project with six different services stresses the importance of a user-centred requirements process that is able to elicit the needs of the users and the context in which the AI-CODE services are to be used. Therefore, the overall requirements analysis is planned as an iterative user-centred design process (Abrás et al., 2004).

The requirements process brings together end-user needs and the technical ambitions of the AI-CODE project. The requirements process defines how novel, easy-to-use services shall be designed and developed in order to take advantage of state-of-the-art technological innovations, while addressing the needs of the users and the intended context of use within the scope of the AI-CODE project.

The initial phase of the requirements elicitation process develops the initial scenarios of use, user stories, use cases as well as the early high-level functional, non-functional and ethical requirements for each AI-CODE service (all of which are being presented in this deliverable). They will be further updated as part of the next planned round of co-creation workshop, in which further end-user feedback from the main target groups shall be obtained to inform the further in-depth technical requirements analysis. In accordance with the workplan, in the deliverable *D3.2 Functional specifications and technical requirements* a more detailed specification of the planned functionalities and technical requirements will be provided.

2.2 Outline of this Deliverable

The deliverable is structured as follows. Chapter 3 provides a high-level overview of the AI-CODE services, their relation to the overarching project goals and the interconnections between them. Chapter 4 details the user-centred requirements analysis and design process that has been adopted in the project leading to this deliverable. This entails the initial definition of the target user groups along with the analysis of user needs, the development of early mock-ups and prototypes for eliciting user feedback and a series of co-creation sessions with end-users - all of which resulted in the definition of scenarios of use, user stories and the user-centred requirements.

The process described in Chapter 4 has been significantly informed by the results from WP2 literature analysis and additional review of existing literature focusing on user-centred requirements of media professionals and fact-checkers. Furthermore, the analysis of overarching journalistic ethical values, standards, and guidelines for trustworthy journalism has been carefully incorporated into the requirements gathering process. This integration ensures that the AI-CODE services not only meet the practical needs of users but also align with the ethical principles that underpin quality journalism. The consideration of ethical frameworks has played a crucial role in informing the ethical requirements for the AI-CODE services. This approach ensures that the development of these tools is guided by robust ethical considerations, addressing potential issues related to AI use in journalism and fact-checking.

Part II of the deliverable then describes the resulting user-centred requirements for each AI-CODE service including the target users, scenarios of use and user stories along with the early functional, non-functional and ethical requirements, as well as early success criteria. Screenshots of the mock-ups and prototypes used in and informed by the feedback from the end-user co-creation workshops are also provided for each service.

3 Overview of AI Code Services

3.1 AI-CODE Services in a Nutshell

As context for the rest of the deliverable describing the user-centered requirements analysis and its results, this section gives a short (“in-a-nutshell”) description of the AI-CODE services listed in Table 1. All of them aim to develop innovative ways of supporting media professionals in trusted information production. This includes in particular novel AI-based services aimed at helping media professionals (such as journalists and fact-checkers) to combat the spread and impact of disinformation by making it easier to detect and mitigate the spread of disinformation, verify credibility of content and sources and critically evaluate the outputs and the use of generative AI (LLMs) to reinforce trust in content from reputable media sources.

Table 1: Overview of AI-CODE services

Service Name	Short Description
1. Disinformation Detection in the Next-Generation Social Media	Identifies disinformation on Fediverse social media platforms.
2. Trustability and Credibility Assessment	Evaluates the trustability and credibility of content and sources found in posts io Fediverse social media platforms
3. Media Asset Annotation and Management (MAAM)	Automatically annotates media content to facilitate organization and retrieval and detects synthetic and manipulated content, including deepfakes
4. Generative AI Interactive Coaching Service & Dynamic Simulator	Helps media professionals better understand how to appropriately use and critically evaluate outputs from LLMs and associated risks, including AI-supported verification of claims and counter-narratives for fighting disinformation.
5. Transparency Service for AI Model Cards	Reports the capabilities, limits, potential risks, and repercussions of different AI models.
6. Personal Companion for Understanding Disinformation	Helps to detect disinformation, hate speech and logical fallacies in articles.

Figure 1 provides an overview of the interconnections and relationships between the various AI-CODE services indicating how each service complements and enhances the others, creating a robust ecosystem of tools for media professionals, journalists, and fact-checkers. Further details about these relationships are given in the following subsections containing a brief description of each service. Understanding the focus of each service and its connections with other services, facilitates understanding the integrated nature of the AI-CODE project and how its services work together to help combat disinformation and improve the quality of information in trusted media production.

D3.1 User-centred requirements definition

	Transparency Service for AI Model Cards	Trustability and Credibility Assessment	Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals	Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals	Media Asset Annotation and Management	Personal Companion for Understanding Disinformation for Media Professionals
Transparency Service for AI Model Cards			X		X	X
Trustability and Credibility Assessment	X			X		
Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals		X				X
Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals		X				
Media Asset Annotation and Management	X			X		
Personal Companion for Understanding Disinformation for Media Professionals	X		X	X		

Figure 1: Overview of connections between AI-CODE services

3.1.1 Disinformation Detection in the Next-Generation Social Media

The Disinformation Detection in the Next-Generation Social Media service consists of two key modules. The **Fediverse Network Moderation Tool** identifies potentially harmful servers in the Fediverse network by categorizing banned servers based on specific keywords, providing an updated list through a user-friendly interface. The **Mastodon Credibility Service**, designed for journalists and fact-checkers, analyzes information credibility on Mastodon instances using data from message reshares, instance bans, and content analysis, offering tools for instance inspection and visual reports on information dissemination trends.

The Disinformation Detection in the Next-Generation Social Media has connections with the **Trustability and Credibility Service** as it analyzes the network or its given nodes (social media sources / accounts) focusing on the specifics of the federated networks. The Disinformation Detection Service could potentially also use some of the same modules or tools as the Trustability and Credibility Assessment (depending on exact technical specification to be performed on D3.1), but applied in a different manner, i.e., not on a single piece of text, but at an aggregated level.

3.1.2 Trustability and Credibility Assessment

The Trustability and Credibility Assessment service utilizes AI-based methods to return credibility and trustability scores from content found in Fediverse social media platforms, providing explainable assessments for both existing and custom content, with the possibility of covering both individual and institutional creators.

The Trustability and Credibility Assessment has connections with the **Media Asset Annotation and Management MAAM**: While the Disinformation Detection in the Next-Generation Social Media analyzes instances in platforms like Mastodon and BlueSky, the service within Trustability and Credibility Assessment

D3.1 User-centred requirements definition

analyzes posts. Differences among the specific nature of its input and its internal processing (aggregation of posts vs. single instances) can be found. This service has the possibilities to showcase the utilities of other services with a compatible functioning or goal. The Media Asset Annotation and Management service will provide additional information that can be interpreted in the direction of credibility and trustability, enhancing the whole experience of usage of the dashboard. In a similar sense the **Personal Companion for Understanding Disinformation** and the **Disinformation Detection in the Next-Generation Social Media** can also add value to the Trustability and Credibility Assessment.

3.1.3 *Media Asset Annotation and Management (MAAM)*

The Media Asset Annotation and Management (MAAM) service is an advanced system that uses AI models for content annotation (labeling), retrieval, and integrity analysis. It's being enhanced with improved forgery detection, explainability features, and user profiling for disinformation analysis. The service is tailored to serve professionals such as journalists and researchers, providing robust tools for managing and verifying media assets. These enhancements aim to bolster the platform's effectiveness in combating media manipulation and disinformation. MAAM has connections to the **Transparency Service for AI-model Cards** as it will use a transparency service to help end-users understand the results of annotation models, including AI-generated content detectors (for synthetic content and deepfakes). MAAM also relates to **Disinformation Detection in the Next-generation Social Media** as Multimedia content collected from next-generation social media can be imported into MAAM to detect cases of AI-generated or manipulated content, such as synthetic images, image forgeries, deepfakes, and more.

3.1.4 *Generative AI Interactive Coaching Service & Dynamic Simulator*

The Generative AI Interactive Coaching Service & Dynamic Simulator consists of two modules. The **Interactive Coaching and Simulation Module (ICSM)** provides an interactive simulation and coaching tool that helps media professionals to better understand the main working principles of generative AI (i.e. Large Language Models) and associated risks, as well as to understand main modalities and principles of LLM use (e.g. prompt construction, system interaction) needed for the production of trustworthy content. This includes interactive guidance for reliability analysis of LLM outputs and support for verifying the veracity of LLM outputs. Users can apply and deepen these learnings in the Counter-narrative Recommendation Module as a concrete LLM application for analysing and countering disinformation. The **Counter-narrative Recommendation Module (CNRM)** offers on-demand suggestions for responding to disinformation in simulated online conversations, utilizing an archive of debunk reports and LLMs for pro-social purposes. It helps users develop effective counter-disinformation strategies and supports claim verification workflows, enabling media professionals to produce timely and accurate responses to false information. Both modules have several interconnecting areas themselves. The main principle is that users can apply their understanding and learnings from the ICSM in the CNRM as a practical LLM application. This includes applying the learnings from prompt construction and the reliability analysis guidance from ICSM, to prompt customization and to the analysis of the generated claim veracity explanations in the CNRM. The learnings from the multiple documents query in the RAG scenario in the ICSM prepare users for using multiple documents as sources for the claim verification in the CNRM.

The ICSM also helps users understand how to identify and address inaccuracies in the CNRM's truthfulness explanations (e.g. modifying prompts by adding specific focus points or claims to correct output errors). They also discover how these methods may not always be effective, due to inherent reliability problems of LLMs. This addresses a fundamental question in system design that is often overlooked in LLM applications: how to support users in identifying, correcting, and recovering from errors in system output.

The Generative AI Interactive Coaching Service & Dynamic Simulator will also integrate the functionality for identification of logical fallacies and claim veracity from the **Personal Companion for Understanding Disinformation for Media Professionals** in the output reliability analysis and link to the **Transparency Service for AI Model Cards**. The CNRM will provide a link to the **Trustability and Credibility Assessment** service.

3.1.5 Transparency Service for AI Model Cards

The Transparency Service for AI Model Cards develops a standardized structure for creating and managing AI model cards, including performance metrics, model limitations and ethical considerations, offering a user interface and SDK for card creation and management. It also includes management functionalities like version control and updates, enabling developers to maintain a history of model improvements. Additionally, the service provides varying levels of detail within model cards to cater to both technical experts and non-technical stakeholders. The Transparency Service for AI Model Cards has connections with the **Media Asset Annotation and Management** as the transparency service will be used to document the detection models integrated into MAAM, helping non-expert users understand the results, explore model limitations, and address ethical issues related to their usage. The Transparency Service also connects with the **Personal Companion for Understanding Disinformation for Media Professionals** as it will document the AI models used in this tool, helping media professionals understand the results, explore model limitations, and address ethical concerns. It will clarify how the model identifies debunked claims, matches claims across various web sources, detects logical fallacies, and identifies hate speech. By offering these insights, the service ensures users can responsibly interpret detection outcomes and trust the tool's recommendations. Finally, the Transparency Service for AI Model Cards also links to the **Trustability and Credibility Assessment** as it will document the models behind trustability dashboard, helping media professionals understand how trustability assessments are made for information sources on the internet and next-generation platforms, like the Fediverse.

3.1.6 Personal Companion for Understanding Disinformation

The Personal Companion for Understanding Disinformation is an LLM-based service assisting journalists and fact-checkers in combating disinformation, helping in writing credible articles, checking content validity. It offers a range of features including assistance in writing credible articles, fact-checking, claim validation, and understanding disinformation spread, while also helping fact-checkers manage their debunking databases. The Personal Companion for Understanding Disinformation has connections to the **Generative AI Interactive Coaching Service & Dynamic Simulator** as the Counter-narrative Recommendation Module could be used from the end user to debunk articles and then add the <claim, verdict, report> to the debunking database of the Personal companion service. In the Interactive Coaching and Simulation Module the end-user can select a set of resources (website articles that match with a claim, debunking reports) and request for an AI-generated report based on these resources. The interactive coaching service could guide the user in this process, e.g., suggest some prompts. The Personal Companion also connects with the **Disinformation Detection in the Next-generation Social Media** as debunked claims found in social media could be fed to the debunking DB of the Personal Companion and therefore processes articles to be checked against these debunked social media claims as well.

3.2 AI-CODE Objectives covered by the Services

How each of the AI-CODE services contribute to the specific objectives of the AI-CODE project is outlined in Table 2. The AI-CODE objectives are:

1. Assess the future developments related to the impact of generative AI and the next-generation social media on the domain of online disinformation and trust
2. Provide a set of content-driven services for disinformation detection and trust assessment in next-generation social media
3. Provide a set of user-driven services dedicated to media professionals to support trusted information production using the emerging technologies
4. Co-design, demonstrate and evaluate innovative approaches and technologies with end users
5. Generate business and social impact in the information ecosystem

AI-CODE objective #1 (Assess the future developments related to the impact of generative AI and the next-generation social media on the domain of online disinformation and trust) is addressed in Workpackage 2 and thus not included in Table 2. All services together form the foundation for achieving AI-CODE objective #5 (Generate business and social impact in the information ecosystem) by providing media professionals,

D3.1 User-centred requirements definition

journalists, and fact-checkers with advanced tools and capabilities that enhance their ability to detect, analyze, and counter disinformation, ultimately improving the quality and trustworthiness of information in the digital landscape. Therefore objective #5 is also addressed by all services in Table 2.

Table 2: Overview of how the AI-CODE services relate to the AI-CODE objectives

Service	AI-CODE Objective				Relation to AI-CODE objectives
	2	3	4	5	
Media Asset Annotation and Management	x		x	x	MAAM is a tool designed to organize media content through AI-based annotation models and detect media forgeries and manipulations through state-of-the-art detectors. This service aids in the detection of misinformation by identifying altered or manipulated media assets. MAAM directly contributes to Objective 2 by offering content-driven services that support disinformation detection in next-generation platforms. It also aligns with Objective 4 through its co-design and demonstration with end users, evaluating the tool's effectiveness in real-world scenarios for detecting misinformation.
Personal Companion for Understanding Disinformation for Media Professionals		x		x	This service enables media professionals to check their articles or external articles with regard to i) claims that have been debunked, ii) claims that exist in other web resources, iii) logical fallacies, iv) hate speech. Also, the media professional can request for an AI-generated report based on selected articles/debunking reports. Thus, it is aligned with Objective 3 of AI-CODE, and with WP6, the focus of which is on prototyping user-driven services to support media professionals in producing reliable information using AI.
Transparency Service for AI-model Cards		x		x	The Transparency Service provides functionalities for the creation, organization, and consumption of AI model cards, tailored to different levels of detail based on the expertise of the end users. This service directly supports Objective 3, which focuses on providing user-driven tools for media professionals to enhance trusted information production using emerging technologies.
Trustability and Credibility Assessment		x	x		This service aims to serve as a dashboard where media professionals will be able to assess the trustability of information sources in the next generation platforms, concretely the Fediverse. Thus, it aligns itself with Smart Objectives 2 and 3 of AI-CODE, and pillars i) and ii) of the WP4 concrete objectives
Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals		x	x	x	The Interactive Coaching and Simulation Module helps media professionals understand and responsibly use Generative AI tools, aligning with journalistic ethics and standards. Developed through co-creation workshops, it addresses AI-CODE Objectives 3 and 4 by providing user-driven services and co-designing innovative approaches with end-users. The Counter-narrative Recommendation Module generates truthfulness explanations for claims based on trustworthy references, supporting Objective 3 by aiding in trusted information production. It also serves as a learning tool for prompt engineering techniques, demonstrating how different prompts affect explanation accuracy.
Disinformation Detection in the Next-generation Social Media	x		x	x	Identification, analysis, and understanding future developments of next-generation decentralized and immersive social media (incl. emerging Fediverse and metaverse environments) in the context of rapid development of generative Artificial Intelligence (with respect to text, image as well as video modalities) and how such a combination can impact the (dis)information space. For Objective 2 the service will offer content-driven services that support disinformation detection in next-generation platforms (social media and fediverse). It covers also the

					Objective 4, evaluating the tool’s effectiveness in social media scenarios for detecting misinformation.
--	--	--	--	--	--

4 User-Centred Requirements and Design Process

4.1 Outline of the Approach

The starting point for the requirements analysis are previous experiences from related projects such as vera.ai¹, TITAN² or AI4TRUST³ that the AI-CODE partners have been involved in. This ensures that the process benefits from previously gained insights and best practices. The requirements process, see figure 2, is designed to harmonize the technological objectives across multiple work packages, specifically WP3, WP4, WP6 and WP7, ensuring a cohesive approach to the design and development of AI CODE services.

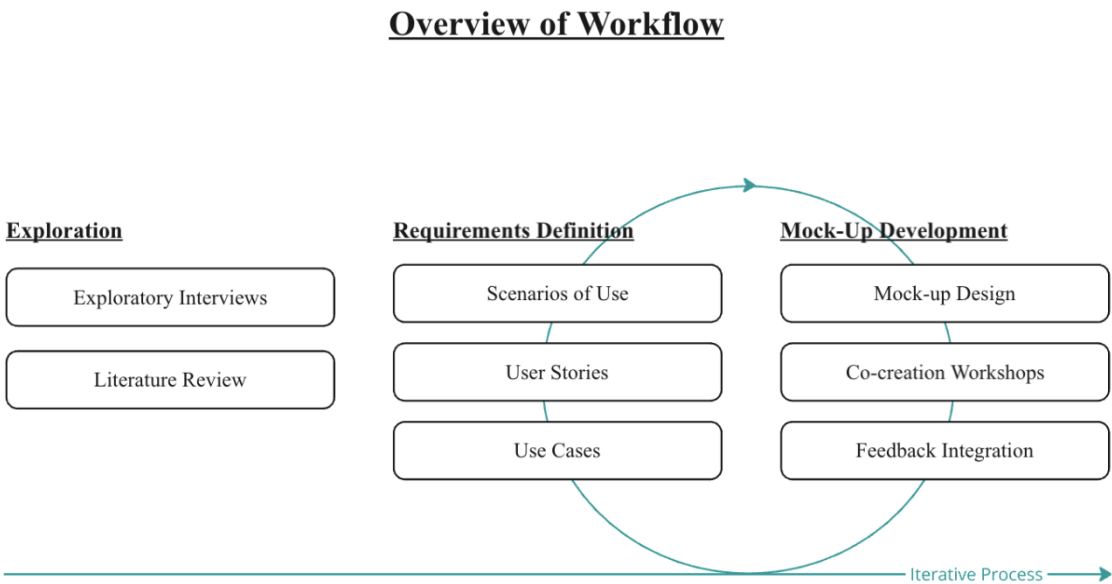


Figure 2: Overview of workflow conducted for the requirements process

To establish an initial understanding of end-user needs and requirements of the broader professional context, the project team conducted a comprehensive exploration phase consisting of:

- Exploratory interviews: A series of in-depth interviews were carried out with 18 media professionals, representing a diverse cross-section of the target user base,
- Literature review: Existing studies on AI needs for media professionals and journalists were reviewed, including an examination of professional ethics guidelines and journalistic standards. This analysis provided valuable context and highlighted industry-specific considerations. The studies reviewed include a large survey of AI in media organisations with 105 news and media organisations from 46 different countries (Beckett at al., 2023), as well as a study of AI in news organisations (134 interview participants from 35 news organizations and 36 international experts from industry, academia, technology, and policy) (Simon, 2024).

Following the interviews and initial user feedback, the project team refined the requirements elicitation process. This refinement naturally led to a multi-tiered approach for capturing and specifying user needs:

¹ <https://cordis.europa.eu/project/id/101070093>

² <https://cordis.europa.eu/project/id/101070658>

³ <https://cordis.europa.eu/project/id/101070190>

D3.1 User-centred requirements definition

1. **Scenarios of Use:** Building on the insights gained from user interviews, the team developed narrative descriptions that illustrate the envisioned usage contexts. These scenarios provide a holistic view of how users might interact with the AI-CODE system, capturing both technological and human aspects of the interaction.
2. **User Stories:** To further detail the user requirements, the team created user stories representing specific combinations of users, contexts, and computer-supported tasks. These stories serve as sub-scenarios, focusing on particular aspects of the system's functionality and user interaction, directly informed by the interview findings.
3. **Use Cases:** As the smallest unit of analysis, use cases were then derived to provide a technical perspective on elements from the user stories. At this stage, a high-level list of use cases for each user story has been developed to define their scope, setting the groundwork for the subsequent creation of mockups and prototypes. More detailed formal use case specifications along with the required system components of each AI-CODE service will be reported in *D3.2 Functional specifications and technical requirements*.

This structured approach ensured a seamless transition from user feedback to concrete requirements, paving the way for the development of initial mockups and prototypes. Building on the initial requirements gathering and methodological approach, the project proceeded to a more interactive phase:

1. **Mock-up Design:** Informed by the developed user stories, mock-ups were created to provide tangible representations of the proposed AI-CODE services.
2. **Co-creation Workshops:** The first round of co-creation workshops was conducted with media professionals. They involved 121 participants including 59 external to AI-CODE (from 12 different professional roles in the media field). These workshops served as a platform for:
 - Presenting the developed mock-ups and user stories
 - Eliciting feedback on the proposed approaches
 - Engaging key target user groups in the design process
3. **Feedback Integration:** The valuable insights gathered from these workshops have been systematically analyzed and incorporated into the project:
 - User stories have been revised and refined for each service
 - Early requirements have been adjusted to better align with user needs and expectations

The iterative nature of this user-centred process ensures continuous refinement and alignment with user needs. The upcoming deliverable, D3.2 "Functional specifications and technical requirements", will encompass:

1. Fully specified (technical) use cases for each AI-CODE service
2. A final specification of both functional and non-functional requirements
3. Detailed formal specifications of AI-CODE technical components and services

This deliverable will serve as a crucial bridge between the user-centred requirements gathering process and the technical development phase, ensuring that the final AI-CODE solutions are both technologically advanced and closely aligned with the needs of media professionals and their professional context.

4.2 Initial Definition of Target User Groups

In the initial phase of the project, the main target user groups were defined for each AI-CODE service. This process involved identifying the primary beneficiaries of each service and outlining the expected benefits for these groups. The initial definitions were based on preliminary research (informed by previous projects and experiences by project partners) and the overall project objectives. Subsequently, these target user group definitions were also subjected to validation through a series of interviews and co-creation workshops. These interactive sessions provided valuable input from potential users and stakeholders, allowing for refinement and confirmation of the initially identified groups. This approach ensured that the project's focus remained aligned with the actual needs and expectations of the intended users, thereby increasing the likelihood of developing services that would be both relevant and beneficial to the target users.

Following the recommendations of the EC Project Review Report to attract more different types of media

D3.1 User-centred requirements definition

professionals (with different backgrounds and with different work experience) to gather more diverse user feedback from users, the definition of the project's main target user groups has been substantially extended for each AI-CODE service (see Table 3).

Table 3: Overview of main target groups per service

Target User Group	Fediverse Network Moderation Tool	Mastodon Instances Inspector	Trustability and Credibility Assessment	Media Asset Annotation and Management (MAAM)	Interactive Coaching and Simulation Module (ICSM)	Counter-narrative Recommendation Module (CNRM)	Transparency Service for AI Model Cards	Personal Companion for Understanding Disinformation
Journalists / Investigative Journalists	X	X	X	X	X	X	X	X
Fact-checkers	X	X		X		X	X	X
Media Organizations	X				X			X
Online community managers	X	X	X			X		
Media Asset Managers				X	X			X
Content Moderators	X	X	X				X	
Media educators & media students				X	X	X		X
PR / communication departments		X						
Brands & Marketing teams		X						
Government Organizations / Policy makers / Regulators	X			X				
Digital content creators of outreach / dissemination / divulgation of present events			X					
Social media managers			X			X		
Social scientists and Researchers in Media Studies/Journalism				X				
Election Monitoring Organizations				X				
Media Archivists				X	X			
Documentary filmmakers			X		X			
Data Scientists (Media Industry)				X			X	
(Publication) Editors							X	X
Technical Editors							X	
Model Card Creators (AI/ML Engineers, AI-System Designer, Researchers, etc.)							X	
Model Consumers (Journalists, Fact-checkers, Developers, etc.)							X	

4.3 Initial Analysis of User Needs

4.3.1 Exploratory User Interviews

Exploratory semi-structured interviews were an important component of a research process informing the user requirements for all services. In recruiting the participants, as well as planning and structuring the interview guide, we were guided by the goal to cover sufficient ground to inform the user requirements, to collect a variety of insights from media professionals coming from different backgrounds, geographical contexts, and carrying out various roles within the media landscape.

Planning and Preparation of the Interviews

The first step in interviews preparation was designing an interview guide. The interview guide was developed through an iterative process, with input from service leaders on topics and issues that would be relevant for them, as well as considering insights from existing studies on user needs for media professionals and fact-checkers. Furthermore, input from partners addressing the ethical perspective in the project ensured the incorporation of relevant ethical principles. Feedback was also collected from several peers from social science research, to ensure that the formulation of questions is not leading, and to identify potential blind spots and biases. The final version of the interview guide covered the following areas; an introduction to the AI-CODE project and the purpose of the interview, media professionals' perceptions and understanding of generative AI, experiences with generative AI in their workflow, ethical considerations and responsibilities related to using generative AI for their work, and a personal outlook on the potential directions of development of generative AI technologies.

Participant Recruitment and Demographics

Interview participants were recruited through the professional networks of project partners. The inclusion criteria required participants to be media professionals working within the media sector of a European country. Ultimately, we obtained a convenience sample, with voluntary participation in the interviews and no external incentives offered. In total, 18 interviews were conducted, involving media professionals from 7 countries (see Figure 1), including the Netherlands (7), Germany (4), Greece (2), Lithuania (2), Croatia (1), Poland (1), and Romania (1). Some of the participants were also internationally involved and have been working within cross-country media landscapes. The participants represented a diverse range of roles within the media landscape (see Table 4). Some individuals hold multiple roles within their company – for example, a person might be both a fact-checker and a team lead. This combination brings a distinct set of responsibilities, and an overarching view of various roles within the media organization. The sample included both junior and more experienced professionals, working in various environments (see Table 4).

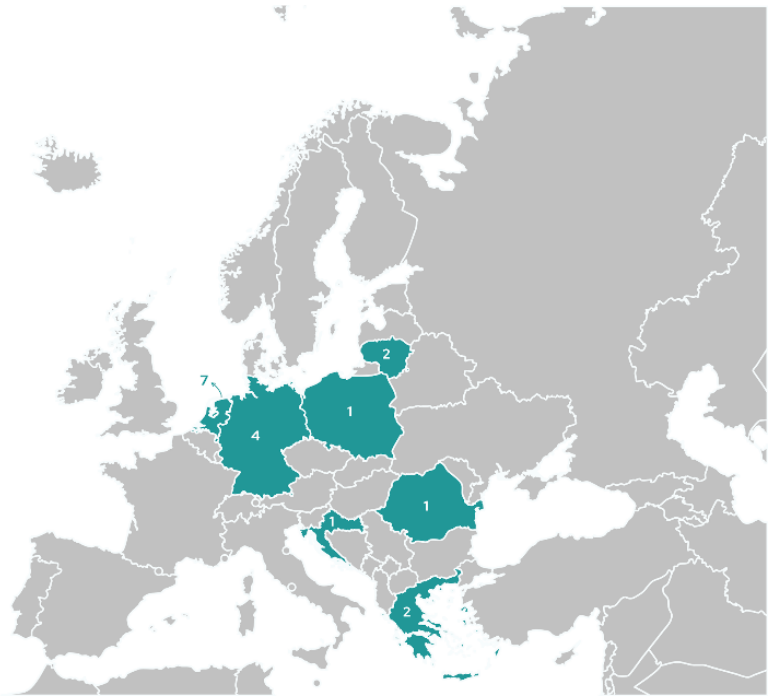


Figure 3: European Countries Represented in Participant Sample

Table 4: Professional Roles of and Types of Organisations Employing Media Professionals in the Sample. Multiple mentions possible.

Role	Number of Participants	Type of Organisation	Number of Participants
Journalist	11	Non-Profit Media Organisation	9
Fact-Checker	4	For-Profit Media Organisation	4
AI-System Designer	3	Public University	2
Editor	3	Non-Profit Cultural Institution	1
Academic Researcher (Field: Media Studies)	1	Public Broadcaster	1
Academic Professor (Field: Journalism)	1	Independent/Freelance	1
Photojournalist	1		
Audiovisual Archivist	1		

Interviewing Process

All interviews were conducted remotely over Zoom during March and April 2024. The average duration of an interview was 1.5 hours. Prior to the interviews, all participants were asked to fill out a short pre-interview online questionnaire (included in annex 1). This questionnaire helped the interviewers gain a better understanding of the participants' professional role, their work contexts, their current use of generative AI technologies, concerns about AI, and experiences with misinformation and disinformation. The information gathered from the questionnaire helped the interviewers to get a better understanding about certain specificities of each participant, which was then used to tailor the discussion during the interviews.

Interview Data Analysis

The analysis of the exploratory interviews was conducted through a rigorous, multi-step process involving three researchers. First, the interview transcripts were cleaned and anonymized to remove any identifying details. All data processing was done in Excel. To deliver the results to the project partners in a most actionable way, the data was first sorted into six main insight categories: Roles, Goals, Needs, Attitudes, Concerns/Problems, Opportunities/Benefits, and Mental Models. These conceptual categories were informed by methods commonly used in user-centred research practices, namely empathy maps and personas, to provide the service partners with relevant and applicable insights regarding their service and its target groups. The sorting process involved analysing each interview by categorizing excerpts (quotes) under specific categories, then coding them with brief, descriptive labels. Excerpts with similar ideas were grouped, even across categories, and each code was mapped to relevant services – allowing multiple mappings per code. Raters applied a broad service-relevance assessment to make sure to include, rather than omit, potentially relevant information for service providers. Each rater coded six interviews, after which inter-rater agreement was assessed by cross-checking coded data; raters reviewed three interviews from each other, flagged uncertain decisions, and reached final decisions through group discussion.

After the individual interview analyses were completed, the coded data was consolidated into a master Excel database. The excerpts were further grouped into thematic categories to facilitate easy browsing by the service partners. A diagram of the code tree is shown in Figure 2.



Figure 4: A diagram of the code tree including four main categories, and all of their respective themes and subthemes that emerged from the interviews data.

Interview Results Delivery to Service Leaders

Service-specific interview insights were delivered to service partners in the format of interactive structured Excel sheets, along with a screencast tutorial on how to most effectively navigate the Excel document, so to find insights most relevant for their service. To better understand the way in which interview data was structured, we provided an overview of all categories, themes and subthemes, together with short definitions of types of insights each theme entails, which could be accessed in the Overview sheet of the Excel document (see Figure 3). Furthermore, the research team used the Claude AI chatbot service to auto-generate concise summaries, or "quick-glances", for each theme and sub-theme. These auto-generated summaries were manually checked and refined by the interview raters to correct wrong or over-simplified and generalised conclusions found in the outputs generated by the Claude AI in certain cases, to omit any errors produced by the system and to ensure overall accuracy and reliability. "Quick-glances" were provided with the purpose for service providers to gain an understanding of which types of insights emerged and can be found within specific themes and subthemes. A preview of "quick-glances" is shown in Figure 4. The service partners were also provided with the ability to filter the insights by category, theme, and sub-theme, making it easier for them to identify the most relevant information for further development of their respective services (see Figure 5). Additionally, small interface adjustments were made to enhance the usability of the shared Excel documents, such as color-coding the categories for easier visual scanning. The extensive documentation and structured presentation of the interview findings aim to ensure that the project partners can easily incorporate the user insights into the ongoing design and development of their services.

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										

Category	Theme	Sub-Theme
AI Literacy		Definitions of AI
		Gen-AI Model Transparency
		Gen-AI Output Reliability
		Gen-AI Training Data Transparency
		Prompting
		Anthropomorphism
		AI Tools
		AI Training
		Data & Privacy
		Colleague Familiarisation
AI Application		Fact-Checking Optimisation
		Language Translation
		Administrative Tasks
		Transcription
		Data Processing
		Editing & Style
		Image and Video Generation
		Integration into Workflow
		AI's Impact on Work Experience
		Participation in AI Tool Development

How to browse the data in this document?

To read this document, it is suggested to first take a look at the "Overview" sheet, where key categories, themes and sub-themes have been listed. When a specific theme/sub-theme has been identified as relevant for your service, and you wish to explore it further, the corresponding data can be found in the corresponding sheet of the category the theme/sub-theme is a part of.

Data in each category sheet is already sorted, so that the inputs are grouped based on the theme/sub-theme. This way, it is easier to find and inspect all data related to a specific theme/sub-theme.

You will notice that the current view is showing only the codes assigned to inputs - if you wish to access quotes related to the codes presented, columns to the right of the respective column can be expanded/unhidden.

Factors around data usage, protection and privacy of said data in AI models/systems.

Overview

Quick Glance

AI Application

Fact-Checking

General Journalism

Career Progression

Uncategorised

Figure 5: A preview of the Overview sheet in the interactive structured Excel document delivered to service partners, containing an explanation on how to best browse the interview data, and how the data is structured.

D3.1 User-centred requirements definition

1	A	B	C	D	E	F	G
2	Category	Theme	Sub-Theme	Quick Glance into Media Professionals' Insights			
3	AI Application	AI Literacy	Definitions of AI	• AI is a technology that can solve problems previously only resolvable by humans, like an 'enhanced version' of Google. • AI systems require initial human input and use algorithms to mimic human thinking, but their processing power is much faster than humans. • The boundaries and appearance of AI systems can reveal their artificial nature. AI is identified by its ability to complete tasks and requests. • AI is viewed as a machine learning system based on human-done data and math to assist with jobs. If something seems intelligent and helpful, it is assumed to be AI. • The term "AI" is considered too broad, with a preference for focusing on specific subfields. AI is human-made and operates differently to help humans, potentially connected to physical data centers. • AI "knows" by scanning and adapting to the information fed into it, learning and recreating things in a human-like manner. • There is uncertainty about the differences between AI, generative AI, and machine learning. • There is awe at the rapid recent acceleration in the AI field, unlike any other. • There is a need to understand how AI works to fully trust its results and identify potential biases like confirmation bias. • Media professionals understand that generative AI generates ideas and offers solutions based on existing data and user prompts. Some of them also describe how AI models predict the next character, deciding which character is most probable, not formulating its own thoughts. • Generative AI is different from other AI that calculates parameters. Large language models (LLMs) find points in a "matrix" closest to the input, not formulating their own thoughts. • Generative AI is trained on data to create new outputs that mimic that data. LLMs respond to			
4			Gen-AI Model Transparency				

+

≡

Overview

Quick Glance

AI Application

Fact-Checking

General Journalism

Career Progression

Uncategorised

Figure 6: A preview of "Quick-Glances" of themes and sub-themes of insights emerged from the exploratory interviews.

AI_Application	Theme	Goals Code	Needs Code	Attitudes Code	Problems Code	Opportunities/ Benefits Code	Mental Models Code	Current Usage Code
72	AI Literacy [Prompting]							Speaks to GPT dismissively
73	AI Literacy [AI Training]							
74	AI Literacy [AI Training]							
75	AI Literacy [Gen AI Output Reliability]							
76	AI Literacy [Gen AI Output Reliability]							
77	AI Literacy [Gen AI Training Data Transparency]							
78	AI Literacy [Prompting]							
79	AI Literacy [Prompting]							
80	AI Literacy [Gen AI Model Transparency]							
81	AI Literacy [Gen AI Model Transparency]							
82	AI Literacy [Gen AI Model Transparency]							
83	AI Literacy [AI Training]							
84	AI Literacy [Prompting]							

Sort Z to A

Sort by colour

Filter by colour

Filter by condition

Filter by values

Select all 25 · Clear

Displaying 26

Familiarisation

✓ Data & Privacy

Digital Archive

Ethical use of AI

Content Ownership

Cancel

OK

+

≡

Overview

Fact-Checking

General Journalism

Career Progression

Uncategorised

Figure 7: Preview of the insights mapped to the "AI Application" category, with the option to filter by themes and sub-themes.

Main Cross-Service Findings from the Exploratory Interviews

While service-specific insights were delivered to the service partners as described above, overarching insights that emerged from the interviews, with cross-service relevancy, are summarised in the following section.

Media Professionals' Current Understanding and Usage Patterns of Generative AI

Interviewed media professionals show varied levels of understanding and use of generative AI, with some likening its function to human reasoning. Most of them describe AI as driven by mathematical and statistical algorithms. More complex mental models of AI could be observed in a few participants with strong technical backgrounds, who, as part of their jobs, participate in developing AI systems themselves. Most of interviewed media professionals are experimenting with generative AI chatbots like ChatGPT and Gemini, while only a minority uses image generators like DALL-E. A few interview participants mentioned using specialized tools for fact-checking such as Vera.ai, or specialised generative AI tools developed in-house. Developing in-house AI tools is viewed as essential for professional journalism, as commercial options are often not tailored to their professional needs. Active involvement in AI development is also seen as crucial for guiding its direction to better serve the goals of media professionals. The most common uses of generative AI among the participants include translation, summarization, idea generation and refinement, overcoming writer's block, writing-style- and spell-checking, text styling, and various administrative tasks. Additionally, for fact-checkers – who represent a distinct group of media professionals in our sample – the use cases for generative

D3.1 User-centred requirements definition

AI differ significantly compared to the majority of participants for whom text-based large language models (LLMs) are most relevant. Fact-checkers face pressure for quick and accurate verification of digital content, highlighting the need for reliable technology to identify AI-generated material, as current detectors are not sufficiently reliable. They recognize the potential of AI in fact-checking to efficiently handle large volumes of content, with some fact-checkers already utilizing LLMs for data coding and labelling. For this group of participants, specialized AI tools for identifying AI-generated content are deemed more relevant for the goal of fighting disinformation than text-focused tools.

Uncertainties in Generative AI Adoption and the Need for Transparent Use & Industry Standards

In general, media professionals expressed a mix of awe at the rapidly developing generative AI technology, as well as the potential opportunities it brings, and a careful caution about it being used irresponsibly. The feeling of uncertainty about generative AI technology is even more exacerbated by the lack of standardized guidelines for ethical AI use within their professions. Therefore, they are charting their own course with generative AI, emphasizing an important guiding principle – transparency in how and when AI is used in their work. In addition to the need for clear industry-wide guidelines on responsible AI use, media professionals also highlight a lack of proper training on how to use this technology effectively. They feel increasingly overwhelmed by the rapid influx of new tools and technologies and emphasize the need for a better understanding of AI to critically assess its outputs, identify biases, and use it responsibly.

Individual and Societal Implications of Generative AI Adoption

Interviewed media professionals expressed significant concerns regarding generative AI's potential societal impact, particularly in areas such as personal data protection and privacy, unreliable outputs producing hallucinations that can lead to inadvertent sharing of misinformation, as well as issues of ownership, attribution, and intellectual property related to LLMs. There is a clear awareness that LLMs inevitably produce hallucinations in their outputs, as they strive to provide satisfactory answers regardless of accuracy. There is greater concern about narrative bias in LLM outputs than intentional misinformation, highlighting the need for disclaimers and plagiarism warnings with LLM-generated content. Additionally, there is apprehension that AI could be manipulated to create harmful content, making it difficult to establish effective guardrails against such misuse. Media professionals express a strong need for clarity on how user inputs are shared and used to improve AI accuracy. Many believe AI tools may collect or sell user data but admit they lack knowledge about these practices. Concerns about using generative AI in their work are spread across two dimensions; individual one, where some feel less satisfied or confident in their own work alongside seemingly "too perfect" AI outputs, with worries about developing an over-reliance on AI tools. Societally, there is a risk of undetectable, unintentional plagiarism, which could harm public trust in media if AI-generated content is published hastily. Journalists stress the importance of using AI in a way that upholds independent journalism and maintain professional standards, calling for a collective effort within the media community. This requires stronger collaboration among journalists, fact-checkers, system developers, and researchers, as well as building bridges between media professionals, innovation teams, and external partners.

4.3.2 Summary of main findings from related user-centred literature

Needs and challenges encountered by journalists and news organisations

Journalism has long since been shaped by technology. As AI technologies evolve, they pose challenges to journalism by influencing workflows, content, and organizational dynamics in three key ways (Cools & Diakopoulos, 2024). First, algorithms may unintentionally introduce bias across the journalistic process. Second, generative AI tools capable of creating text, images, and videos from prompts can undermine content accuracy and credibility (Jones et al., 2023). Third, the lack of transparency of certain AI algorithms potentially limits diversity of perspectives, editorial independence, and human decision-making (Beckett & Yaseen, 2023). This creates a need to support journalists in their daily workflow to deal with the challenges of (generative) AI technologies. To uncover these needs, we discuss the role of AI in journalism from four perspectives, after which we summarise the needs:

D3.1 User-centred requirements definition

- AI as implemented by news organizations in the daily workflow (Simon, 2024; Diakopoulos et al., 2024),
- the role of companies offering AI platforms (Simon, 2024),
- responsible use of AI in news organisations (Diakopoulos et al., 2024).
- obstacles encountered by fact checkers (Hrckova et al., 2022).

Use of AI in news organisations

AI is already being used in news organisations on a large scale. Simon (2024) identified four motives for news agencies to employ AI in their work processes:

- technological developments,
- market pressure
- industry dynamics
- uncertainty, hype and hope.

However, these four motives as explained by Simon (2024) appear to be closely related as part of a process. First, technological developments are considered a motive just because they ‘happen’. This appears to drive the uncertainty, hype and hope, which are explained as efforts to use AI just because it *could* do something special; it has potential. The hype is also driven by industry dynamics, which Simon (2024) describes as peer pressure. The hopes of AI’s potential drive the market pressure, which centres around the hope to compensate for the collapse of the old business model by increasing the number of subscribers and efficiency of the journalistic workflow.

When it comes to the use of AI in the journalistic workflow, the efficiency expectations are related mostly to news production (Simon, 2024). Examples are finding connections between large datasets, generating texts for social media, headlines, story outlines, rewriting text. Additionally, AI is used for generating images, videos, text to speech, transcription, translation, generating metadata and enhancing the user experience (Diakopoulos et al., 2024; Simon, 2024). AI has made this much more accessible, which leads to more input for stories, which increases quality and output (Simon, 2024).

Other categories of AI usage are information gathering and sensemaking, which encompasses news discovery, research, ideation and curation. Automated fake news detection was also mentioned (Diakopoulos et al., 2024). Interestingly, there is both a strong interest in delegating activities to AI and resistance to getting help from AI, especially around creative tasks. The most prominently mentioned concerns feeding the hesitation to accept help from AI in the workflow are potential introduction of bias and lack of accuracy, originality, transparency and relevance. Diakopoulos et al. (2024) suggest mitigating this by formulating different success criteria for assessing the output quality of tasks performed by AI versus tasks performed by humans. This way, it becomes clearer what to delegate to AI and what to humans. Diakopoulos et al. (2024) propose that new interfaces are needed to support this type of decision making and journalistic work.

A special note is made regarding the use of LLMs (Simon, 2024). Summarizations are often not that good and it takes journalists and fact checkers a long time to check them. This is something they take seriously, as they are afraid of reputation damage if AI inaccuracies should get ‘out there’. A particularly poignant quote from a U.K. editor is: “The job of journalism is to find stuff not on the internet already. Artificial intelligence won’t be able to do that.” (Simon, 2024, pg. 21).

The role of AI platforms and platform companies

AI is primarily used in business and distribution tasks but is increasingly applied to news creation and production as it advances. Major platforms like Google, Meta, Microsoft, and Amazon dominate this space, offering AI tools and infrastructure that news organizations rely on due to limited resources for in-house development. This dependency undermines autonomy, as platforms dictate service access and pricing. Switching platforms is costly and disruptive, especially when platforms discontinue acquired services, affecting workflow continuity.

The lack of transparency in AI models (the "Black Box" issue) raises concerns about bias in training data, privacy, and adherence to journalistic values. AI also poses a strategic threat by using news organizations’

D3.1 User-centred requirements definition

content without compensation, improving the very systems that could diminish the journalist's role.

Smaller news organizations, particularly in non-Western regions, face greater challenges due to limited resources for AI adoption. While platform AI offers convenience, over-reliance risks dependency, inequality, and loss of journalistic independence.

Responsible use of AI in the organisation

On a higher level, Diakopoulos et al. (2024) identified concerns related to losing control over AI output, lack of human oversight and lack of disclosure when generative AI is used. They found that respondents think that uses such as generating entire pieces of content, generating interview questions or replicating artists' styles in image generation should be banned altogether. Respondents wanted to receive proper training in how to use generative AI more responsibly to overcome these concerns.

Current strategies for responsible use of generative AI focus on monitoring the output rather than monitoring the input and the models themselves. In other words, there was limited critical engagement with ethical and legal concerns around the training data and the model, and the trustworthiness of the technology. It thus seems that generative AI is pressuring individuals to learn new skills to keep up with the developments, for which respondents also fear there might not be enough time and resources to facilitate this learning.

Obstacles encountered by fact checkers

Hrckova et al. (2022) discovered that fact checkers' most pressing issues are related to monitoring the online space, selection of potential false claims, communication and avoiding duplicity, verification of content credibility and the dissemination of fact checks.

Fact checkers emphasise the need for tools to automate alerts, filter information, search for sources, and identify multiple versions of misinformation, particularly from hard-to-access documents, statistics, or social media. They also lack a dedicated messaging platform, relying mostly on Facebook, and often fail to optimize fact-checking output for online visibility, which limits collaboration with larger international organizations (Hrckova et al., 2022).

Additionally, fact-check selection criteria are often subjective, prioritizing societal impact over virality but influenced by political, cultural, and audience factors. Eastern European fact checkers face additional hurdles, such as language barriers that complicate searches and reliance on Western fact checks (Hrckova et al., 2022).

Needs

From the issues discussed by Simon (2024) and Diakopoulos et al. (2024), we derive the following lists of needs for news organisations that are relevant to this project:

- more transparency on the workings of AI models
- gain transparent insights into what AI can and cannot do
- equality between news organisations
- assistance with monitoring and scanning different media
- design of new interfaces to support journalistic work with gen AI
- formulate clear quality criteria for assessing the output quality of a task performed by AI versus a task performed by humans
- learning prompt engineering to formulate prompts that maximize efficiency and productivity
- receive training on how to use generative AI responsibly

For fact checkers in particular, Hrckova et al. (2022) identified the following list of challenges, which the services to be developed in this project will address:

- Searching for the image manipulations
- Searching for the source of evidence for verification
- Searching for the source of evidence in videos
- Search within the existing fact checks
- Searching for other versions of misinformation

D3.1 User-centred requirements definition

- Seeking factual claims suitable for fact checking

Guidelines and policies on AI use in media organizations

In order to put the requirements for the AI services into the context of practice we carried out a survey of AI guidelines created by public service media organizations in Europe. We found some of the documents through a web search, and others through the European Broadcaster's Union AI Ethics Groups Guidelines Repository (EBU, 2023). The EBU is a knowledge-sharing membership organization of European public service media organizations and has a long history of addressing the impact and use of AI in the media field through conferences, interest groups, publications, and research projects (EBU, 2024). The documents shared in the EBU AI Ethics Group Guidelines Repository reflect a variety of institutional responses to the threats and opportunities that media organizations are encountering with AI and generative AI. Some of the documents offer only general guidance and others outline institutional policies in a detailed way. We decided to analyze only documents that include clear and detailed policies because they offered more opportunity to clarify major areas of similarity and comparison between institutional approaches.

We analyzed the following documents to understand what the major areas of concern are in using generative AI are, and how organizations are devising and instituting frameworks to handle these concerns:

- BBC Editorial Guidance on the use of Artificial Intelligence (UK)
- Czech Radio Rules for Using Artificial Intelligence (Czech Republic)
- LRT Public Enterprise Lithuanian National Radio and Television Guidelines for the Use of Artificial Intelligence (Lithuania)
- NRK's Guidelines for Editorial Use of Artificial Intelligence (Norway)
- Radio Free Europe/Radio Liberty/ Standards of Ethical Journalism/ Use of Generative Artificial Intelligence
- SVT: General policy for AI usage at SVT (Sweden)
- SWR Editorial guidelines for "artificial intelligence" (Germany)
- ZDF Guiding principles for the use of Generative AI (Germany)

As we analyzed the documents, we distilled four major categories that media organizations are addressing in their policies and frameworks. They include:

- 1) Prohibited and allowed use of generative AI
- 2) Requirements related to transparency
- 3) Guidelines related to data, intellectual property, and security
- 4) Assigning legal and ethical responsibility when using generative AI.

The approach to these topics is handled differently in each institution, and below we will outline these differences and discuss them in turn.

D3.1 User-centred requirements definition

Table 5 illustrates various approaches that different media organizations have to the use of different kinds of Generative AI, and to transparency requirements related to the use of AI in their policy documents.

D3.1 User-centred requirements definition

Table 5: Overview of media organizations' approaches to different kinds of Generative AI

Organization	Generative AI for Text Generation	Generative AI for Image Generation	Generative AI for Audio Generation	Generative AI for Research	Transparency Requirements for use of Generative AI
BBC	Prohibited for news content.	Prohibited except for illustrative purposes.	Allowed for synthesized voice; must not replicate others.	Allowed for insight or analysis, but not for direct content creation. Risk analysis procedures in place.	The audience should be informed that AI has been used and how.
Czech Radio	Subject to prior approval; generally for creative purposes and not news.	Allowed for illustrations or fictitious content.	Allowed; digital voiceprints are created under a licensing agreement texts written by individuals.	Allowed for generating source codes for Software and Apps.	Requires prompts to be logged; visuals must include acknowledgments and watermarks.
LRT	Allowed; AI-generated text uploaded to the web portal must be reviewed by a responsible employee	Allowed, but prohibited to use AV content to manipulate, discriminate, or mislead.	Allowed, but prohibited to use AV content to manipulate, discriminate, or mislead.	Allowed for filtering AI-generated content published online.	Fragments generated by AI tools must be clearly marked.
NRK	Prohibited except for (disclosed) automated journalism (sports results, stock listings, real estate sales, etc.)	Prohibited for photo-like content unless illustrating AI use; transparency required.	Prohibited for impersonating voices; ensures confidence in real journalist voices.	Allowed for generating summaries, SEO, and other non-critical text (need not be labelled if edited by journalists).	Requires transparency for AI-generated visuals and automated text summaries.

D3.1 User-centred requirements definition

Radio Free Europe/ Radio Liberty	Prohibited for all published or broadcast text, including headlines and captions.	Prohibited.	Prohibited.	Allowed for research, brainstorming, and task planning. Emphasizes fact-checking for AI-generated research outputs.	Not specified.
SVT	Prohibited for headlines, intros, or fact boxes unless approved.	Prohibited use of realistic quality in news contexts.	Prohibited to use sound (e.g., voices) in news contexts.	Allowed, should be read by a human before it becomes part of a publication. Should be fact checked, and bias should be considered.	If exceptions are granted and generative AI is used, it should be made clear to audiences and the AI model should be disclosed.
SWR	Allowed; every piece of AI-generated content must be approved by a human.	Allowed; every piece of AI-generated content must be approved by a human.	Allowed; every piece of AI-generated content must be approved by a human.	Allowed; every piece of AI-generated content must be approved by a human.	Must disclose what AI tools were used, and place in a "How we worked" fact box.
ZDF	Allowed; must follow a two-source rule where generative AI does not count as a source.	Allowed; must follow a two-source rule where generative AI does not count as a source.	Allowed; must follow a two-source rule where generative AI does not count as a source.	Allowed; must follow a two-source rule where generative AI does not count as a source.	Content created with generative AI is identified as such for users .

Guidelines related to data, intellectual property, and security.

Media organizations face complex issues related to the use of generative AI and the legal and ethical handling of data, intellectual property, and security. The BBC (2024) explains the interconnected issues: “There may be legal and commercial rights issues affecting whether the user of the AI or developer of the AI owns or is liable for any output created by it” (p. 3). The policy documents include various approaches to how organizations must handle concerns related to the categories of potential rights infringements of others, the handling of private information in a careful manner, and security and cyber resilience.

Potential rights infringements of others: Media organizations understand that the output created by generative AI systems may contain training data that could infringe copyrights of others if published. Responding to this potential liability, LRT (2024) states that if using an AI tool, a user must “ensure that the

D3.1 User-centred requirements definition

AI tool has obtained permissions or licenses to use the content used to train it for learning purposes” (p. 3), and in order to mitigate any legal liability, LRT states that “Contracts with the provider of the AI tool must include a guarantee in case of third-party/author claims” (p. 3). Organizations recognize that in the future there may be disputes related to potential rights infringements that are not now considered. Many organizations are responding to this lack of clarity about rights by restricting the use of generative AI, and others are using it only in very specific cases that can be controlled, such as Czech Radio’s use of Digital Voiceprints created with contracted individuals, to avoid any potential future disputes.

Handling private information in a careful manner: Media organizations understand that content they share with generative AI may become part of its training data, and that they must be careful to protect this data. For example, NRK (2023) stipulates that “Unpublished editorial material (such as interview transcripts, etc.) should not be shared with such services” (p. 2). And while NRK Journalists are allowed to use transcription services, they must use them in local storage environments.

Security and cyber resilience: Most of the documents do not recognize the considerable risks related to cyber attacks and vulnerability that AI tools may present to media institutions. Czech Radio (2024) mentions the following risks: “AI tools may face cyberattacks that may lead to data leakages, system damage or stopping of operation” (p. 5) and “AI tools may contain security vulnerabilities that may be misused for attacks” (p. 5). In order to deal with these kinds of risks, LRT (2024) creates the guideline that “AI tools must be used in LRT operations after they have been fully vetted and evaluated for their cyber resilience, IT security and risk profile” (p. 4). Additionally, to attempt to safeguard the organization’s infrastructure, they allow “the use of free/public AI tools only from personal accounts/without linking them to LRT accounts/information assets” (p. 4). Security concerns are an area that has received little attention in the media community’s public-facing documents.

Assigning legal and ethical responsibility when using generative AI

The AI policy guideline documents address the complicated question of how institutions are assigning legal and ethical responsibility for use of generative AI inside of their own organizations. Some organizations have also created internal groups to handle questions related to the use of AI. The approaches in this section reveal where responsibilities lie, and where decisions are made within organizations.

Table 6: Overview of responsibilities and committees in place at different media organisations

Organization	Responsibility Assignment	Committees/Groups
BBC	Senior editorial figures oversee AI deployment and usage; editorial line managers manage AI use in teams.	AI Risk Advisory Group (AIRA) provides detailed advice on editorial and non-editorial AI risks. Members include legal, data protection, and infosec experts.
Czech Radio	Employees are personally liable for AI-generated outputs, similar to other tools (e.g., web searching). Competent managers assess risks and mitigation.	AI Use Strategic Committee carries out risk and impact assessments for AI use.
LRT	General Manager ensures implementation of AI guidelines, while personal responsibility remains with those involved in development and use of AI content.	Not specified.
NRK	Not specified	Not specified
Radio Free Europe/ Radio Liberty	Not used, not specified	Not specified
SVT	Heads of departments provide support for correct AI tool usage. Legal	Not specified

D3.1 User-centred requirements definition

	responsibility for published content lies with a person defined by freedom of speech laws.	
SWR	Content created using AI must be approved by a human before publication. The human retains control and responsibility, but specific roles are not clarified.	Interdisciplinary team of editorial, legal, and technical expertise
ZDF	Editorial departments are responsible for generative AI use in programming, planning, and communication.	Not specified.

Conclusion

Along with the areas listed above, the documents share general points of agreement on how media organizations need to approach AI – with clear concerns about how generative AI presents threats to media independence, journalistic integrity, and the trust that the public has in the media that is essential to healthy democracies. Public service media organizations express their understanding that their use of generative AI needs to align with their public service journalism ethics and values, and that part of their mission is to explain generative AI to audiences as societies move through the rapidly changing impact of artificial intelligence on all sectors of our lives, including media. The documents reflect how organizations understand that the speed of AI development requires an approach that can shift as technology, law, and regulations change, and that humans need to retain oversight and responsibility in the process. Media organizations express specific concern about serving the needs of vulnerable groups, including children, and making sure that biased training data doesn't encourage or amplify discrimination.

These documents also reflect that media organizations recognize the fundamental existential challenge that generative AI presents to public service media through threats of job loss and challenges to maintaining intellectual property due to unlicensed use of content in generative AI models. Other risks to be more fully examined include one that Czech Radio (2024) explains as: "Risk of loss of control: AI may be used to control humans without them being aware thereof. This may lead to loss of freedom and democracy" (p.6).

4.3.3 Summary of ethical frameworks that informed the ethical requirements for the services

Ethics matters greatly in designing AI systems, due to the manifold challenges AI systems might present, ranging from possible impacts on labour and (creative) production (Bordot 2022), the potential for intensifying surveillance and infringing on privacy rights (Kalluri et al. 2023), to reproducing harmful biases (Buolamwini, & Gebru 2018), and epistemic challenges and responsibility attribution (Coeckelbergh 2020). Generative AI has only amplified these concerns. In an effort to ensure that ethics is not an afterthought in the AI-CODE project, ethical design requirements were identified alongside the user driven design requirements in the first year of the project. The list of ethical requirements has been identified through a two-step process.

Step 1: Ethical design requirements in the Ethics Guidelines for Trustworthy AI

The starting point for identifying the ethical design requirements for the AI-CODE project were the Ethics Guidelines for Trustworthy AI (Ethics Guidelines) published in 2019 by the European High Level Expert Group on AI (AI-HLEG) (European Commission 2019). The work of the AI-HLEG was selected as the starting point because of two main reasons: i) as a European Horizon project, the work of the European Commission on the ethical implications of AI are the intuitive and logical starting point; ii) the Ethics Guidelines are rather influential and the review of ethical guidelines by Corrêa et al. (Corrêa et al. 2023) shows, that there is overall a convergence around ethical principles such as those suggested by the AI-HLEG.

The AI-HLEG states that an AI system can be deemed 'trustworthy' if it is lawful, ethical, and robust (AI-HLEG, 2019, p. 2). This is further elaborated on in seven key requirements that need to be met so that the development, deployment and use of AI systems can be considered trustworthy; namely (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5)

D3.1 User-centred requirements definition

diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.

The ethical design requirements for AI-CODE were drafted based on these seven requirements, by asking the question: “What is needed in the design of the system to be able to say that one has fulfilled this requirement?” The AI-HLEG was also tasked with developing an Assessment List for Trustworthy AI (ALTAI) which will be used later in the project.

Step 2: Tailoring the design requirements to AI-CODE

Since the AI-HLEG Guidelines are rather broad and not context specific, the initial design requirements were reviewed and elaborated on with service-specific ethical concerns in mind. These service-specific ethical concerns were identified through a brainstorming and mapping exercise with the entire AI-CODE consortium on a shared Miro board. In the first stage of the brainstorming, the CEPS research team reviewed literature on the ethical aspects of the use cases. Building on this literature, as well as relying on their previous expertise in the ethics of AI and in collaboration with the project’s ethics advisor, a preliminary list of ethical concerns for each service was identified. In a second step, the developers of the services reviewed the preliminary brainstorming, clarified possible misconceptions of the services, and contributed their own questions and ethical concerns.

These two steps have produced the preliminary list of ethical design requirements for AI-CODE given in Table 7. The design requirements will be elaborated further in D9.5 where they will be furthermore extensively operationalised. Therefore, the work presented here is preliminary. Concrete changes to the design of the AI-CODE services for ethical soundness have been taken based on the feedback of the Ethics Advisory Board (EAB) in their first meeting in M12.

Table 7: Preliminary list of ethical design requirements

#	Ethical requirements	Related to Trustworthy AI requirement
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system	Human agency and oversight (HAO)
HAO-02	The capabilities of the application shall be communicated clearly to the user.	HAO
HAO-03	The application shall include a detection and response mechanism for undesirable adverse effects of the AI system.	HAO
HAO-04	The application shall take into account the risk of stimulating dependency in users.	HAO
TSA-01	The application shall have a risk assessment system i.e., risks, risk metrics and risk levels of the AI system are defined.	Technical robustness & safety + Accountability (TSA)
TSA-02a	The accuracy, reliability, and reproducibility of the application shall be monitored, documented, and communicated to the user over time.	TSA
TSA-03	The data used to build the application shall be up-to-date, of high quality, complete and representative of the deployment environment.	TSA
TSA-04	The application shall be resilient to cyber-attacks.	TSA
PDG-01	The application shall minimize the amount of data collected on users.	Privacy and Data Governance (PDG)
TRA-01	The system input and output of the application shall be traceable through documentation.	Transparency (TRA)
TRA-02	The decisions of the application shall be understandable by the user (XAI) as far as reasonably achievable.	TRA

D3.1 User-centred requirements definition

TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system.	TRA
FAI-01	The application shall have procedures in place that avoid unfair bias that apply to data, design, and output.	Fairness (FAI)
FAI-02	The application shall take the diversity and representativeness of the end user into account.	FAI
FAI-03	The application shall have a mechanism that allows for flagging issues related to bias, discrimination or poor performance.	FAI
ENV-01	The application shall identify, evaluate, and where possible mitigate environmental impacts (embed a carbon tracker) and impact on human working arrangements and democracy.	Environmental and societal wellbeing (ENV)
ACC-01	The application is designed so that it is auditable to third parties i.e., the development process and sourcing of training data are traceable and the AI system's process, outcomes and impacts are logged consistently.	Accountability (ACC)
TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated	Consortium
ENV-02	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects	Consortium
HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case	Consortium

4.4 Development of Early Mock-ups and Prototypes

As part of the user-centred requirements elicitation process, early mock-ups and prototypes were developed. This approach is a common and integral component of user-centred design methodologies. The creation of mock-ups and early prototypes serves to materialize concepts and ideas for a given service, making them tangible and easier to evaluate.

These visual and interactive representations were utilized for multiple purposes:

1. **Communication:** They provided a common reference point for discussions among team members, stakeholders, and users.
2. **Illustration:** The mock-ups and prototypes helped to visually demonstrate proposed features and functionalities, making abstract concepts more concrete.
3. **Validation:** Early ideas were tested and validated through these representations, allowing for timely identification of potential issues or areas for improvement.
4. **Feedback Elicitation:** They served as effective tools to gather specific and actionable feedback from target users during the co-creation workshops.

The development of these mock-ups and prototypes was carried out in the corresponding service tasks T4.1, T4.2, T4.3, T6.1, T6.2 and T6.3. Subsequently, they were employed in task T3.4 for the preparation and execution of co-creation workshops.

This iterative process facilitated the development and revision of user stories and requirements. By presenting tangible representations of the proposed service, users could more easily envision its functionality and provide valuable insights. This feedback was then incorporated into the refinement of user stories and the adjustment of requirements, ensuring that the final service concept and design will align closely with user needs and expectations.

The use of mock-ups and prototypes in this manner demonstrates a commitment to user-centred design principles, fostering collaboration between developers and end-users throughout the development process. This approach typically leads to more user-friendly and effective final products, as it allows for continuous

D3.1 User-centred requirements definition

validation and improvement based on direct user input. Figure 8 depicts a small sample of the mock-ups and prototypes developed by the project. Further screenshots along with detailed descriptions can be found in the individual chapters for each service in part 2 of this deliverable.

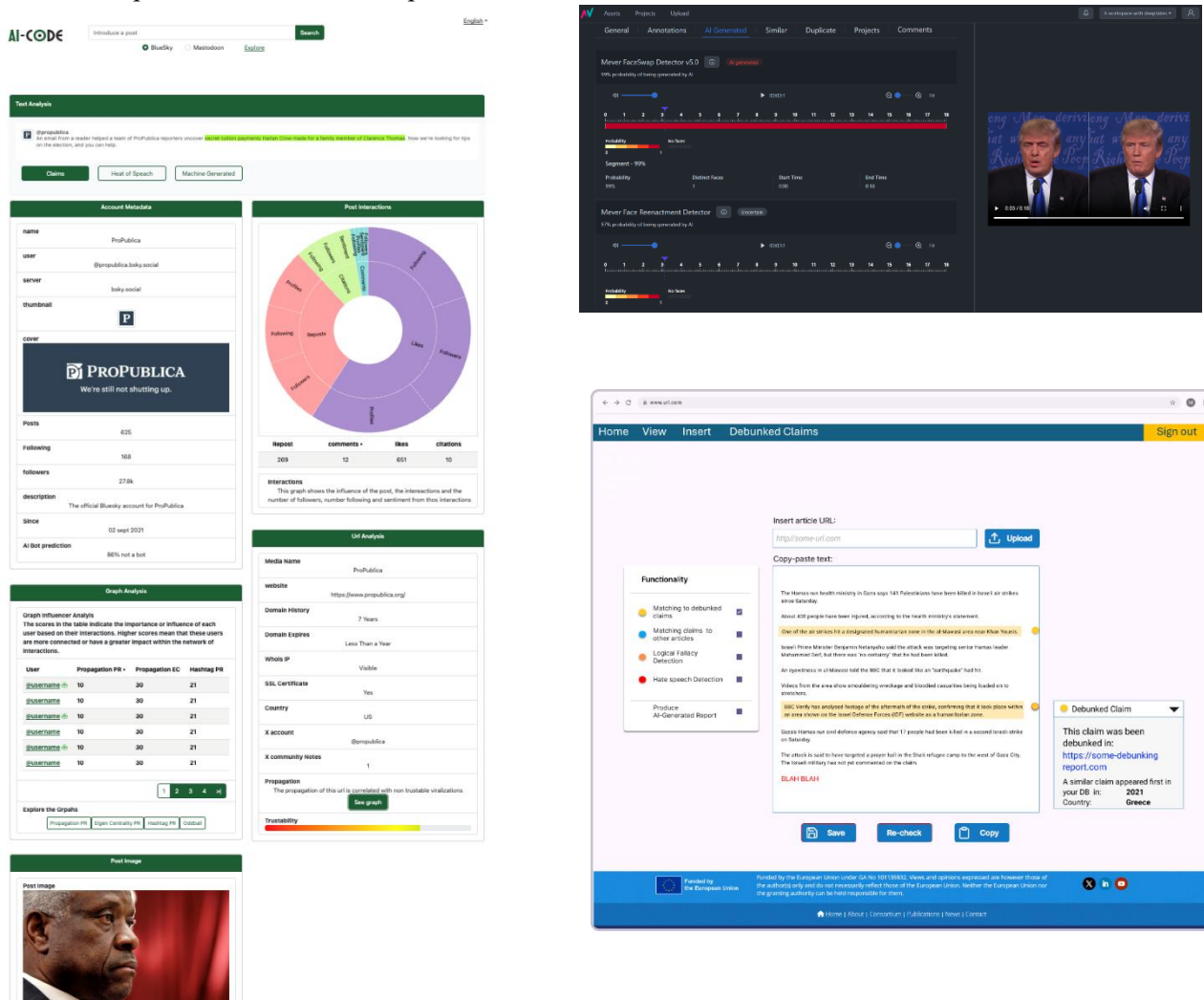


Figure 8: Selected screenshots of service mock-ups: Trustability and Credibility Assessment (left), Media Asset Annotation and Management MAAM (top right) and Personal Companion for Understanding Disinformation (bottom right)

4.5 Co-Creation Workshops

4.5.1 Workshop methodology and approach

As part of the design process, a series of online co-creation workshops were conducted by each of the services to test their current concepts and develop user-centred requirements further. While the content of the co-creation workshops was tailored to individual services and their goals, an overarching methodology was employed, as provided by EIPCM through a preparation session, along with self-made Miro workspace templates, see Figure 9, and template slides to be adapted by each service. Workshops were held online (Zoom, MS Teams) and Miro was used as a useful software for online collaboration.

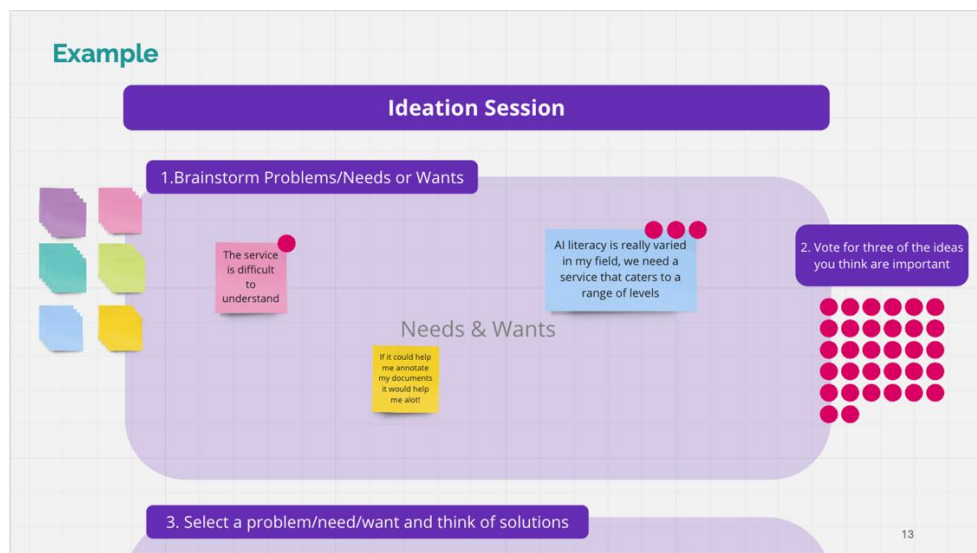


Figure 9: Example Miro use case (update this caption too) from the co-creation coaching workshops facilitated by EIPCM

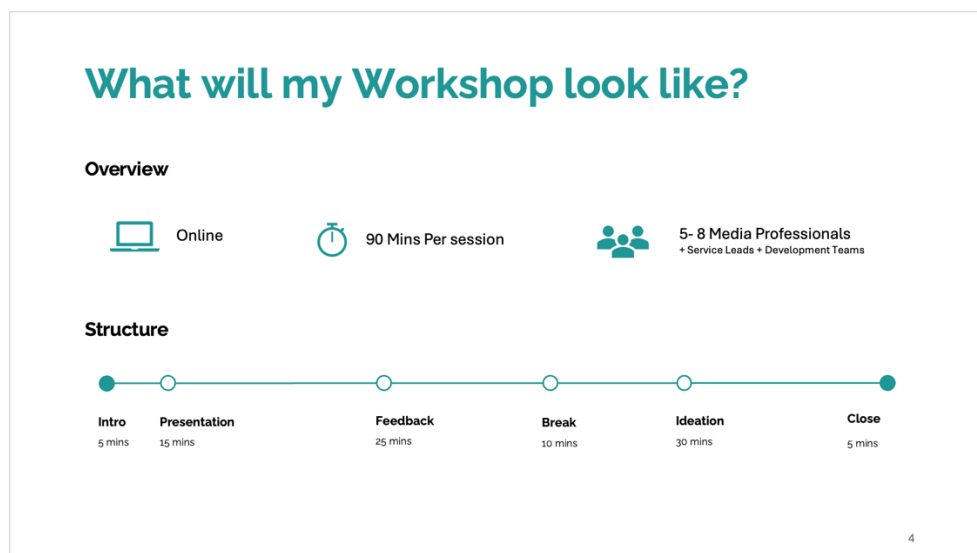


Figure 10: Overview of main steps in the overall methodology set by EIPCM and adapted by services

The basic structure, seen in Figure 10, is as follows:

First, session leaders would introduce the AI-CODE project as well as the service partner, including goals and backgrounds. The broad concept of the service was explained, along with the aims of the session and general principles, see figure 5, pertinent to understanding the need/basis. Current mock-ups or prototypes were shown to participants, to illustrate proposed functionalities and visualise ideas clearly. A feedback session was then conducted to ask participants of their experience with content presented so far. As an example, EIPCM listed the following categories for their feedback session:

- Content feedback
- Proposed Content
- Misconceptions
- Presentation Style
- Worries & Concerns
- Aha Moments
- Additional Suggestions

D3.1 User-centred requirements definition

Service leaders asked questions specific to their individual goals for the session. Lastly, it was encouraged for services to conduct an ideation session, figure 6, where participants could, for example, further develop thoughts discussed in the feedback session. Leaders without prior facilitation experience were given coaching and support by EIPCM to encourage participation and discussion through techniques such as asking individuals for answers and asking open questions. EIPCM also participated in each session to support the moderator in the workshop facilitation and feedback elicitation from the participants.

Example Use Case

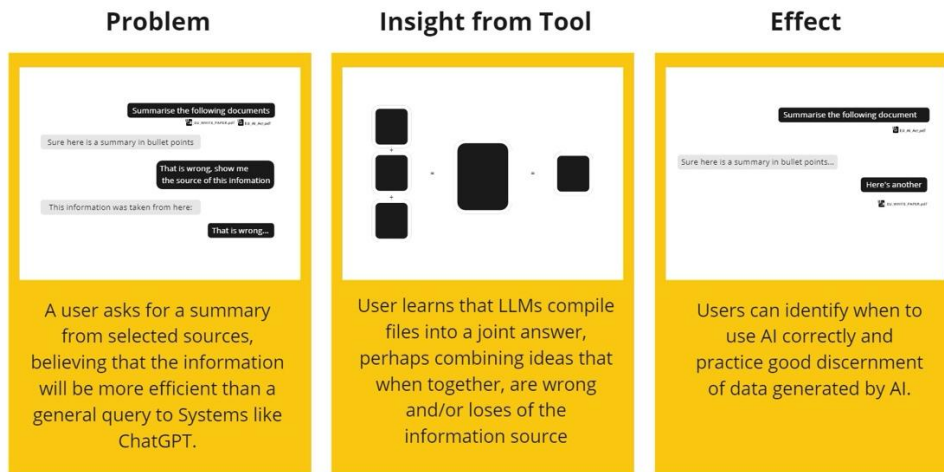


Figure 11: Example Use-case used by EIPCM in their workshop to illustrate the principles of their service



Figure 12: EIPCM's Ideation session where users identified points of intervention for their service with users

Feedback and ideas from the co-creation workshop was used by services to improve their functionalities or communication. As aforementioned, service leaders who felt that they could gain more from additional workshops were able to implement changes and conduct new workshops with participants. From the session, services could better understand how their service was received by users; additional points to consider and potentially new features to implement.

4.5.2 Summary of main findings from end-user feedback

In total 11 co-creation workshops were conducted, with seven initial workshops held between 07/06/2024 - 28/06/2024 and additionally four workshops conducted between 02/08/2024 – 30/09/2024. The workshops had 121 participants in total, 59 of them external to the AI-CODE project. 56 % of the participants were men and 44 % woman, covering a substantial diversity of roles of media professionals engaged across modules,

D3.1 User-centred requirements definition

confirming the relevance of the tools to various stakeholders in the media profession, and in research (see Table 8). Workshops were primarily held over Zoom, in addition Miro was used as a useful software for online collaboration.

Table 8: Roles of participants to co-creation workshops

Participant Role	Co-Creation Workshops					
	Disinformation Detection in the Next-Generation Social Media	Trustability & Credibility Assessment	Media Asset Annotation and Management	GenAI Interactive Coaching Service & Dynamic Simulator	Transparency Services for AI-Model Cards	Personal Companion for Understanding Disinformation
Journalists		✓		✓	✓	✓
Fact-checkers			✓	✓		✓
Media / Data Analysts		✓		✓	✓	
Innovation Managers	✓			✓	✓	
Communication Manager	✓			✓		
Digital Marketing Manager	✓					
AI Ethics Specialist					✓	
Policy Directors					✓	
Election monitoring experts			✓			
(Online) Community managers				✓		
Media / Journalistic education	✓			✓		✓
Media Archivists				✓	✓	
Researchers (AI, Disinformation, Media)	✓	✓	✓	✓	✓	

The Disinformation Detection in Next-Generation Social Media service received feedback focusing on service accessibility outside Europe, interoperability with existing services, and emphasis on content analysis

D3.1 User-centred requirements definition

beyond metadata. New ideas for potential features included enhancing transparency regarding the analysed sources, providing tools to verify the sources (especially important for OSINT researchers), and providing a trustworthiness scores for the instances.

For the Trustability and Credibility Service, participants provided extensive feedback on the Trustability and Credibility Assessment UI, emphasizing the need for clear navigation, understandable metrics, and fine-grained classifications beyond binary true/false options. Visual elements and multi-format analysis received positive feedback, with users supporting an aggregated score system that maintains access to individual analyses for users with varying expertise levels. Missing features were identified, including trustability level labels and external AI-CODE services integration, while a valuable "history of research" feature was discovered to enhance application reusability and response capability. The feedback will guide improvements to explainability, display methods, and overall user experience across different expertise levels.

For the Media Asset Annotation Management (MAAM) service, participants validated the importance of AI-generated content detection and near-duplicate identification for election monitoring and fact-checking manipulated political content, with a specific need for side-by-side comparison tools to analyze altered segments in long political videos. Audio analysis capabilities were identified as critical due to the rising threat of deepfake audio in disinformation campaigns, while participants suggested a layered interface approach to balance technical detail with clarity for users of varying expertise levels. Collaboration features emerged as a top priority, with teams currently relying on inefficient manual processes like Excel sheets for content annotation and expressing strong interest in centralized collaborative tools with shared annotation spaces. Key improvement opportunities included developing better comparison tools for long-form video content, enhancing audio forensic capabilities, addressing bias in detection models toward Western media content, creating built-in collaborative annotation environments, and implementing features to prevent duplicate analysis efforts within teams.

The Interactive Coaching and Simulation Module workshops included demonstrations of LLM prompting challenges and interactive mock-ups. Participants appreciated the relevance and clarity of the content, emphasizing the importance of understanding fundamental concepts of LLMs, such as temperature and system prompts, to effectively utilize AI in journalism. Many expressed a desire for more detailed explanations tailored to varying levels of prior knowledge, as well as a community-driven feature for topic suggestions. Misconceptions were addressed, with participants realizing that LLMs are not as complex as they initially thought and that politeness in prompts does not affect output quality. Concerns were raised about accessibility for smaller media teams and the sustainability of the service as LLMs evolve, stressing the need for core concepts that remain relevant. The interaction style received positive feedback, with users enjoying the "behind the curtain" approach and suggesting more storytelling elements. Lastly, participants shared "aha" moments, particularly regarding the tendency of LLMs to fabricate answers rather than admit uncertainty.

For the Counter-narrative Recommendation Module, participants provided positive feedback on the tool's ability to save time and streamline daily tasks by offering quick access to relevant information and aiding in real-time information verification on social media. They suggested enhancing usability by increasing flexibility in providing source materials, such as allowing users to input links and upload files directly. Users also recommended presenting multiple verdicts simultaneously to showcase different writing styles and arguments, as well as providing context around relevant document sections to improve interpretability. Additionally, there was interest in incorporating more variation in verdict wording and integrating features that assess source trustworthiness and track previously debunked claims. These suggestions are being actively addressed in the tool's development.

The Transparency Service for AI Model Cards workshop revealed that model cards are often too technical for non-expert users. Participants suggested developing simplified "light" versions with layered, hyperlinked content that allows users to access deeper information as needed. They confirmed the importance of including key elements in model cards, such as performance metrics, training data descriptions, and ethical considerations, while also recommending indicators for dataset diversity and model limitations. Visual enhancements proposed included color-coded sections, ranked lists, and intuitive data visualizations to improve usability, especially in the performance evaluation section. Additionally, participants highlighted the necessity for training resources, including a "zero card" to help users effectively read and interpret model

cards, and suggested improvements for better workflow integration and a dual-level interface for both experts and non-experts.

With regards to the Personal Companion, its claim detection and matching component received positive feedback, although there was some skepticism regarding the reliability of logical fallacy detection due to linguistic nuances. Some participants noted that the fallacy detection feature could be beneficial for coaching junior journalists. Suggestions for user interface improvements were implemented in the second iteration of mock-ups, which included functionality enhancements such as the ability to store debunked claims in a shared database, search functionality for claims, and the presentation of results with reliability indicators. Extensive discussions with domain partner DW helped tailor the tool to user needs, identifying role-specific profiles for various user groups, including journalism students, editors, and fact-checkers. For instance, it was suggested that functionality for matching claims to reputable sources should be predefined for students to avoid overwhelming them, while allowing journalists and fact-checkers the flexibility to edit their trusted sources.

4.6 Definition of scenarios of use and user stories

The development process for scenarios of use followed an iterative approach starting from use case scenarios outlined in the project proposal. These broad scenarios were subsequently revised based on insights gained from exploratory interviews conducted with potential users. After this revision, mock-ups were created to visualize the scenarios, which allowed for a clearer understanding of how users would interact with the service. The scenarios of use underwent further refinement based on feedback gathered during co-creation workshops, ensuring that they accurately reflected user needs and expectations.⁴

Similarly, user stories were developed through an iterative process to define specific capabilities that a service shall offer, including the purpose for which users would utilize those capabilities. Typically, user stories follow a format such as "As a [type of user], I want [some goal] so that [some reason]" (Cohn, 2004). The initial user stories drew from experiences gained in previous projects (such as vera.ai, TITAN, and AI4TRUST), ideas presented in the project proposal, and insights obtained from exploratory interviews. These user stories were then revised based on additional feedback from co-creation workshops, allowing for a more comprehensive understanding of user requirements.

This iterative approach to developing scenarios of use and user stories, along with the creation and revision of mock-ups, ensures that the final product is closely aligned with user needs and expectations. It facilitates continuous refinement based on user feedback and helps bridge the gap between user requirements and technical implementation.

⁴ To clarify terminology, it is important to distinguish between scenarios of use and technical use cases. Scenarios of use refer to user-centred use cases that describe what target users would use a given service for in a specific application context. They provide a narrative context for how the service might be utilized in real-world situations (Rosson & Carroll, 2022). In contrast, technical use cases specify in detail what needs to be implemented for a given user story to be realized, focusing on the technical requirements and steps necessary to enable specific functionalities (Cockburn, 2000).

Part II: Results per service

5 Disinformation Detection in the Next-Generation Social Media

5.1 Service Description

5.1.1 Scope and Objectives

The growth of decentralized social networks, such as Mastodon and other platforms within the Fediverse, has created both opportunities and challenges. These networks offer users more control over their data and communities but also introduce complexities in moderation, content management, and transparency. To address these challenges, this service combines two essential tools: the **Fediverse Network Moderation Tool** and the **Mastodon Instances Inspector**.

The **Fediverse Network Moderation Tool** supports administrators in maintaining safe, inclusive spaces by streamlining moderation tasks and fostering cross-instance collaboration. Meanwhile, the **Mastodon Instances Inspector** provides valuable insights into the structure, activity, and health of Mastodon instances, enabling informed decision-making and better network management.

Together, these tools empower administrators, moderators, and users to navigate the decentralized social web effectively, ensuring a balance between freedom and accountability.

Fediverse Network Moderation Tool

This tool is designed to help the discovery of potentially harmful servers in the Fediverse. By utilizing the publicly available moderation strategies enforced by individual servers in the Fediverse, this tool categorizes banned servers across specific keywords (e.g., spam, harassment, racism, etc), and tracks how frequently a server has been banned under each category.

The tool enables users to discover harmful servers through these indexed keywords, which are typically manually assigned by the server administrators from the various existing instances. With a user-friendly and intuitive interface, both experts and non-experts can easily access a of servers linked to specific keywords of interest and import them in the Mastodon admin panel.

One of the tool's main advantages is its ability to provide a regularly updated list of servers, keeping the pace with the evolving nature of decentralized social media.

Mastodon Instances Inspector

This service is designed to assist journalists and fact-checkers in verifying the credibility of information on Mastodon instances (i.e., servers).

It will be using three sources of data: i) the networks messages reshare across different instances; that will be continuously collected ii) the network of bans performed by instances' moderators, that will be reconstructed periodically based on public data, tracking the instances that are banned, those who are banning them and the reason for those bans; and iii) the content of messages shared within by users of a specific instance, that will be downloaded on request, analysed, and then deleted.

The Mastodon Instance Inspector will offer tools that allow users to inspect specific instances, assess their reliability based on moderation actions and analyse them based on the reasons for which they were flagged or banned. The service tracks how information flows through the Mastodon network. Fact-checkers can access a database of banned instances (banning network) to validate claims. The service also generates visual reports, simplifying the interpretation of trends in information dissemination.

5.1.2 Supporting critical thinking

Far from providing simplistic labels, the service presents rich contextual information that invites interpretation and inquiry. By surfacing how instances relate to one another, who they block, and how information flows between them, the tool encourages users to question sources, consider motivations behind bans or reshares, and understand the dynamics that shape online discourse.

By allowing users to compare instances — in terms of moderation behaviour, content patterns, bot activity, and trust signals — the service encourages analytical thinking. Users are prompted to examine how different communities govern themselves, what kind of content they prioritize or allow, and why some instances are more isolated or interconnected than others.

5.2 Target Users and Expected Benefits

5.2.1 Fediverse Network Moderation Tool

Target users	Expected Benefits
Primary target users	
Media Professionals, Media organisations, Journalists	- Keep media professional's servers safe in the Fediverse - Prevent harmful interactions from other servers and services in the Fediverse - Avoid spreading of harmful content
Fact-checkers	Discover servers in the Fediverse matching "harmful" keywords (e.g., misinformation, harassment, etc.)
Server Admins & Content Moderators	- Proactively block harmful instances before they interact with their community. - Use keyword-based filters to apply moderation aligned with their server's values. - Collaborate with other admins by aligning moderation strategies.
Government organizations and policy-makers	- Investigate the evolution of the platform for shaping new policies (e.g., DSA, GDPR) - Monitor how moderation and banning practices evolve across instances, identifying gaps in self-regulation or risks of abuse. - Identify patterns of misinformation spread on a decentralized network — crucial for national security and election integrity oversight. - Support initiatives around transparency, accountability, and trust in digital communication spaces.
Main stakeholders	
Research community	Researchers can build on the moderation strategies collected by the tool for research/academic purposes

5.2.2 Mastodon Instances Inspector

Target users	Expected Benefits
Primary target users	
Journalists	- Improved accuracy and reliability in reporting by verifying the quality of information circulating on Mastodon instances quickly. - Increase efficiency by reducing time spent on manual research

D3.1 User-centred requirements definition

Fact-checkers	• Access to comprehensive and up-to-date data sources supporting the validation of claims by evaluating the quality of the instance of which a user is a member. • Verification of the instance's source based on the aggregation of the decision of the bans of all Mastodon's moderators to enhance the robustness of fact-checking efforts. • Discover servers in the Fediverse matching "harmful" keywords (e.g., misinformation, harassment, etc..)
Brands & Marketing Teams	• Discover the most suited instances for promoting products or services • Discover high-trust, high-engagement instances suitable for brand-safe outreach. • Avoid association with harmful or controversial content.
PR & Communication Teams, Community Managers	• Evaluate the trustworthiness and moderation practices of Mastodon instances before joining, ensuring brand safety and alignment with organizational values. • Avoid reputational risks by identifying and steering clear of poorly moderated or controversial instances. • Map strategic entry points by targeting instances with high legitimacy, diverse audiences, and positive reputations. • Use content trends, dominant languages, and engagement patterns to tailor messaging and communication strategies for specific communities.
Main stakeholders	
Research community	• Build on the inspection capabilities offered by the tool for research/academic purposes

5.3 Scenarios of Use

5.3.1 Fediverse Network Moderation Tool

Scenario 1: Network Moderation and Media Professionals

Tom is the Social Media Manager of a Media Company aiming to join the Fediverse for trying a new social experience and improve users' reach. After creating a new server and joining the Fediverse, Tom discovers that, by default, any server can interact with the newly created one, including harmful ones. The potential risk is that such a new server will be flooded by harmful interactions or content. Tom wants to prevent the new server from receiving bad interactions, but he does not know how to discover bad servers to filter them out using the Mastodon Admin Console. Tom, hence, uses the network moderation tool to obtain a curated list of servers to moderate.

Tom accesses the tool and declares a list of keywords that describe the type of servers to be avoided (e.g., racism, harassment, spam, etc). The tool finds the list of servers typically moderated for the reason declared by Tom and prepares a Mastodon-compatible file to enforce moderation. Tom downloads the files and uploads it in the Mastodon Admin Console, thus effectively enforcing the moderation towards the "bad" servers identified by the tool upon keyword declaration from Tom. This process ensures the server curated by Tom will be safer in both content and interactions.

Scenario 2: Instance Investigation and Fact-Checkers

Chloe works for a Fact-Checking Company that investigates the development and growth of new or alternative social media platforms. These include decentralized social media platforms which are more difficult to analyze than centralized ones (e.g., X), as they consist of different servers and not a single point-of-access. For example, Chloe is tracking fake news about US Elections being viral throughout the Web, and she wants to discover whether this is flowing across the Fediverse, and which are potential servers that are involved in spreading it to analyze and fact-check them. However, given the scattering of servers throughout the Fediverse (and hence the Web), Chloe finds very difficult to discover them and track the spreading of such news.

Therefore, Chloe uses the network moderation tool to get a list of servers matching her criteria (i.e., typically banned for fake-news and misinformation spreading) aiming to figure out potential spreaders in the Fediverse. The tool finds the list of servers flagged to be involved in fake-news and misinformation spreading, and returns them to Chloe, also providing details about the number of times such servers have been flagged for these keywords. Chloe uses this list to investigate and fact-check such instances, thus providing new insights into how such kind of content spreads in alternative/new social media platforms like Mastodon.

Scenario 3: Server Administrator Streamlining Moderation Practices Across Instances

Alex is the administrator of a mid-sized Mastodon instance focused on digital art and online creativity. Over time, the instance has grown significantly, attracting a diverse user base. However, with this growth, Alex begins noticing an increase in spam accounts, unwanted interactions, and harmful content originating from certain external instances. Manually moderating these interactions has become time-consuming and inconsistent. Alex needs a scalable solution to maintain a healthy environment for users without having to investigate and block harmful instances one by one.

Alex turns to the Fediverse Network Moderation Tool for help. Through its intuitive dashboard, Alex explores keyword-based categories like “spam,” “hate speech,” and “NSFW content.” The tool presents a curated list of instances that have been frequently banned by other servers for these reasons. Alex can review the metadata, see how often each server has been flagged, and read community-provided tags and descriptions. After selecting the relevant categories, Alex exports a Mastodon-compatible moderation file and uploads it directly into the server’s Admin Console. In minutes, the instance is updated with a robust, community-informed blocklist. Alex sets up automated updates from the tool to regularly refresh the blocklist, ensuring the server stays protected against newly identified harmful actors.

Scenario 4: Policy Maker Monitoring Cross-Border Influence and Moderation Gaps

Sophie is a digital policy advisor for a European regulatory body focused on media integrity and online harms. Ahead of an upcoming election, her team is tasked with identifying potential threats in the online information space — including disinformation campaigns and foreign influence operations spreading through decentralized platforms like Mastodon. Using the Fediverse Network Moderation tool, Sophie examines moderation and banning patterns across a broad range of instances. She identifies clusters of servers with weak moderation practices that are repeatedly flagged for hate speech, election misinformation, and harassment.

By analysing the ban network, she uncovers a subset of instances that have been banned by numerous well-established servers — a red flag that indicates reputational issues. Sophie compiles her findings into a briefing for lawmakers and digital oversight committees, highlighting moderation blind spots and recommending stronger cross-jurisdictional cooperation. Her report not only strengthens public policy around decentralized networks but also supports transparency and accountability efforts, ensuring that self-governing systems remain secure and democratic.

5.3.2 *Mastodon Instances Inspector*

Scenario 1: Investigative Journalist Tackling Misinformation Using Network Analysis

Emily, an investigative journalist, is working on a story about a disinformation campaign surrounding a political issue she is interested in which is circulating on Mastodon. Concerned about the credibility of the sources, she uses the Instance Inspector tool to investigate a particular instance that appears central to the narrative. The tool reveals the instance has been flagged for spreading misinformation and hate speech, prompting Emily to avoid using content from this source in her article.

To gain deeper insights, Emily uses the tool's AI-driven content analysis to assess recent posts from the flagged instance. The analysis confirms that many posts have been marked as unreliable. Confident that avoiding this instance is the right decision, Emily shifts her focus to understanding how the disinformation has spread across the broader network. Using the Network Analysis tool, she gets the details on the moderation history of these instances, showing that several have been banned for misinformation and other offences and similarly the moderation behaviour of those instances who are responsible for banning them.

By analysing these trends, Emily gains a clear picture of moderation dynamics between different parts of the network. Finally, she generates visual reports that highlight the key patterns in the spread of disinformation and compiles her findings into a detailed article. Emily's report not only exposes the disinformation campaign but also explains why she excluded certain unreliable sources from her work. Her use of AI tools and network analysis ensures that her report is both credible and data-driven, maintaining the trust of her readers.

Scenario 2: Fact-Checker Using Tools to Track Banned Sources and Disinformation

Alessandro is a fact-checker at a media organization, responsible for ensuring that journalists use credible sources. Recently, a surge of online misinformation surrounding a trending news story catches his attention. To investigate, Alessandro uses a set of tools to track banned sources, identify questionable content, and trace the spread of false information.

He begins by checking a database of banned instances, looking up a particular source mentioned in a recent article. The database shows this source has been banned multiple times for spreading misinformation. Confident in his findings, Alessandro marks the instance as unreliable and moves on to other sources involved in the misinformation campaign. Next, he tracks the spread of misleading content across the network. He identifies a few key sources that have been frequently reposting false information. After analysing the content, he compiles a report on these questionable sources and shares it with his team to help them avoid using unreliable information.

Finally, he traces the origin of the misinformation. By mapping out how the story spread across the network, he identifies the original source and several clusters that amplified the false claims. He creates a report summarizing how the disinformation moved through the network, helping his organization understand the scope of the problem. By the end of the day, Alessandro's work ensures that his team is well-informed about which sources to avoid, helping maintain the accuracy and credibility of their reporting.

Scenario 3: Brand Ensuring Safe Engagement and Audience Alignment

Julia is the Digital Marketing Lead at a sustainable fashion brand preparing to expand into the Fediverse as part of an eco-conscious campaign. She's intrigued by Mastodon's decentralized communities, which offer an alternative to traditional social media — but she knows that choosing the right instance to engage with is critical for maintaining brand integrity and reaching the right audience. To guide her decision, Julia uses the Mastodon Instances Inspector. She starts by researching several large and mid-sized instances that focus on environmental issues, digital culture, and activism. The tool provides her with a clear picture of each instance's moderation behavior, reputation in the network, and history of bans received or issued — helping her steer clear of spaces that might expose her brand to harmful or controversial associations.

Julia can also explore the most spoken languages within each instance, ensuring her messaging aligns with the dominant user base. She also analyzes topic trends and recurring hashtags, helping her understand what kinds of content resonate most with the community — whether it's circular fashion, climate activism, or DIY culture. Through on-demand content snapshots, she also identifies the type of claims and messages

being shared — ensuring that her team avoids platforms where misinformation is prevalent or values don't align. Using these insights, Julia narrows down a selection of instances that are not only trustworthy but also demographically and thematically aligned with her campaign. With this data-backed approach, Julia launches her brand's Fediverse presence on the right footing: aligned with community values, tuned into cultural nuances, and positioned for meaningful engagement.

Scenario 4: PR Manager Ensuring Brand Safety Before Entering the Fediverse

Darius is the head of digital communications for a major nonprofit organization. His team is exploring Mastodon as a new communication channel to reach niche, values-aligned audiences without relying on traditional platforms. But first, they need to assess the reputational and moderation risks of different instances before selecting one to join. Using the Mastodon Instance Inspector, Darius filters potential hosts by moderation quality, language, audience demographics, and reputation within the Fediverse. He quickly rules out several instances that have been flagged for inadequate content moderation or spreading misinformation — factors that could damage the organization's public image.

The tool helps Darius identify a handful of high-reputation instances with strong moderation records and diverse, active communities. He also consults recent content trends, examining which topics are gaining traction in these spaces, and tailors the organization's launch messaging accordingly. With clear data in hand, Darius presents a confident strategy to his leadership team. By choosing the right instance and avoiding problematic servers, the organization ensures brand safety, maximizes reach, and sets the tone for a meaningful presence in the decentralized web. The insights from the tools become an ongoing resource for planning, engagement, and risk management.

5.4 User Stories

5.4.1 Fediverse Network Moderation Tool

User Story #1	Title: Network Moderation
Type of user	Media Professionals, Media Organisations, Journalists
Action	As a Media Professional or Media Organisation (and Journalists), I want to join the Fediverse with a new server and keep it safe from harmful interactions, so that I can establish my server as a trustworthy instance.
List of Use Cases	- Set up network moderation on the Mastodon Admin console by discovering potentially harmful instances in the Fediverse based on certain criteria (i.e., keywords), and prevent receiving harmful interactions from these servers - Keep the network moderation strategy of the maintained server up-to-date by periodically discovering new instances that are flagged as harmful during time
User Story #2	Title: Server Discovery
Type of user	Fact-Checkers
Action	As a Fact-Checker, I am interested in discovering the servers in the Fediverse typically banned for certain reasons, so that I can analyze them and fact-check their content.
List of Use Cases :	Discover servers that are relevant to certain keyword-based criteria (e.g., misinformation, fake-news) to be able to analyze and fact-check their content Discover potentially harmful instances in the Fediverse that can start spreading phenomena with misinformative content or other harmful contents, to be able to monitor them or inspect them
User Story #3	Title: Community Safety
Type of user	Server Admin
Action	As a Server Admin, I want to protect my community by importing curated moderation rules and blocklists, so that I can streamline moderation without reviewing each instance manually.

D3.1 User-centred requirements definition

List of Use Cases : • Identify and block instances frequently flagged for specific harmful behaviors using keyword filters • Export Mastodon-compatible moderation files to enforce safety rules efficiently • Stay aligned with moderation best practices by using data aggregated from the broader Fediverse	
User Story #4	Policy Analysis
Type of user	Policy makers
Action	As a Policy Maker, I want to analyze moderation trends across the Fediverse to identify risks, gaps, or influence patterns, so that I can inform digital policy and platform governance.
List of Use Cases : • Analyze instances frequently flagged for disinformation, hate speech, or harassment • Identify clusters with weak moderation as part of risk assessment or compliance oversight	

5.4.2 Mastodon Instances Inspector

User Story #1	Title: Instance inspector
Type of user	Journalist
Action	As a Journalist, I want to join the Fediverse as a user and be sure I am getting attached to a reliable instance
Use Cases : • Search for a specific instance and get the credibility status within the bans network. • Run an automated AI driven analysis of the content that circulated recently on the instance Provide a report on the instance to evaluate its reliability as a source, informed by the reasons for bans and from AI flags for potentially unreliable content.	
User Story #2	Title: Network analysis for dis/misinformation insights
Type of user	Journalist
Action	As a Journalist, I want to understand the federated context of Mastodon instances, generating detailed stats and reports from network analysis, so that I can gain insights into trends and patterns related to potential flows of mis/disinformation between instances
Use Cases : • Monitor sources through the repost and ban network to identify interesting patterns. • Use the network analysis to identify clusters of mis/disinformation sources. • Identify key instances in the dissemination of information, particularly those potentially spreading misinformation. • Analyse the movement of instances across different moderation reasons {such as Harassment, Hate-Speech, Misinformation and Spam} • Quantify the polarization based on the moderation activities of instances to understand if instances are being banned for similar or different reasons. • Generate visual reports (graphs, charts) highlighting statistical trends and patterns of diffusion of (mis/dis)information from specific instances informed by the reliability of the web domains of the news circulating. • Compare historical data to detect the evolution the network Export detailed stats and reports for further analysis or publication.	
User Story #3	Title: Access to the database of bans
Type of user	Fact-Checker
Action	As a Fact-Checker, I want to access directly the database of banned instances to ensure its reliability by directly verifying the accuracy of the bans produced by independent moderators of different instances.
Use Cases : • Search for relevant instances based on their credibility status within the banning network. • Access historical data on bans to evaluate the evolution of bans.	

D3.1 User-centred requirements definition

Highlight and flag discrepancies in the evaluation of bans by different banning instances Provide a summary of the bans produced or received by an instance, including confidence levels estimated by the agreement between different banning instances	
User Story #4	Title: Tracking questionable sources
Type of user	Fact-Checker
Action	As a Fact-Checker, I want to use the bans network to track and flag questionable sources, so that I can alert journalists and media professionals.
Use Cases : - Monitor sources through the repost and bans network to identify interesting patterns. - Run an automated AI driven analysis of the content that circulated recently on the instance Provide a report on the instance to evaluate its reliability as a source, informed by the reasons for bans and from AI flags for potentially unreliable content.	
User Story #5	Title: Tracing information origin and dissemination
Type of user	Fact-Checker
Action	As a Fact-Checker, I want to utilise the network analysis to trace the origin and dissemination of information, so that I can identify and address misinformation trends.
Use Cases : - Analyse the most spoken languages and trending content topics within each instance. - Use content trends to inform brand messaging tailored to specific communities.	
User Story #6	Title: Assessing most common topics and languages
Type of user	Marketing team
Action	As a Brand Strategist, I want to evaluate the credibility, user demographics, and trending topics of specific Mastodon instances, so that I can safely plan ethical and well-targeted communication campaigns.
Use Cases : - Search for a specific instance and get the credibility status within the bans network. - Use network analysis tools to map the flow of information from its origin across various sources. - Identify key instances in the dissemination of information, particularly those potentially spreading misinformation. - Generate visual reports (graphs, charts) highlighting statistical trends and patterns of diffusion of (mis/dis) information from specific instances informed by the reliability of the web domains of the news circulating.	
User Story #7	Title: Safe Mastodon adherence
Type of user	PR team
Action	As a PR Manager, I want to analyze the moderation behavior and content culture of Mastodon instances, so that I can avoid reputational risks and protect my organization's image.
Use Cases : - Search for specific instances and get trust summaries informed by ban data and AI signals. - Review the moderation consistency of each instance to avoid inconsistent or toxic environments. - Examine language use and popular discussion topics to ensure alignment with brand values. - Flag instances that show polarization or inconsistent moderation for exclusion from outreach plans.	

5.5 Early Requirements

#	Early high-level functional requirements	How will the requirement be addressed in the service design?
1	The system shall allow users to input keywords to identify and curate a list of servers to moderate	Users will be able to input custom keywords (e.g., "racism", "harassment",

D3.1 User-centred requirements definition

	based on specific harmful interactions (e.g., racism, harassment).	“misinformation”) to identify instances associated with specific harmful interactions.
2	The system should be able to regularly (e.g., daily) crawl the Fediverse for getting the most up-to-date network moderation information from existing servers, that will be used to build the lists	The system will regularly crawl publicly available metadata from Mastodon and other ActivityPub-compatible instances to maintain an up-to-date view of moderation practices and inter-instance bans.
3	The system shall include a database for storing and updating the metadata about network moderation publicly available from each instance	The system will include a database for managing the metadata about network moderation
4	The system shall generate and provide Mastodon-compatible files for enforcing moderation on identified servers.	The system will allow the users to export block lists that adhere to Mastodon format.
5	The system shall display historical data on server flags, including the number of times servers have been flagged for specific reasons.	The systems will provide the users with historical data on banned instances.
6	The system shall implement AI-driven content analysis to assess and report on recent posts from instances flagged for misinformation or hate speech.	Leveraging AI-driven models, the system will analyze recent user-generated content (on-demand) from flagged instances, detecting and reporting indicators of misinformation, hate speech, or machine-generated content.
7	The system shall enable users to search for specific instances and retrieve their credibility status within the banning network.	Users will be able to search for any instance and retrieve a trust summary based on its position in the banning network, content quality metrics, and moderation history.
8	The system shall provide tools for tracking the spread of content across the Fediverse, identifying key sources involved in misinformation dissemination.	The system will track how content — particularly misinformation or high-engagement posts — spreads across the network, assessing the infodemic risk.
9	The system shall generate visual reports that highlight trends and patterns in information spread and moderation dynamics between instances.	Interactive visualizations will highlight moderation dynamics, ban networks, and content spread.
10	The system shall allow users to flag discrepancies in ban evaluations across different instances and provide summaries of bans produced or received by an instance.	Users will be able to highlight inconsistencies in ban behaviors (e.g., divergent treatment of similar instances) and view summaries of moderation actions performed <i>by</i> and <i>against</i> each instance.
11	The system shall implement network analysis features to quantify moderation participation and polarization among instances based on their banning history.	Built-in graph analysis tools will quantify moderation engagement and centrality enabling deeper insights into governance patterns across the decentralized network.

#	Early high-level non-functional requirements	How will the requirement be addressed in the service design?
1	The application shall be accessible and useable for non-technical audiences	The user interface will be designed with accessibility in mind, offering clear visual cues, simplified workflows, and glossary support to ensure usability by non-technical audiences.
2	The application shall be well documented and described	Documentation shall be provided, including usage instructions, methodology explanations, and glossary terms, to support onboarding, transparency, and reproducibility.

D3.1 User-centred requirements definition

3	The application shall be responsive, with low computation time	Backend services will be optimized for performance to ensure minimal latency and fast response times.
4	The system shall provide a user-friendly interface that facilitates easy navigation and interpretation of data for non-technical audiences.	The interface shall prioritize intuitive navigation and legibility, using clean visual design, contextual tooltips, and guided workflows to simplify interaction with complex data.
5	The system shall ensure scalability to accommodate a growing number of instances and users without performance degradation.	The system architecture will support increasing traffic, data volume, and user load.

5.6 Ethical Requirements

#	Ethical requirements	How will the requirement be addressed in service design?
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system.	Our service will provide aggregated results and the user will not directly interact with the AI system.
HAO-02	The capabilities of the application shall be communicated clearly to the user.	Detailed documentation will outline what the system can and cannot do, including any limitations or assumptions in its analysis.
PDG-01	The application shall minimize the amount of data collected on users.	The system will not collect persistent user data. On-demand content analysis is triggered explicitly by the user and does not store user-identifying information.
TSA-03	The data used to build the application shall be up-to-date, of high quality, complete and representative of the deployment environment.	For on-demand analysis, the system fetches content in real time from publicly available posts. Data is processed immediately and discarded after use to ensure freshness, privacy, and relevance.
FAI-02	The application shall take the diversity and representativeness of the end user into account.	Interface design and content categories will reflect a range of languages, cultures, and moderation contexts. User testing with diverse stakeholder groups will inform design iterations to ensure inclusivity and accessibility.

5.7 Early Success Criteria

System Effectiveness benchmarking:

- The Mastodon Instances Inspector will reliably report details on the moderation history of banned instances.

User evaluation:

- Media professionals will be able to effectively discern potentially harmful Mastodon instances based on analyzing information flows through the Mastodon Network.
- Media professionals will be able to efficiently discover potentially harmful servers in the Fediverse based on specific keywords (e.g., spam, harassment, racism, etc)

- Fact-checkers will be able to efficiently access the services' database of banned instances to validate claims.
- The visual reports generated from the service will help Media professionals to efficiently share the interpretation of trends in information dissemination.

5.8 Early Mock-ups of the Service

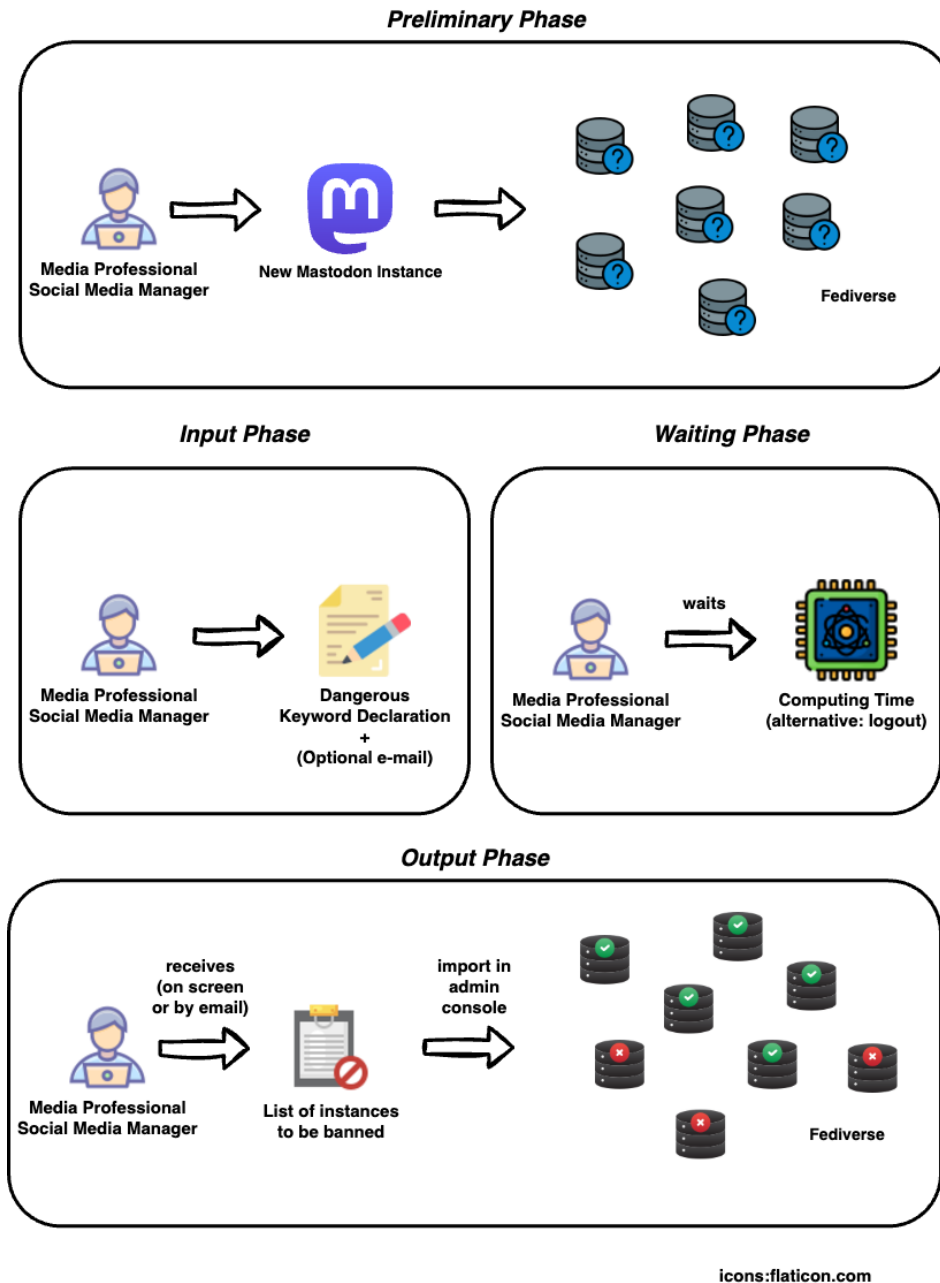


Figure 13: Visual diagram of the main functionalities offered by the Fediverse Network Moderation Tool. After creating a new instance and joining the Fediverse, admins (e.g., Media Professional Social Media Manager) can use the functionalities offered by this tool to determine the list of instances to be moderated (e.g., based on certain keywords) via the Mastodon admin console, in order to avoid receiving harmful interactions from them.

6 Trustability and credibility assessment

6.1 Service Description

6.1.1 Scope and objectives

This service will consist of a set of tools (AI-based methods and models utilising deep learning, NLP, image matching, etc.) able to automatically detect potential credibility signals that can be later combined by various assessment algorithms into credibility indicators to decide on the trustability and credibility of content and sources. As input, media professionals will be able to provide either: 1) an existing news article/a social media post (e.g., to assess the trustability/credibility of a content/source they aim to write); or 2) a custom piece of content (e.g., to assess the trustability/credibility of their own work). As output, various forms of explainable trustability and credibility assessments will be provided. The service will cover credibility signals related to both content as well its source (either at a level of individual creators or at a higher level of social network groups/pages/channels or web domains).

6.1.2 Supporting critical thinking

This service stimulates critical thinking by equipping media professionals with detailed insights into the credibility and trustworthiness of the found content. Rather than simply labeling a piece of information as “true” or “false,” the system identifies and presents nuanced credibility signals such as linguistic patterns, visual inconsistencies, and source reliability indicators that users can evaluate and interpret. By analyzing these, users are encouraged to analyze evidence and consider multiple perspectives before making decisions or judgements.

Furthermore, the service allows users to test and reflect on the credibility of both external content and their own work. This dual perspective encourages self-assessment and critical thought, helping professionals recognize potential biases in their reporting and decision-making processes. The nature of this service fosters the activity of its users, by stimulating an iterative process of evaluation, interpretation, and revision. In this way, this service could be considered stimulating for deeper reasoning and more informed, critical engagement with the information landscape.

6.2 Target Users and Expected Benefits

Target users	Expected Benefits
Primary target users	
Fact-checkers / investigative journalists	- Better opportunities to check facts when comparing sources of information online - Easier to share verifiable facts - Increase the trustability of the pieces of news they diffuse (media diffusion)
Community managers	- Better knowledge about new posts/users in Social Networks - Fast method to have preliminary trustability metric - Automated identification of misleading profiles or contents
Digital content creators of outreach / dissemination / divulgation of present events	- Better chances to find reliable sources for coverage amidst noise during current events - Improvement of the quality of their content - Effectiveness on timely identification of reliable sources
Social media managers / Content moderators	- Better knowledge about new posts/users in Social Networks - Find reliable information for their own posts
Documentary filmmakers	Increased trustability of the sources used

Main stakeholders	
Online news portals run by media organisations	- Assurance of the credibility of their own productions - Improvement of information gathering and, therefore, of their news production chain
Independent online news portals	Assurance of the credibility of their own productions
Secondary benefitters	
Governments	- More control on hot topics spread online - Assessment of public opinions - Assessment of dangerous disinformation

6.3 Scenarios of Use

Scenario of Use #1

Jenna, a media professional who is preparing an inform about the candidates to presidency of her country, has found important information. Nevertheless, it comes from an unfamiliar online source, and she knows the importance verifying its reliability in this age of misinformation. In order to get to the truth, she opens the newly implemented trustability assessment service.

She gathers all references related to the source. First, she copies the URL of the article, looks if the source has been linked and shared in popular social media platforms, and retrieves their hottest links. She saves these multimedia elements as well.

The trustability assessment service's interface guides Jenna through the process step-by-step. She inserts the URLs into the corresponding fields, specifies the types of media included in the article, and adds the other required inputs. Immediately, she clicks the starting button to initiate the assessment.

The output report indicates that the source had a history of spreading unverified information, the high share count was driven by bots, and the multimedia elements were traced back to unrelated events.

Relieved to have avoided a potentially flawed story, Jenna knows now to discard this source. The trustability assessment service thus ensures that only accurate information will reach the public.

Scenario of Use #2

Alex is scrolling through his social network feed when he comes across a post from an unknown user that catches his attention. The post makes some intriguing claims, and Alex is curious about its credibility. Suspicious about the authenticity of the information, Alex decides to use the social network's Credibility Assessment Tool.

Alex inputs the post and starts the tool. After a few seconds, it processes the post and displays a simple credibility metric using a colour bar. The colour bar ranges from red (low credibility) to green (high credibility). In this instance, the bar shows yellow, indicating moderate credibility.

Intrigued about the moderate rating, Alex wants to understand the reasoning behind it. He clicks on the detailed analysis option provided by the tool. The detailed analysis reveals various components considered in the credibility assessment: speaking about source reliability, the unknown source has a mixed history of factual reporting; when speaking of content verification, the claims made in the post are partially supported by reputable sources; and when speaking about the engagement and interaction, the post has been shared widely, but many shares come from questionable accounts.

Supported with this information, Alex decides to approach the post with caution, recognizing the need for further verification from trusted sources before sharing it.

6.4 User Stories

User Story #1	Title: Professional investigation of online sources of information
---------------	---

D3.1 User-centred requirements definition

Type of user	Media professionals
Action	As a media professional, I need to get an idea of how much I can trust a certain source of information online, so I can write a valuable piece of news based on it.
List of Use Cases	1.1: Identify the online source of information to perform credibility analysis 1.2: Input the link to the source into the service 1.3: Execute the tool and navigate through the charts and explanations of the results 1.4: Draw conclusions on the credibility of the sought information based on 1.3
User Story #2	Title: Checking social media
Type of user	Social Media User
Action	As an active social media user, I want to guarantee that my daily information and communication experience online is reliable, so that I can better understand what is happening in the world and how my acquaintances are doing so far.
List of Use Cases	2.1: Locate new social media post to verify trustability 2.2: Input the link to the post into the service 2.3: Execute the tool and navigate through the charts and explanations of the results
User Story #3	Title: Consulting history of searches
Type of user	Previous user of the service
Action	As a previous user of the Trustability and Credibility Assessment Tool who has my own personalized space within the service I want to check at the history of past searches, so that I can recall previous usages and compare them with new ones.
List of Use Cases	3.1: Open the UI of the service and log in with the own credentials and password 3.2: Click on the History tag and browse through the list of previous searches
User Story #4	Title: Consulting popular searches
Type of user	Previous user of the service
Action	As a previous user of the Trustability and Credibility Assessment Tool who has my own personalized space within the service, I want to check the most popular searches by all users, so that I can compare them to my history and discover new insights on information online.
List of Use Cases	4.1: Open the UI of the service and log in with the own credentials and password 4.2: Click on the Popular Searches tag 4.3: Browse through the list of popular searches

6.5 Early Requirements

#	Early high-level functional requirements	How will the requirement be addressed in the service design?
1	The service must be able to analyse both content and context to assess the amount of credibility.	Analysis combined from content shared and user history, interaction and metadata.
2	The service should allow the users to identify low credibility content.	An aggregated “final score” is provided to the user as a summary of all analysis performed. This statement provides a grade in which the content is stated as high credibility/low credibility
3	The credibility assessment tool must return an output form of labelled levels of credibility, ranging from low to high.	A range of labelled levels of credibility is provided, together with a “color based” indicator from red to green.

D3.1 User-centred requirements definition

4	The credibility assessment tool should return an aggregated score, with possibility to check individual results if desired	Yes, the initial aggregated score is provided and this can be inspected to see the weight and result of each individual analysis
5	The credibility assessment tool should return an explainable output	Explanations and clarifications are given through the User Interface
6 a)	The service, when oriented to the users, must show the outputs of the analysis	Outputs from analysis are offered to the user
6 b)	The service should allow the user to customise the dashboard to display only the results of the output modules they are interested in.	Possibility to hide sections provided to the users
7	The service could redirect to another service in AI-CODE which already performed another analysis	Interconnections with other tools will offer the option to access the proper tool (link) when accessible
8	The service should provide optional brief explanations to the credibility metrics used	Explanations and clarifications are given through the User Interface
9 a)	User should log to the platform, and have a history of the searches	History of searches available
9 b)	The history of the searches should display the previous searches of the logged user	History of searches available
9 c)	The history of the searches could allow comparison between the metrics previously retrieved	The history of searches provides a clear view on how the previous analysis was scored in terms of trustability
10	The platform should have a history of the searches that can be updated with new ones	History of searches updates when new ones are created
11	The service must be accessible by UI and REST API	User interface and REST API publicly available

#	Early high-level non-functional requirements	How will the requirement be addressed in the service design?
1	The service must work in English language	The service works in English
2	The service must be available on Mastodon and Bluesky	The service is available for Mastodon and Bluesky
3	The service showing could show the trustability assessment for each component and then show transparently how the overall score is calculated	The aggregated score is shown together with the fraction contributed from each individual analysis to the final score
4 a)	The service could have a “Help me” section	“Help me” section available
4 b)	The “Help me” section could include definitions on what each tool does for a better understanding for the user	“Help me” section offering definitions and clarifications on tool analysis and usage
5	The application shall be well documented and described	Documentation on usage/workflow and maintenance are offered to users
6	The application shall be accessible and useable for non-technical audiences	The service aims to grant a clear final score with the required output analysis. If users have more expertise or want to get a more in depth understanding, they can navigate deeper in the UI.
7	The application shall be easy to use	The service only requires a post URL from Bluesky or Mastodon to be used.
8	The application shall be easy to learn	The service requires little input from the user and provides graphical and textual support to make it easier to understand the application and to learn how to use it.
9	The application shall be used in a browser	User Interface publicly available through the browser.

10	The application shall provide differentiated support for beginner and expert users as two main types of user cohorts within the target user group of the application.	The system is designed to be intuitive and straightforward. By simply entering the link to a post, users can extract all the information that was gathered to generate the credibility score. Each section includes a brief description explaining how the corresponding module works, along with an option to expand the view for more detailed, technical insights. For more advanced users the platform also allows for deeper exploration of the underlying data and methods, enabling more comprehensive analysis and interpretation.
----	---	--

6.6 Ethical Requirements

#	Ethical requirements	How will the requirement be addressed in the service design?
HAO-02	The capabilities of the application shall be communicated clearly to the user.	The service will provide concise descriptions and explaining what each functionality does and what the results stand for.
FAI-01	The application shall have procedures in place that avoid unfair bias that apply to data, design, and output.	Models will be retrained when significant bias is perceived, and if it is not possible, models will be changed accordingly
PDG-01	The application shall minimize the amount of data collected on users.	The service has been designed so only textual information, and their associated metadata is read, and only the results of their analysis are stored.
TRA-01	The system input and output of the application shall be traceable through documentation.	The design and operation of the services, including numerical data about its training and parameters, alongside the nature of its operation, will be thoroughly documented.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system.	A disclaimer will be made available clarifying the service is a support and should not be taken as a final decision.

6.7 Early Success Criteria

System Effectiveness benchmarking

- The credibility assessment tool will reliably evaluate the trustability of content and apply verifiable scores from low to high.

User evaluation

- Media professionals will be satisfied with the service's reliable evaluation of the trust and credibility of inputs, including Mastodon and Bluesky content.
- Media professionals will find the outputs created by the service comprehensible.
- Media professionals will find the service an effective way to build a more trustworthy social media feed.

6.8 Early Mock-ups of the Service

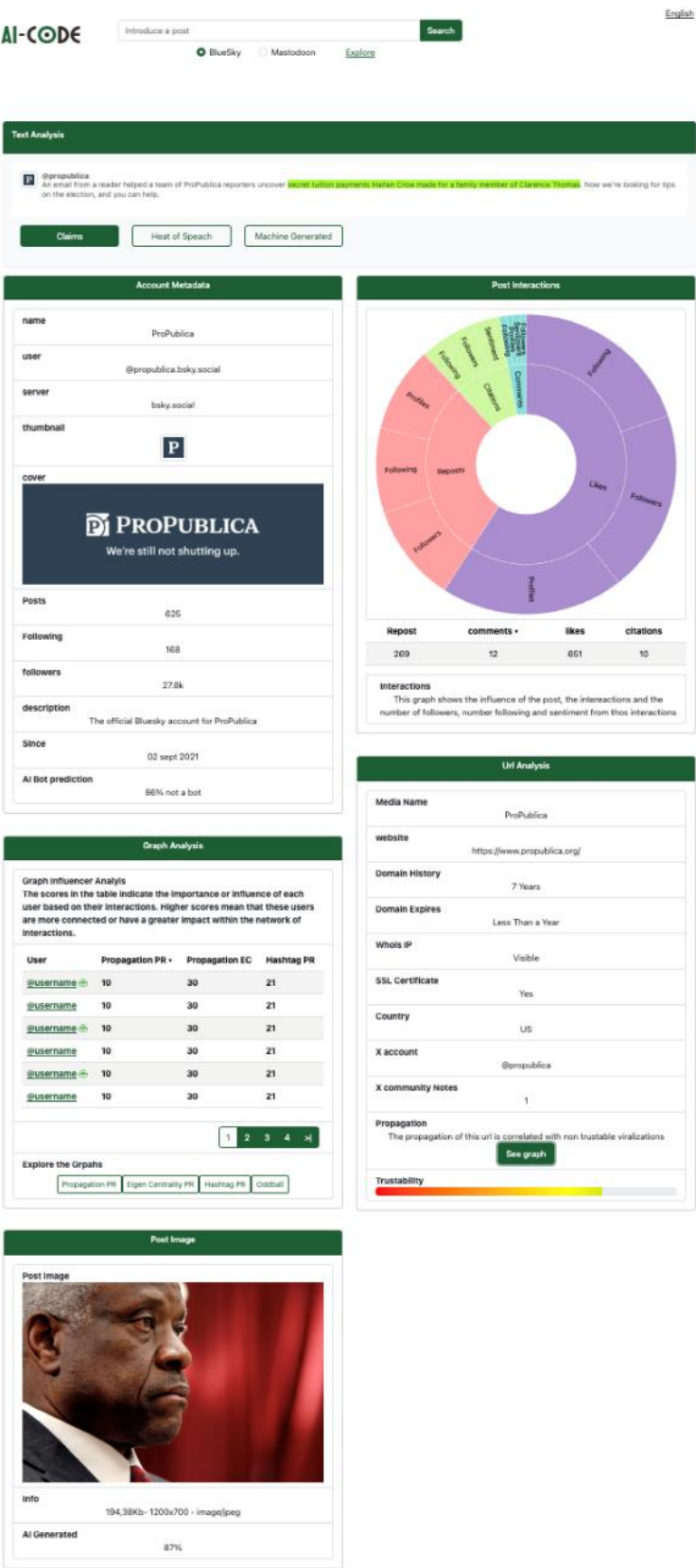


Figure 14: Main page of the service

D3.1 User-centred requirements definition

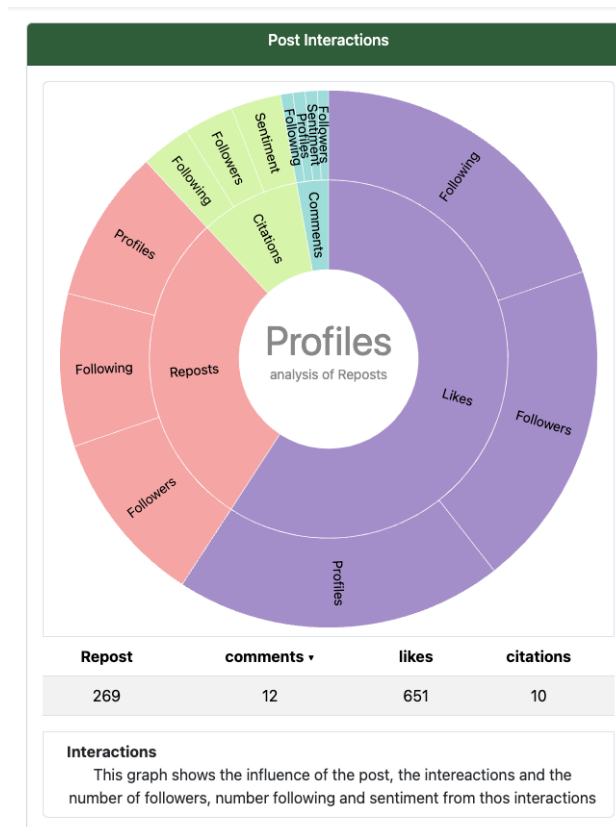


Figure 15: Sunburst diagram of analysis

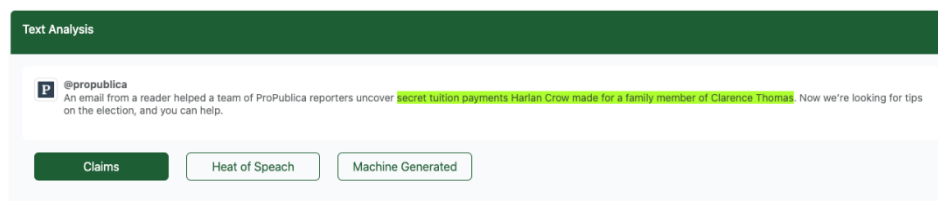


Figure 16: Window displaying results of Text Analysis

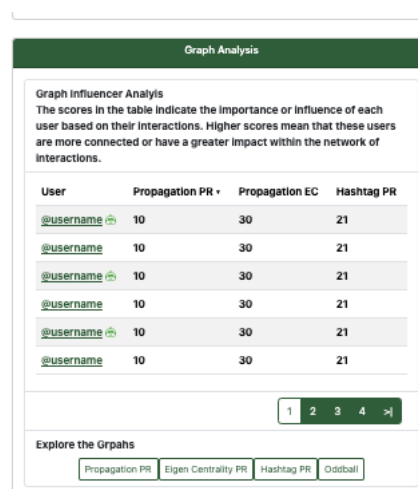


Figure 17: Window displaying results of Graph Analysis

D3.1 User-centred requirements definition

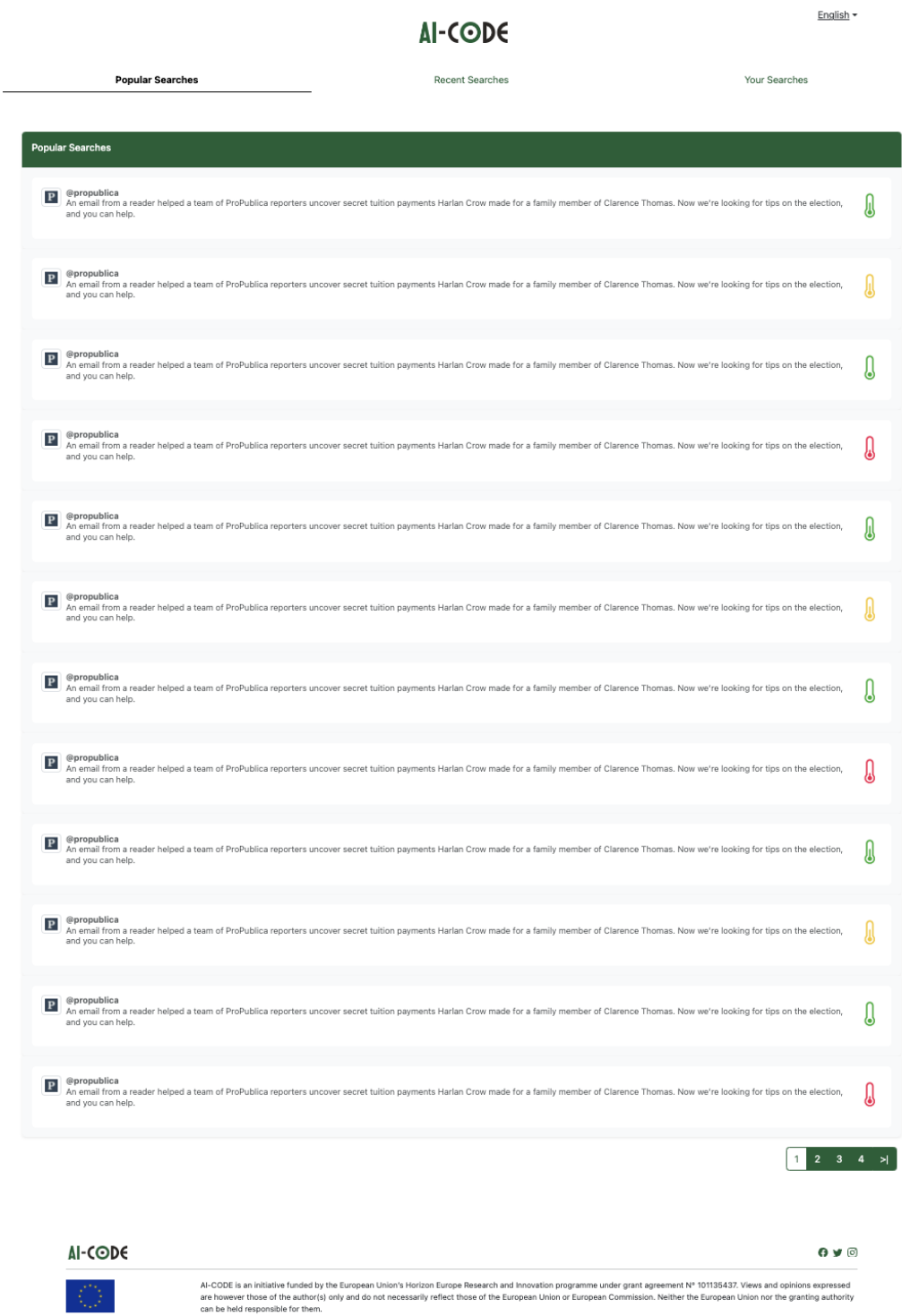


Figure 18: Early version of the window displaying history of searches

7 Media Asset Annotation and Management

7.1 Service Description

7.1.1 Scope and objectives

The Media Asset Annotation and Management platform (MAAM) is an advanced service, initially developed in the context of the H2020 MediaVerse project, which forms the initial technological basis for advanced and trusted multimedia asset management and retrieval in AI-CODE. It is designed to annotate and manage media content using AI models. It applies labels to uploaded media based on various categories such as objects, actions, persons, news-related tags, and moderation tags. The platform supports powerful retrieval capabilities through text-based and visual similarity searches, enabling users to efficiently filter and find relevant media content, e.g. via its Near-Duplicate Detection (NDD) functionality, which identifies variations of the same multimedia asset that may be visually similar but not identical. The system also integrates several on-demand AI models for synthetic content detection, manipulation detection, deepfakes identification, and other forms of media integrity analysis.

Planned developments in AI-CODE aim to further enhance the platform's effectiveness in combating disinformation by incorporating more advanced forgery detection, media manipulation analysis tools, and improved versions of existing detectors. In addition, explainability will be strengthened through the integration of model cards (from T6.2), offering greater transparency around how AI models make decisions. User profiles will also be enriched to capture users' expertise in disinformation analysis, making the platform more relevant for experts in the field. Tailored to serve journalists, fact-checkers, social scientists, researchers, and political science researchers, the enhanced MAAM platform will provide these professionals with robust tools to manage, analyze, and verify collections of media assets—particularly in the context of fighting disinformation and media manipulation.

7.1.2 Supporting critical thinking

MAAM promotes critical thinking by combining multiple AI detection models with transparency features such as confidence scores and model cards to provide users with a deeper understanding of generative algorithms and their risks. By analysing synthetic or manipulated multimedia content through different detection approaches, users gain insights into how effective generative algorithms are at creating synthetic media, how different manipulation techniques work (e.g., altering voices and/or faces in videos), and how AI detectors operate. Model cards further explain the strengths, weaknesses, and intended use of each detection model, preventing blind reliance on AI-generated outputs. This approach encourages users to critically interpret detection results, compare different models, and assess the real risk posed by generative media, helping them to make informed decisions rather than passively accepting automated detections.

Additionally, MAAM's collaboration features enhance critical thinking by facilitating interactions between expert and non-expert users. Non-expert users, such as journalists or fact-checkers, can gain a clearer understanding of the disinformation landscape by engaging with AI-generated predictions and expert evaluations. At the same time, expert users benefit from exposure to how non-experts interpret and understand detection results, revealing gaps in explainability and public understanding of AI tools. This interaction challenges experts to reassess the limitations of state-of-the-art detection approaches and redefine their methodologies to bridge the gap between technical accuracy and real-world application.

7.2 Target Users and Expected Benefits

Target users	Expected Benefits
Primary target users	
Fact-checkers	• Misinformation Mitigation: Strengthen efforts to combat misinformation by identifying fake media before it spreads.

D3.1 User-centred requirements definition

	• Deepfake Identification: Utilise specialised AI models to detect synthetic media, including deepfakes. • Forgery Detection: Leverage cutting-edge tools to identify subtle image and video manipulations. • Comprehensive Reports: Generate detailed verification reports to support fact-checking conclusions. • Collaboration and Feedback: Enhance accuracy and efficiency by collaborating with other users, sharing insights, and exploiting feedback to cross-verify findings and improve detection results.
Journalists	• Media Verification and Content Integrity: Easily understand and utilize reports from fact-checkers that verify the authenticity of images and videos, including detecting deepfakes, forgeries, and manipulated media. This ensures that journalists can confidently report on issues like misinformation and political propaganda, with clear and explainable insights guiding their content decisions. • Content Moderation: Automatically tag offensive, disturbing, or sensitive content, helping to manage and filter content, and protect viewers from potentially traumatizing content. • Assessing User Expertise: Evaluate and consider the expertise of users providing feedback on potential disinformation, such as fake or synthetic content.
Social scientists and Researchers in Media Studies / Journalism	• Efficient Content Organisation: Easily annotate and categorize media assets using AI annotation. Social scientists and researchers can have multimedia collections automatically tagged by MAAM with relevant metadata, enabling faster sorting and analysis. This reduces manual work when dealing with large collections of media assets. • Enhanced Media Search: Simplify the search in large media collections with advanced query and retrieval capabilities. • Specialized Analysis Tools: Leverage advanced AI models for specific types of analysis, such as synthetic content detection, deepfake identification, and manipulation detection. These are valuable for social and political scientists studying misinformation, media manipulation, or the impact of synthetic media on public discourse and political processes.
Media asset managers	Efficient Content Organisation: Enables media asset managers to structure and manage large volumes of media using metadata and AI-generated annotations. Enhanced Media Search and Retrieval: Enables rapid access to specific asset subsets through filtering, tagging, and semantic search capabilities. Efficient Content Moderation: Supports automatic detection and flagging of inappropriate or manipulated content.
Policy Makers and Regulators	• Misinformation Tracking and Fact-based communication: MAAM facilitates fact-based communication strategies by identifying disinformation trends that require public intervention and action. • Empirical Evidence for Policy Making: MAAM provides insights that help policymakers understand the scale and nature of synthetic content threats.
Students	• Digital Literacy and Analytical Skills: Students gain hands-on experience with AI-powered media analysis tools, enhancing their ability to critically assess digital content and interpret AI-generated results. • Collaboration and Research Opportunities: Students can participate in fact-checking projects, work with experts, and use MAAM to document findings and contribute to disinformation research.
Main stakeholders	
Media Organisations	• Risk Mitigation: MAAM helps media organisations detect and prevent the use of fake or manipulated media, reducing the risk of disseminating false information.

	• Improved Operational Efficiency: With automatic annotation and content management features, MAAM streamlines the process of organising, retrieving, and verifying media assets, leading to more efficient workflows and reduced time-to-publish.
Policy Makers	• Enhanced Public Trust: By supporting the adoption of tools like MAAM, policy makers can help increase public trust in media by ensuring that published content is verified and free from manipulation.

7.3 Scenarios of Use

Scenario of use #1: Monitoring and Verifying Media Content During Elections

Richard, the Founder and Lead Researcher of an NGO focused on monitoring election integrity and fighting disinformation, is preparing for the upcoming elections in a country of the global South. To ensure that the elections are free from manipulated media, Richard turns to the MAAM tool, which allows him to manage and verify media content related to the election. He starts by setting up a new workspace within MAAM, focusing on the specific regions and content he needs to analyse. Since MAAM doesn't collect content directly from social media or other external sources, Richard gathers relevant images and videos himself and uses the tool's batch upload feature to input them into the system. As the election period intensifies, Richard relies on MAAM's AI-powered detection tools to analyse these uploaded files for signs of manipulation, such as deepfake videos, altered or fully synthetic images. He carefully reviews the flagged content, using his expertise to assess its potential impact. After completing his analysis, Richard uses MAAM to compile a clear and detailed report, which he shares with his team and other important stakeholders. This report highlights verified cases of synthetic media and offers insights that can help protect the integrity of the election, reflecting Richard's dedication to maintaining fair and transparent electoral processes.

Scenario of use #2: Large-Scale Analysis of Deepfake Risks in Videos with Public Figures

Marko, a post-doctoral researcher focused on media manipulation and misinformation, is working with a large dataset of videos featuring public figures. His dataset includes both original videos and suspected deepfakes. Marko's goal is to analyse this large-scale dataset to identify which public figures are most at risk of being targeted by deepfake technology and to understand how these deepfakes might be embedded into other content.

To begin his analysis, Marko uses MAAM's near-duplicate detection feature (NDD), specifically designed to operate at the video segment level. This allows him to identify parts of videos that have been embedded into other videos, based on both visual and audio matching. For example, MAAM helps Marko detect cases where a segment of a speech given by a public figure has been extracted and inserted into a different context or video, possibly to mislead viewers. Once the near-duplicate segments are identified, Marko applies MAAM's deepfake detection models to these segments. These models analyse the flagged video segments for any signs of manipulation, such as unnatural facial movements, inconsistencies in lighting, or anomalies in voice patterns that are often indicators of deepfakes. With the deepfake detection results in hand, Marko takes his analysis a step further by leveraging MAAM's person identification feature. He cross-references the deepfake detection results with the system's ability to recognize the public figures featured in the videos. By doing this, Marko can determine which individuals are most frequently targeted for manipulation, and which have the highest risk scores according to the deepfake detection models. Finally, Marko uses MAAM to compile his findings into a comprehensive report. This report includes a detailed breakdown of the identified near-duplicate segments, the deepfake risks associated with each, and an analysis of which public figures are most at risk.

Scenario of use #3: Detecting Video Manipulations with Near Duplicate Detection

Jonathan, an investigative journalist, is working on a story involving a high-profile government official. While researching, he comes across a short video circulating on social media that shows the official making a controversial statement. The content of the video seems unusual, and Jonathan suspects that the video might have been manipulated to mislead the public. Jonathan's primary goal is to verify a specific manipulation by comparing a suspicious video to an original, authentic one. His investigation is about exposing the misleading

D3.1 User-centred requirements definition

narrative created by altering a public official's statement.

To verify his suspicions, Jonathan turns to the MAAM tool. He uses the Near Duplicate Detection (NDD) feature, which allows him to search his organisation's extensive media collection for the original version of the video. Within moments, MAAM identifies a matching video in his collection that closely resembles the circulated one. Jonathan carefully compares the two videos and quickly notices that while the visuals are nearly identical, the audio in the circulated video has been altered. With this discovery, Jonathan is able to confirm that the video has been manipulated by splicing in altered audio to create a misleading narrative. Armed with this evidence, he prepares to expose the manipulation in his report, highlighting the importance of verifying media content in an era of widespread disinformation.

Scenario of use #4: Identifying Out-of-Context Media with Visual Similarity and Reverse Image Search

Kate, an independent journalist, is participating in a collaborative workspace within the MAAM tool, where other journalists and fact-checkers upload their findings related to ongoing disinformation campaigns. One day, a new image was uploaded on social media, showing a disabled man with a caption claiming that his condition was caused by a recent vaccine. The claim is alarming and has the potential to spread widely, so Kate decides to investigate further.

To verify the authenticity and context of the image, Kate uses MAAM's Visual Similarity feature. This tool allows her to search for visually similar images within MAAM's public workspaces and across her own collections. The feature quickly identifies that the same image has been used in various other stories, each with a different context. Intrigued, Kate decides to dig deeper. She employs the reverse image search functionality within MAAM, and through this search, Kate traces the image back to its original source: a news article published years ago, where the image was used in a completely unrelated context, having nothing to do with vaccines or health-related issues. With this information, Kate can conclusively demonstrate that the image is being misused to spread false information. She documents her findings and shares them within the MAAM workspace, helping other users understand the true origin of the image and preventing the spread of the misleading claim.

Scenario of use #5: Determine whether a video is a deepfake by collaborating with an expert.

John, a journalist working at a media organisation, receives a video that appears suspicious and could potentially be a deepfake, with multiple voices raising concerns about its authenticity. With the urgency to ensure the credibility of the content before using it, John uploads the video into the MAAM tool and initiates a deepfake detection analysis. The MAAM tool generates a report, but the results show a high degree of uncertainty, with conflicting indicators that make it difficult for John to confidently interpret whether the video is a deepfake. Recognizing his limited expertise in AI and machine learning, John decides to collaborate with a colleague within the MAAM who specialises in deepfake algorithms.

He reaches out to Sarah, an expert in the field, and shares the video and the MAAM-generated report with her through the platform's collaborative tools. Sarah reviews the report and cross-examines the evidence from several deepfake detection algorithms provided by MAAM but also references to results from other external tools. With her expertise, she carefully analyses the nuanced details that John initially found unclear. After a thorough examination, Sarah concludes that while the video is not a deepfake, it has undergone significant post-processing, which might have caused the initial red flags. She explains the specific types of processing detected and their likely impact on the video's authenticity. John, now armed with a clear understanding thanks to Sarah's insights, decides to report on the video with accurate information, highlighting that although it is not a deepfake, it has been altered in other ways.

Scenario of Use #6: Repository of Problematic Multimedia Content for Verification

An international media watchdog organisation leverages MAAM to create a comprehensive archive of disinformation media, including deepfakes, fully synthetic, and manipulated images and videos. The team, composed of journalists, media professionals, and fact-checkers, regularly uploads known problematic media assets into this centralised repository within MAAM.

To enhance the accuracy and detail of the repository, they use MAAM's advanced detection algorithms to analyse each asset. These tools provide additional visual and audio clues, such as identifying subtle visual manipulations or detecting discrepancies in audio tracks. This detailed analysis is stored alongside the original

D3.1 User-centred requirements definition

media, creating a thorough documentation of each disinformation case. This curated repository becomes an essential resource for the organisation. When new suspicious media arises, team members can upload it to MAAM and compare it against the archived collection. MAAM's retrieval capabilities efficiently identify similarities between the new content and the previously stored cases, helping the team quickly determine if the new media is likely manipulated or related to known disinformation use cases. By utilising MAAM, the organisation not only builds a robust repository of disinformation cases but also equips itself with a powerful tool for verifying new media content, enhancing their efforts to combat misinformation and protect public trust.

Scenario of Use #7: Geolocation Estimation of Images to Detect Image Misrepresentation

Ayesha, a political science researcher from abroad, investigates viral images claimed to show protests in a city she's unfamiliar with. Suspecting misinformation, she uses MAAM's geolocation estimation feature to verify the images. The platform analyses visual and metadata cues, comparing them with known landmarks and environmental data, helping Ayesha determine if the images are misrepresented. Since Ayesha doesn't know the city well, the tool is crucial for identifying discrepancies. She collaborates with local researchers through MAAM, sharing geolocation data and producing a report that debunks the false narrative.

7.4 User Stories

User Story #1	Title: Identifying and Analysing Manipulated Video Content Using Near Duplicate Detection and Deepfake Analysis
Type of user	Fact-checker
Action	As a fact-checker , I want to use the Near Duplicate Detection feature to find and compare similar videos, so that I can identify and expose any manipulations or alterations in video content to ensure accurate and truthful reporting.
List of Use Cases	• #1.1: Search for near duplicate videos and video segments based on visual and aural similarity. • #1.2: Compare visual similarity of videos at segment level. • #1.3: Identify video differences at audio level. • #1.4: Generate deepfake reports for the retrieved near duplicate videos.
User Story #2	Detecting and Reporting Manipulated Media During Elections
Type of user	Fact-checker
Action	As a lead researcher at a fact-checking organisation , I want to use AI-powered detection tools to scan uploaded media for signs of manipulation and compile detailed reports on verified synthetic media, so that I can identify disinformation and share my findings with my team and stakeholders to protect the integrity of the election .
List of Use Cases	• #2.1: Batch Import of assets into the MAAM platform. • #2.2: Select specific AI models to analyse uploaded assets. • #2.3: Utilize AI models to detect manipulations, forgeries, and fully synthetic content in images and videos. • #2.4: Create detailed reports with visual clues and manipulation scores.
User Story #3	Investigating and Debunking Out-of-Context Media Claims
Type of user	Journalist
Action	As an independent journalist , I want to use the Visual Similarity and Reverse Image Search features in MAAM to verify the authenticity and context of an image, so that I can identify and expose any misleading use of the image and prevent the spread of false information.
List of Use Cases	• #3.1: Visual similarity search for images, across personal and public workspaces.

D3.1 User-centred requirements definition

• #3.2: Reverse Image Search to external databases to trace images back to their original sources. • #3.3: Document findings and clues for out-of-context cases. • #3.4: Share findings and clues for out-of-context cases	
User Story #4	Collaborating to Verify Potential Deepfake Content
Type of user	Journalist
Action	As a journalist , I want to collaborate with a deepfake expert within the MAAM workspace to analyse and interpret the results of a deepfake detection report, so that I can accurately determine the authenticity of the video and ensure the credibility of my reporting.
List of Use Cases	
• #4.1: Analyse videos with multiple manipulation and deepfake detection methods. • #4.2: Find with an Expert. • #4.3: Collaborate with an Expert • #4.4: Create reports with visual clues for manipulation and deepfake detection.	
User Story #5	Media Asset Management
Type of user	Social Scientist
Action	As a social scientist , I want to use MAAM as a repository of media assets for disinformation research, so that I can store, organize, and analyse large volumes of media content , facilitating collaboration with other researchers and enabling efficient tracking, verification, and analysis of disinformation patterns and trends.
List of Use Cases	
• #5.1: Centralised content storage for all media assets, such as videos, images, audio files, and documents. • #5.2: Automated asset annotation and metadata generation for media assets organization. • #5.3: Collaboration within and across teams.	

7.5 Early Requirements

#	Early high-level functional requirements	How will the requirement be addressed in the service design?
1	The system shall allow users to import large batches of media files, including images and videos, through various methods such as direct uploads, media link imports, or integrating content from supported platforms for processing and analysis.	The system will provide real-time progress updates on large-scale batch uploads using a dynamic status tracker. Notifications will be triggered upon completion, keeping users informed without requiring manual checks. The system will handle large-scale imports efficiently using background job processing of Spring. Tasks will be queued and executed in parallel to optimize performance. Users will be able to import media via direct file uploads and URL-based imports with CSV files.
2	The system shall integrate multiple AI models to detect and localize image manipulations, including fully synthetic images and partially manipulated content, and provide detailed reports with detection scores. ⁵	MAAM follows a service-oriented architecture, where AI annotation / detection models are deployed as independent services that communicate with MAAM via APIs. This modular approach allows for easy integration of additional models, ensuring scalability and adaptability to new techniques. Image detection models follow a standardized

⁵ The MAAM service already integrates several detectors for identifying synthetic and manipulated images. These will be further expanded to include state-of-the-art generative and detection models.

D3.1 User-centred requirements definition

		request/response structure, enabling seamless integration of new models without significant effort, ensuring scalability and adaptability throughout the duration of the project.
3	The system shall detect video manipulations, particularly deepfakes, by applying several advanced AI algorithms. It will analyse videos for signs of identity swapping, facial reenactment, and other manipulations, offering detailed reports with probability scores and localized detection of manipulated segments. ⁶	Similar to image models, video detection models follow a standardized request/response structure, enabling effortless integration of new models throughout the project.
4	The system shall provide a Near Duplicate Detection functionality to identify both visually and aurally similar videos within the user's collection and public collections, with the capability to detect similarities at different segments of the video. ⁷	Integrate a state-of-the-art audio similarity model in the NDD service. Redesign the UI to support visual matching at different levels of the video.
5	The system shall provide an intuitive user interface (UI) that allows users to compare media at the segment level, highlighting subtle visual or audio manipulations. The system shall enable users to navigate between different segments of the media, offering side-by-side comparisons and visual markers to indicate areas of potential alteration.	The UI design will provide side-by-side comparisons with visual markers, allowing users to navigate and analyze detected similarities efficiently. Best practices in UI/UX design will be followed to ensure an intuitive, user-friendly interface.
6	The system shall enable users to search for visually similar images across multiple collections within the MAAM workspace and shared public collections, rather than being limited to a single collection. ⁸	The data model will be updated to support workspace access levels (private/public). Visual similarity retrieval from Elasticsearch will be updated to include workspace information in the ES queries.
7	The system shall integrate with external databases and image repositories to perform reverse image searches and trace the original source of media content.	When a user initiates a reverse search, the system will generate a search query using the image and redirect to Google's reverse image search endpoint. Additionally, integration with Yandex's reverse image search will be considered.
8	The system shall provide collaborative workspaces for sharing media, reports, and findings with colleagues. It will include access rights management, allowing users to assign permissions such as read-only, comment, or edit, ensuring secure and controlled collaboration.	Middleware in the Spring Boot backend will enforce access policies on every request for workspaces.
9	The system shall include expert's profile with detailed information such as their areas of specialization, past contributions, published reports, and experience in relevant fields.	To support this, the MAAM data model will be extended with user profile entities and relationships, while the UI will be enhanced to provide dedicated user profile pages designed for a seamless and user-friendly experience.
10	The system shall allow users to search for and connect with experts within the MAAM platform based on specific expertise (e.g., deepfake detection).	An expert directory will allow authenticated users to search for and browse profiles of other users registered in the MAAM platform based on expertise tags, roles (e.g., researcher, journalist, moderator),

⁶ MAAM already supports two deepfake detection models for face swap and face reenactment. This capability will be expanded by incorporating additional manipulation techniques, as well as the detection of fully synthetic videos.

⁷ MAAM already supports NDD but currently provides results at the level of entire videos. It will be improved to support segment-level detection and comparison as described in functional requirement #5.

⁸ Visual search is already supported. It will be further enhanced by leveraging advanced visual embeddings, allowing for more accurate and efficient image retrieval based on deep learning representations.

D3.1 User-centred requirements definition

		and activity/contributions. Several connection mechanisms (e.g., follow, message, notifications) will be added to initiate communication, invite experts to a workspace or request collaboration.
11	The system shall offer tools to create detailed reports that include visual clues, detection scores, and expert analysis, with support for exporting reports in various formats (e.g., PDF, Word) for sharing with stakeholders.	Different libraries will be used for generating PDFs and Word documents to support exporting. Integration with the transparency service will ensure that each report provides a high-level overview of the detection models, including their purpose, performance, and limitations.
12	The system shall include a mechanism to report the uncertainty or confidence level of AI detection results, helping users interpret ambiguous findings.	The data model will be updated to support user feedback by allowing users to submit reports or comments on the provided results.
13	The system shall provide a dedicated analysis page that allows users to upload and analyze a single media asset without the need to keep the asset within the MAAM system.	The MAAM UI will be redesigned to support drag-and-drop functionality for quick analysis, giving users the option to either keep the analyzed asset in the system or discard it after review.
14	The system shall provide model attribution for detecting synthetic images, including fully synthetic images and inpainting results. Detectors will identify the specific method or algorithm used, such as GAN variants, diffusion models like Stable Diffusion, or other techniques believed to have generated the manipulated or synthetic content.	The Synthetic Image Detection service integrated into MAAM will be expanded with an attribution model aimed at identifying the generative model behind a synthetic image. Due to the experimental nature and currently low performance of such detectors, this attribution model will operate as an additional, on-demand component.
15	The detection models integrated into MAAM shall provide a confidence score to quantify the likelihood of the detection.	Each detection or prediction will include a confidence score that quantifies the system's certainty. An additional evidence-level label (e.g., low, medium, strong) will be derived through calibration based on false positive rate.
16	The system shall integrate model cards for all AI models used within MAAM, including annotation models and disinformation detection models.	The transparency service will include an API that exposes model details in a structured format, allowing other components to retrieve information such as model purpose, architecture, training data, performance metrics, and known limitations.
17	The model cards shall provide detailed descriptions of each model, including information on the model's purpose, underlying algorithms, training data, performance metrics, limitations, and any relevant usage guidelines.	MAAM will follow the structure defined by the Transparency Service, and its UI will be redesigned to expose these fields for each integrated model.
18	The system shall implement role-based access control to ensure that only authorized users can access specific features, media content, and analysis results.	Spring Security will be configured to restrict access to endpoints based on user roles. A custom Spring middleware will be used to enforce workspace-level access checks based on the user's role within the targeted workspace.
19	The system shall support a wide range of media file formats (e.g., JPEG, PNG, MP4, MOV) to ensure compatibility with various user-provided content.	Various backend libraries and plugins (e.g., FFmpeg for video, Pillow for images) will be integrated to handle media pre-processing and post-processing as needed.

D3.1 User-centred requirements definition

#	Early high-level non-functional requirements	How will the requirement be addressed in the service design?
1	The system shall be able to scale to handle large volumes of media files, supporting simultaneous analysis for multiple users without significant degradation in performance.	Built on Spring, the system will leverage features like Spring Batch to enable asynchronous and parallel execution of processing tasks in a fault-tolerant manner, ensuring efficient workload distribution and high availability.
2	The system shall be built with a modular architecture to facilitate easy updates, maintenance, and the integration of new features or models.	The system already follows a service-oriented design, with MAAM serving as the main service that interacts with other services using multiple communication protocols, including REST APIs and gRPC. All new functionalities will be incorporated following the same approach, ensuring consistency, scalability, and easy integration.
3	The system shall provide a user-friendly interface that is easy to navigate, allowing users with varying levels of technical expertise to effectively use the tool.	The UI, implemented in Nuxt, will follow best practices in design and development, including clear navigation structures, accessible components, and responsive layouts to ensure ease of use for users with varying levels of technical expertise. To ensure this, user feedback on the current design and alternative mockups will be gathered through pilots and demos, and will be carefully considered during the design refinement process.
4	The system shall have an uptime of at least 99.9% over any given month, ensuring that it is reliably available for users during critical periods such as election monitoring.	The system will be deployed on a cloud-based Kubernetes cluster with load balancing to ensure high availability and fault tolerance. Monitoring and alerting will be handled using Prometheus and Grafana, providing real-time performance insights and automated alerts to minimize downtime.
5	MAAM shall provide differentiated support for beginner and expert users as two main types of user cohorts within the target user group of the application.	The UI will be redesigned to support a scaffolding approach for presenting AI detection results. Users — particularly non-experts — will initially see an overall assessment of the likelihood that an asset is synthetic or manipulated, with more detailed information available on demand. Additionally, specific features such as feedback on detection results and detailed assessment tools will be accessible only to expert users.

7.6 Ethical Requirements

#	Ethical requirements	How will the requirement be addressed in the service design?
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system	The interface will include clear labels and messages indicating when users are engaging with AI-generated results.
HAO-02	The capabilities of the application shall be communicated clearly to the user.	The system will provide concise descriptions and tooltips explaining what each AI feature does and its intended use. This will be based on the integration of the transparency service.
PDG-01	The application shall minimize the amount of data collected on users.	Only essential user data required for functionality will be collected.

D3.1 User-centred requirements definition

TRA-01	The system input and output of the application shall be traceable through documentation.	All operations will be logged and documented, including records of inputs, outputs, and the exact settings and parameters used during the application of each detection model.
TRA-02	The decisions of the application shall be understandable by the user (XAI) as far as reasonably achievable.	This will be supported by the integration of the transparency service and model cards. Intended use and training/evaluation data will help users interpret what to expect from each model prediction. The MAAM UI will reflect the different levels of detail provided by the transparency service.
FAI-01	The application shall have procedures in place that avoid unfair bias that apply to data, design, and output.	Bias mitigation strategies will be applied during the training and implementation of detection models, with fairness checks integrated into the evaluation process. The transparency service will document all relevant details, including dataset composition and model performance, to ensure visibility into potential sources of bias.
TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated	Accuracy metrics and known limitations of the integrated models will be displayed through the transparency service and model documentation.
HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case	The definition and implementation of these oversight modes will be guided by recommendations from the ethics advisory board, whose suggestions will help ensure responsible deployment and alignment with ethical best practices.

7.7 Early Success Criteria

System Effectiveness benchmarking

- The MAAM will achieve high detection rates for deep fakes and synthetic images, with minimal false positives and false negatives.

User evaluation

- Media professionals will report satisfaction with the usability and trustability of the MAAM system.
- Media professionals will be able to efficiently work with the MAAM system's tools for collaborative task assignment, progress tracking, and asset sharing.
- Media professionals will effectively gain a deeper understanding about how AI systems used in the MAAM tool work through the integration of standardized model cards in the MAAM.

7.8 Early Mock-ups of the Service

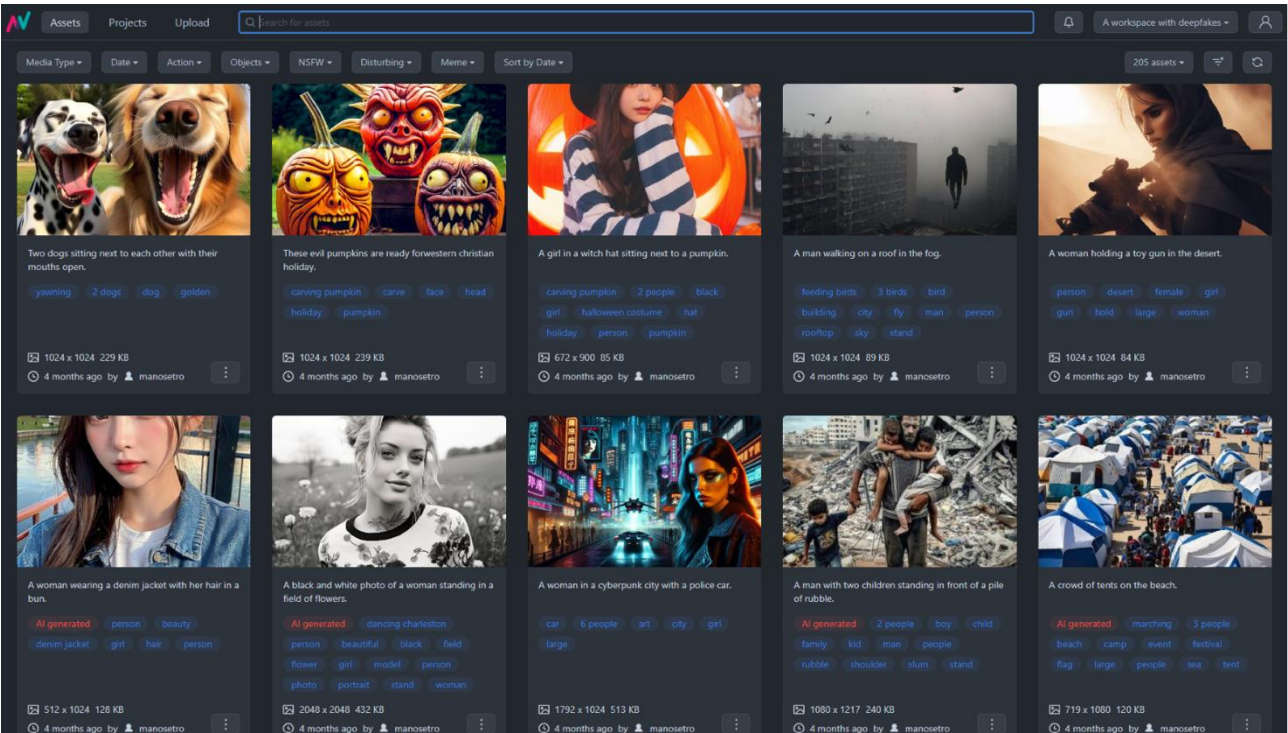


Figure 19: A screenshot from the MAAM platform showcasing an asset grid with a paginated view of media assets within a workspace.

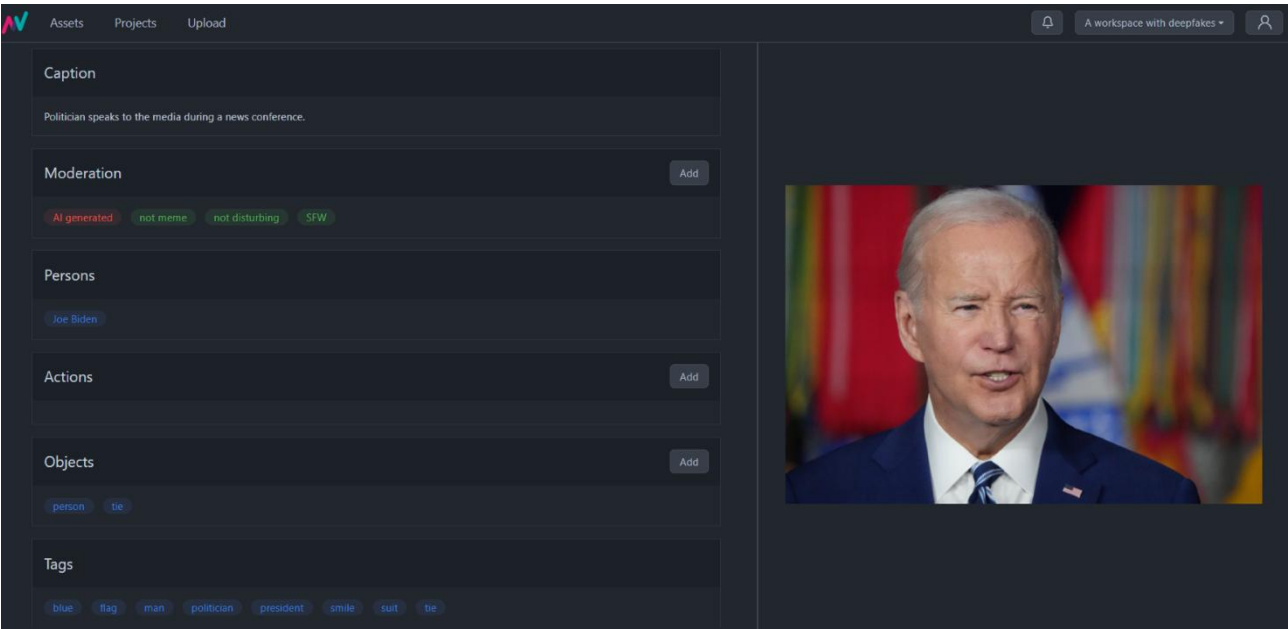


Figure 20: Asset Annotations for image uploaded and analysed by MAAM. Annotations include moderation tags, persons, actions, objects and news related tags.

D3.1 User-centred requirements definition

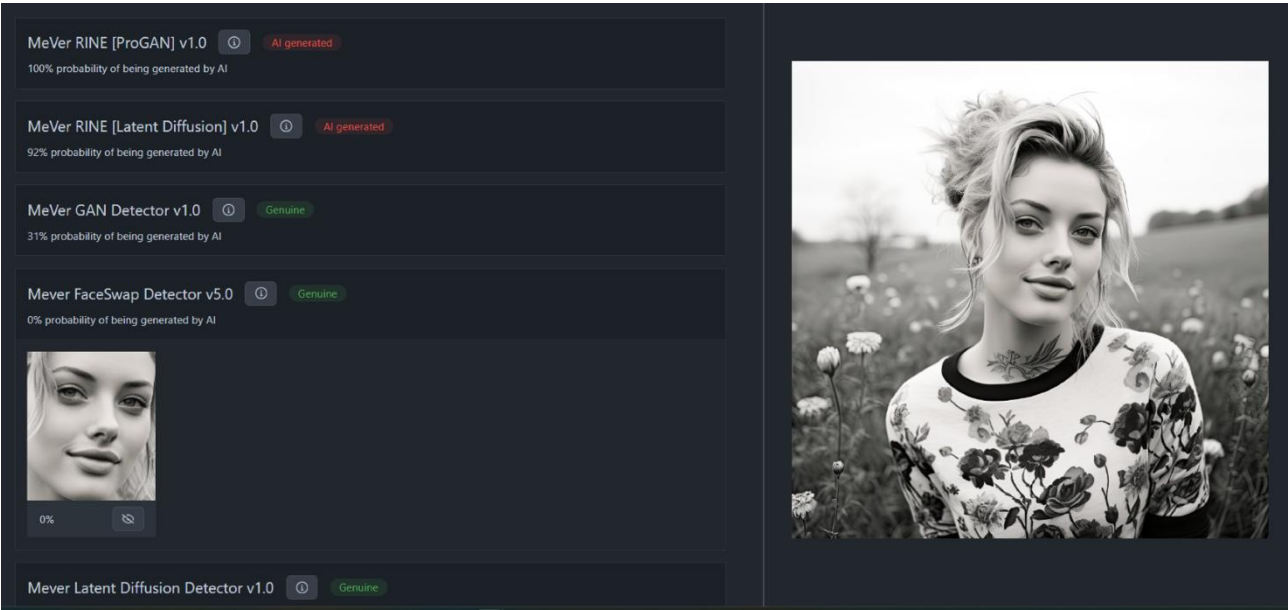


Figure 21: Reports on synthetic image detection generated by models already integrated into the MAAM system.

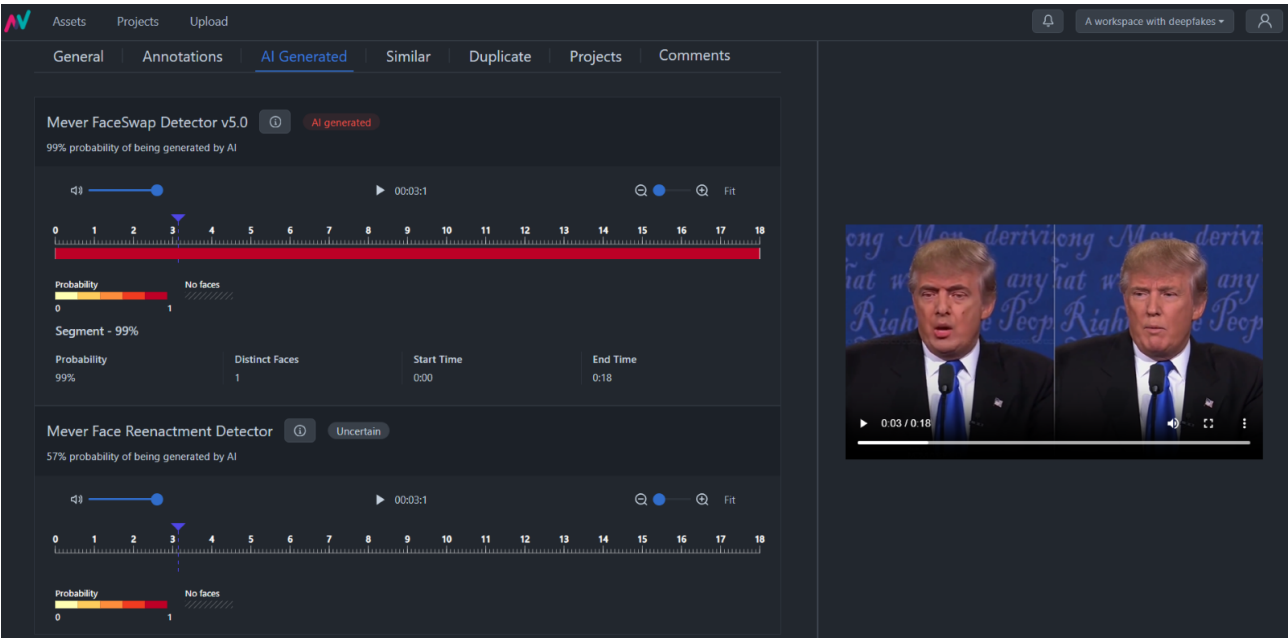


Figure 22: Deepfake Video Report with two detectors: FaceSwap and Face Reenactment

8 Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals

8.1 Service Description

8.1.1 *Scope and objectives*

The Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals helps media professionals to understand how generative AI can be used in trusted media production and countering disinformation. The two main components of this service, the **Interactive Coaching and Simulation Module (ICSM)** and the **Counter-narrative Recommendation Module (CNRM)**, are designed to work in tandem, providing media professionals with a holistic learning and continuous usage experience. Users can begin with the Interactive Coaching and Simulation Module to gain a foundational understanding of main principles of generative AI systems (LLMs) and prompt engineering. This knowledge can then be directly applied in the Counter-narrative Recommendation Module, where users can see how these principles inform the generation of effective responses to disinformation. Conversely, users of the Counter-narrative Recommendation Module can access the Interactive Coaching and Simulation Module to deepen their understanding of the underlying mechanisms that drive the Counter-narrative Recommendation Module. The Interactive Coaching and Simulation Module also integrates selected functionalities of text veracity verification from the Personal Companion for Understanding Disinformation service. This interconnected approach ensures that media professionals can develop both theoretical knowledge and practical skills in using generative AI responsibly and effectively in their work.

Interactive Coaching and Simulation Module

The Interactive Coaching and Simulation Module provides an interactive simulation and coaching tool that helps journalists and media professionals to develop a suitable mental model of how generative AI works (LLMs) and understand main principles of prompt construction in order to use generative AI in line with journalistic ethical values and standards, i.e. produce trustworthy content. This also includes interactive guidance for reliability analysis of LLM outputs. Users can interactively experiment with main LLM parameters, elements of prompt construction and types of user-system interaction. They can visually observe and analyse their impact on the type, quality and reliability of the output in a guided and self-exploration mode in pre-defined scenarios. They can apply these learnings in a real-time LLM application in the Counter-narrative Recommendation Module, thereby gaining practical insights into LLM behaviour and potential risks and counter-measures in content generation.

Counter-narrative Recommendation Module

The second Counter-narrative Recommendation Module of the service represents a development and adaptation of the AI4TRUST results. In particular, within the simulation environment, media professionals and end-users will be provided with a tool that can explain why a claim is or is not true and suggest—on demand—possible responses to disinformation posts within a simulated online conversation. This tool will allow the users to develop a practical understanding of the main topics and strategies of how to effectively respond to disinformation. The adaptation of this module requires the availability or production of an archive of debunk reports and is rooted on the paradigm that LLMs can also have a pro-social use.

8.1.2 *Supporting critical thinking*

Both modules of the service will play a crucial role in preparing and training media professionals to master the responsible and effective use of generative AI in producing high-quality, trusted information. Together, they will reinforce the development of a correct mental model of how generative AI systems such as ChatGPT work in a way that enables the development of necessary competences for the safe use of generative AI for creation of high-quality content for media professionals. This interactive simulation (enriched by supporting educational resources) will enable journalists, fact-checkers, and other media workers to use generative AI tools safely and responsibly. They will learn how to understand these tools, how to interact with them and

how they can anticipate future developments in this technology. Only then will they be able to exploit the full potential of such tools in order to fight the spread of disinformation and to (re-)gain trust in media offerings.

The Interactive Coaching and Simulation Module promotes critical thinking by enabling users to experiment with different prompting techniques and system parameters in a controlled environment. By directly observing how variations in prompts and parameters affect output quality, media professionals develop an analytical mindset that questions the reliability of generative AI systems rather than accepting outputs at face value. This experiential learning approach, combined with guided reliability checklists and verification tools, cultivates systematic assessment habits that transfer to real-world applications.

The Counter-narrative Recommendation Module further enhances critical evaluation skills by demonstrating the process of fact-checking against reliable sources. As users engage with the module to verify claims against authoritative documents, they practice identifying misleading information and constructing evidence-based counter-narratives. This practical application reinforces the importance of source verification and encourages media professionals to apply similar rigorous analytical approaches when encountering potentially false information in their daily work, ultimately strengthening their ability to distinguish between credible and misleading content.

8.2 Target Users and Expected Benefits

8.2.1 Interactive Coaching and Simulation Module

Target users	Expected Benefits
Primary target users	
Journalists	• Better understanding of main principles of how generative AI works (LLMs) and of the associated risks of using generative AI for better mental models and trustworthy use. • Better understanding of how to critically evaluate the output and the use of LLM systems in trusted media production. • Increased understanding of how the modality of use and interaction of LLMs influences the reliability of generated outputs. • Increased understanding of the types and sources of problems in use of LLMs for trustworthy content production (e.g. hallucinations and misinformation, LLM confirmation bias)
Journalism and media students	• Understanding main working principles of Generative AI for a better grasp of how to use it responsibly in journalism and media content workflows • Through guided experimentation with prompt construction, optimization and LLM output analysis, future media professionals learn to critically evaluate the reliability and quality of LLM outputs and AI generated content • Practicing LLM use and prompt construction prepares students for real-world applications in media production • Increased understanding of the types and sources of problems in use of LLMs for trustworthy content production (e.g. hallucinations and misinformation, LLM confirmation bias)
Journalism and media educators	• A ready-to-use tool they can use for Generative AI Literacy training and education of journalism and media students and professionals (as part of in-house professional training and development modules in media organisations) • Guide students in comparing outputs from different prompts or sources, helping them understand how variations influence outcomes
Documentary makers	• Understand how to use LLMs responsibly to produce trustworthy narratives aligned with ethical standards

D3.1 User-centred requirements definition

	Experimenting with prompts enables documentary makers to explore innovative ways to frame narratives while maintaining factual integrity
Media asset managers	- Better understanding of generative AI's risks and opportunities, enabling them to guide their teams effectively - Ensure that their teams produce reliable content that aligns with journalistic standards, enhancing organizational credibility
Main stakeholders	
Media organisations	Improved effectiveness in producing and publishing trustworthy content by better informed personnel

8.2.2 Counter-narrative Recommendation Module

Target users	Expected Benefits
Primary target users	
Journalists	- Support in producing answer to be used on social or other online media - Support in debunking false claim - Reducing time required to identify false claims - Easier access to information relevant to the claim from the documents
Fact-checkers	- Fast access to explanation of why a claim should be considered false or partially false (if that's the case) - Easier access to information relevant to the claim from the documents
Online community managers	- Support in producing answers to be used in online communities - Support in faster veracity assessment of large quantities of online claims - Reducing time required to identify false claims
Media educators	- Increase awareness on how different sources of information can lead to different counter-narratives - Compare different sources of information to investigate if the information provided leads to the same conclusion
Main stakeholders	
Media organisations	- Improved claim verification workflow - Faster engagement with general public for social correction activities

8.3 Scenarios of Use

8.3.1 Interactive Coaching and Simulation Module

Scenario of use #1

Rachel, a seasoned science journalist, is keen to explore the potential of Generative AI and how it can be used for trustworthy journalism. Her role involves investigating scientific developments, interpreting complex data, and communicating these findings to the public in an accessible manner. She is particularly interested in how AI can assist in ensuring the accuracy and reliability of the information she disseminates.

Rachel begins by accessing the Interactive Coaching & Simulation Module service, where she is met with an introduction screen, introducing her to the service. She then moves on to selecting the entry point question "How do systems like ChatGPT answer my questions?" This step is crucial for her to understand the main operational principles behind Large Language Models (LLMs) and their application in journalism. The coaching service launches a simulator containing a control panel, output panel and a guidance panel. Rachel reads the brief introduction given and then starts to explore how prompt construction works to learn what she needs to consider for her own practice. She tries out different settings for prompt design criteria and observes

D3.1 User-centred requirements definition

what changes in the corresponding generated prompt they cause. The criteria include choices such as, human-like/non-human interaction style, prompt specificity and context. After sending the prompt, the simulator outputs a response. Rachel continues to change and send prompts to compare the resulting LLM outputs, doing so with the help of the reliability checklist and the system's function that highlights differences in output. Rachel then toggles to system controls, where the simulator changes the prompt parameters to default system parameters. She experiments with changing the system prompt and temperature settings, observing how these adjustments affect the specificity and creativity of the outputs. Rachel realizes that higher temperature settings might introduce more variability, potentially increasing the chance of hallucination.

After getting an intuitive understanding for the effect of prompt construction choices and system parameters, Rachel moves on to the next part. While doing that, the system encourages her to re-send the first prompt she constructed again. Rachel compares the outputs displayed for the same prompt entered at the beginning and repeated at the end of her interaction sequence. Supported by the highlighting function, she notices differences between the two outputs, realising how system outputs for the same prompt can change over time. Rachel then moves to the self-exploration stage where she can apply her learnings in constructing her own prompts without guidance of the system highlighter and reliability checklist. She notices that she can, however, activate an output verification function, which highlights potentially incorrect claims in the text, making it easier to spot potential problems in the reliability of the output. An accompanying explanation tells her that these results are provided by another AI-CODE service that can verify the veracity of claims in a text. She makes a note to try out that service in more detail later (Personal Companion for Understanding Disinformation). The coaching service also points Rachel to try out the "Counter-narrative Recommendation Module" which she can use to verify the veracity of a specific claim and receive a narrative explaining why the claim can or can't be considered true, based on a proposed source. Impressed by the quality of the generated texts and the claim verification service, Rachel is curious to learn more about how LLMs can generate human-like texts and where the problems in inaccuracies ("hallucinations") come from.

Accordingly, she explores the module where she can observe how the mechanism of next-word prediction works and how she can query LLMs on multiple documents (the RAG method). She also learns how the ability of LLMs to perform specific tasks (e.g. summarize texts) depends on their training on a range of pre-defined examples (called supervised learning). She learns that LLM training includes a set of examples of common tasks users might ask them to perform, consisting of task instructions (prompts, questions) and corresponding example results (e.g. summary of a given text, answer to a question). She also explores how she can apply a similar principle in her prompts to improve the quality of generated results (few shot prompting).

Rachel concludes the session thinking she now has a good intuitive understanding of some of the main factors in prompt construction, system parameters and LLM operation that influence the quality and reliability of outputs from generative AI. She also has a pretty good idea why she can't rely on LLM outputs to be correct by default and how she can apply similar principles and techniques for checking the reliability of information from her journalistic training to verifying the reliability of LLM outputs. This allows her to better understand where the risks of using generative AI stem from. She notes that the related AI-CODE services for verifying the veracity of claims could help her to perform that task more efficiently and effectively. This should help her in ensuring that the content she generates with the help of LLMs in different journalistic tasks is trustworthy and complies with journalistic standards.

Scenario of use #2:

Sarah, a political correspondent for a national newspaper, is preparing a series of articles on the upcoming presidential election. She's heard concerns about potential bias in AI-generated content and wants to understand how LLM-generated content might influence public opinion.

Sarah logs into the Interactive Coaching and Simulator Module to explore how supervised learning affects output. She reads the explanation of the training process and examines examples of LLM tasks, noting similarities and differences between system outputs and examples of supervised training content. Once understanding the main principle, Sarah moves to the next screen, where she is introduced to the concept of few-shot learning. Drawing the link between her learnings before, Sarah understands on how examples decided by AI companies affect the output given for common tasks and how she can use few-shot prompting to affect the outputs herself. She experiments with example prompts given.

D3.1 User-centred requirements definition

As she compares the outputs, Sarah notices subtle differences in tone and emphasis depending on the examples provided. This leads her to the "self-exploration" page, where she crafts prompts mimicking various news outlets' styles. She's surprised to see how the AI adapts its language and framing based on these prompts. Concerned about the implications, Sarah delves into the section on how prompts construction affects LLM output. She tweaks the prompts multiple times using the parameter controls, documenting how the responses evolve with iteration. This exercise highlights the dynamic nature of LLMs and their potential to easily shift narratives.

To round out her investigation, Sarah explores the self-exploration module where she can test some political prompts on provided example documentation. She constructs prompts about controversial election topics and analyses the outputs for factual accuracy using the provided guidance checklist. This process reveals several instances of "hallucinations" – plausible-sounding but fabricated details about candidates and policies. Armed with this knowledge, Sarah feels better equipped to critically evaluate AI-generated political content and to educate her readers about the potential influences of LLMs on election coverage.

Scenario of use #3

Alex is an investigative journalist working on a sensitive story about corporate corruption. He's considering using an LLM to help analyse large volumes of leaked documents but is worried about maintaining the confidentiality of his sources.

Alex starts by exploring the system's data privacy section. He reads the overview of data collection practices for typical open source LLM models (as a summary of information from selected model cards of AI CODE services using them) and grows increasingly concerned about the potential risks of using an LLM to analyse confidential documents. To gain a deeper understanding, he accesses the "Transparency services for AI-model cards" feature, following the guidance to explore the training data behind popular LLMs. Alex learns that documents that are uploaded to an LLM are stored by the underlying system and concludes that he should not risk uploading the leaked documents.

Alex then starts to be concerned about his general interactions with LLMs. Wanting to ensure he can use the LLM effectively while protecting his data, Alex moves on to the prompt construction module. He practices crafting prompts that extract useful insights from his documents without including identifying details. He learns techniques for communicating with the LLM using more abstract or coded language, further reducing the risk of exposing sensitive information.

Now curious about the effects of using documents to retrieve information from LLMs, Alex investigates the module exploring how LLMs compile information from multiple documents. With his newfound prompt construction skills, Alex can more easily construct question prompts and send them to the system, with different numbers documents he has selected. He notices how sometimes the output has factual inaccuracies, especially when compared to the retrieved knowledge presented in the sidebar. He realised the added risk of using LLMs to answer questions, especially on sensitive topics, as reliability cannot be guaranteed.

Finally, Alex re-enforces his new prompt construction skills by trying out the 'guess the prompt' module. The system gives Alex a generated output analysing a piece of legislature from an economic perspective, with a hidden prompt that he must figure out. First, Alex analyses the text to spot key terms and narratives that could help him determine parts of the hidden prompt. Submitting a guess, the system outputs a response that Alex can compare to the original. This process continues until Alex feels that he has gotten the main points of similarity in the text and then he reveals the hidden prompt. Alex can identify from the previous module key parameters of the prompt that he has guessed correctly and other parts that he had missed. By analysing the outputs, he develops a strategy for using the LLM to assist his investigation while maintaining strict confidentiality.

Confident in his ability to leverage AI tools responsibly, Alex feels prepared to proceed with his high-stakes investigation, knowing the circumstances in which he can safely use LLMs.

Scenario of use #4

As part of its professional development strategy, a national media company integrates the ICSM module into its staff training programme to support several key learning goals: enabling journalists and content creators to use generative AI in alignment with journalistic ethical standards; raising awareness of the limitations and

D3.1 User-centred requirements definition

risks of large language models (including hallucinations and biases); exploring the potential of AI in producing trustworthy journalism; and building an intuitive understanding of how prompt construction, system parameters, and model behaviour influence output quality and reliability.

Through the ICSM, the staff is introduced to key concepts such as prompt design, system settings, and the behaviour of LLMs in media applications. Through guided experimentation and scenario-based learning, participants explore how to craft effective prompts, adjust system parameters, and critically evaluate AI-generated outputs. They learn how to test their own prompts and how to optimize them to get more reliable and more trustworthy results. The training encourages practical engagement and reflection, helping staff understand how their input choices shape the relevance, tone, and accuracy of content.

Ethical considerations are woven throughout the training. Using the ICSM, participants engage in discussions on topics such as editorial bias, or the importance of human oversight in AI-augmented content creation. These discussions are supported by scenarios relevant to the media industry, prompting critical thinking about how generative tools should—and should not—be used in public communication.

By integrating the ICSM module into its training approach, the media company ensures that its employees gain greater confidence in engaging with LLMs, a shared language for discussing generative AI, and a stronger commitment to upholding editorial standards while exploring the creative possibilities of emerging technologies.

8.3.2 *Counter-narrative Recommendation Module*

Scenario of use #5

Mark is a Media Professional who wants to address a claim posted in the comment session of the information platform he is working for (or under a post of the information platform's social media account). The claim is about official data on the European program for decarbonising road transport. The official report containing the information needed is more than 100 pages long and Mark needs a fast way to confront the claim with the report to write an answer.

Mark logs into the application, navigates to "Counter-Narrative Generation," and inserts the claim "We did it! All the work done up to now is enough to reach climate neutrality. We don't need to give up on combustion cars!". He then proceeds to upload the original document of the "Transport and environment report". Given the enthusiastic tone of the comment, in the custom prompt he asks to answer with an empathetic tone and clicks the verification button.

The application analyses the document and highlights a few passages indicating that the claim is misleading and generating a possible answer stating that "I understand your enthusiasm, but while current policies are projected to reverse this trend, they are not sufficient to achieve the EU's 2050 climate neutrality target. This is a long way from the 90 % reduction by 2050 that is needed from transport to achieve the overall 2050 climate neutrality target."

Mark edits this verified verdict to use it as an answer to the user posting it promoting accurate reporting based in verified information.

Scenario of use #6

Andrew is a Fact Checker who wants to assess the truthfulness of a claim that is spreading on several social media with the aim of writing a debunking article. The claim states that "If a long trail from a plane extends for miles and lasts for hours, it is a chemtrail."

Andrew logs into the application, navigates to "Counter-Narrative Generation," and inserts the claim. He then proceeds to upload a reliable document covering the topic of the claim as the "Report on aircraft contrails department for transport". After that he clicks the verification button.

The tool analyses the document and highlights a few passages indicating that the claim is misleading and generates a short explanation stating that those trails are water vapour and that their size and the amount of time they remain visible depends on atmospheric factors.

D3.1 User-centred requirements definition

Andrew repeats the process with additional relevant documents, such as a document on condensation trails from the meteorological service, to assess the claim with more sources and obtain more information and insights.

Andrew then uses the verified verdicts, and the passages extracted from the documents, as starting point to write its debunking article.

8.4 User Stories

8.4.1 Interactive Coaching and Simulation Module

Overarching User Goal: understand how LLMs generate answers to prompts, including the main parameters (that can be conveyed without technical knowledge) and training principles that influence its output, so that the user can critically evaluate and ensure that it is using it appropriately and effectively for different tasks.

User Story #1.1	Title: System Transparency – Understanding Next Word Prediction
Type of user	Journalist
Action	As a journalist, I want to learn how LLM outputs are generated by next-word prediction principles, so that I can understand how LLMs are able to write in a human-like manner.
Use Cases	- Visualize word selection probability distribution in sentence generation. - Experiment with parameter adjustments to observe variations in generated text - Demonstrate how prediction mechanisms can lead to inaccuracies or hallucinations in LLM outputs
User Story #1.2	Title: System Transparency - Exploring System Parameters
Type of user	Journalist
Action	As a journalist, I want to learn about the main system parameters affecting the results of LLM models, such as system prompts and temperature settings, so that I can better understand what factors influence how LLMs generate answers to my prompts.
Use Cases	- Display introduction explanation of system parameters - Set system prompt and temperature settings using the control panel. - Display effect on output from changing system prompt and temperature settings
User Story #1.3	Title: System Transparency - Exploring Supervised Learning
Type of user	Journalist
Action	As a journalist, I want to learn how examples of inputs and outputs for specific tasks are used in training LLMs so that I can understand how LLMs can perform specific tasks like summarization and how I can use a similar technique in my prompt instructions (few shot learning).
Use Cases	- Display explanation of the LLM supervised learning training process - Display examples of LLM prompts and outputs for selected tasks with related examples of training data for supervised learning - Demonstrate how similar principles can be applied in prompts for less common tasks, via a guided exploration of few shot prompting - Provide self-exploration of few shot technique without guidance
User Story #1.4	Title: System Transparency - Exploring Information Retrieval
Type of user	Journalist
Action	As a journalist, I want to understand the differences in how LLMs and traditional search engines retrieve information, including how they attribute sources, so that I can make informed decisions about which tool to use for different types of research and reporting tasks.

D3.1 User-centred requirements definition

Use Cases	
	• Provide introduction and explanation of learning fragment • Provide a choice of documents to upload to the LLM and generate question prompts for demonstrating Retrieval Augmented Generation (RAG) • Display process of going to database and seeing examples of sentences fed into the model • Analyse and highlight differences and similarities between the outputs of the different prompts • Provide explanation of RAG workflow for the prompt/output example

Overarching User Goal: understand how to assess the quality of output from LLMs, to ensure the accuracy, veracity and impartiality of a user's reporting/journalistic pieces.

User Story #2.1	Title: Output Reliability – Main factors of Prompt Construction
Type of user	Journalist
Action	As a journalist, I want to understand main principles of prompt construction, so that I can better interact with the LLM and receive more accurate and reliable outputs.
Use Cases	
	• Provided the different parameters of prompt construction • Compare different outputs generated and highlight differences • Provide outputs for users to guess it's possible initial prompts
User Story #2.2	Title: Output Reliability - Evaluating Changing Output Results over Time
Type of user	Journalist
Action	As a journalist, I want to understand how the same inputs can result in different LLM outputs over time, so that I can be aware of the limitations in the reliability of LLMs.
Use Cases	
	• Compare outputs generated from the same prompt at different points in the interaction sequence. • Provide self-exploration with guidance initially hidden
User Story #2.3	Title: Output reliability - Exploring Few Shot Prompting
Type of user	Journalist
Action	As a journalist, I want to learn about how decisions made during the supervised learning stage of LLM training influence the model's output, so that I can understand how the model is capable of executing certain tasks, and so I can use the same principles to prompt for better outputs.
Use Cases	
	• Display explanation of the LLM supervised learning training process • Display examples of LLM prompts and outputs for selected tasks with related examples of training data for supervised learning • Demonstrate how similar principles can be applied in prompts for less common tasks, via a guided exploration of few shot prompting • Provide self-exploration of few shot technique without guidance
User Story #2.4	Title: Output Reliability - Exploring LLM Induced Misinformation
Type of user	Journalist
Action	As a journalist, I want to learn about how LLMs can output misleading/incorrect information ("hallucinations"), so that I can better critically evaluate the information they provide.
Use Cases	
	• Provide parameters to for users to construct prompts with. • Highlight differences between outputs, including factual inaccuracies such as hallucinations • Link to personal companion service to further explore factual accuracy of generated outputs • Display guidance checklist, detailing information about factual accuracy to guide analysis of outputs

D3.1 User-centred requirements definition

- Provide self-exploration where users construct their own prompts without prompt parameter controls
- Encourage to apply learnings with link to counter narrative recommendation module

Overarching User Goal: understand how input data is utilised by LLM companies, so that a user can assess the lack of transparency in data-use and can protect his/her sources, maintain the trust of the audience, and ensure compliance with relevant laws and regulations.

User Story #3	Title: Addressing Data Privacy Concerns - Explore LLM Model Cards for End-Users
Type of user	Journalist
Action	As a journalist, I want to understand how input data is utilised by LLM companies, so that I can assess the lack of transparency in data-use and can protect my sources, maintain the trust of my audience, and ensure compliance with relevant laws and regulations.
Use Cases	• Display a brief overview of common disclosed data collection practices by major LLM companies. • Provide link to “Transparency services for AI-model cards” service to explore and observe what kind of data is used in LLM training and collected during LLM usage and how the transparent that is for the end-users • Provide guidance on how to use model cards service

Overarching User Goals: understand how LLMs process and output natural language so that a user can interact with the system in an appropriate manner and critically evaluate interactions with it.

User Story #4	Title: System Interaction & Anthropomorphism – Understanding Prompting Style
Type of user	Journalist
Action	As a journalist, I want to understand the difference between using human-like and non-human-like language to instruct the system to perform a certain task, so that I can communicate effectively and efficiently with these systems.
Use Cases	• Generate prompt according to different parameters selected and subsequently generate output from the prompt. • Compare and highlight different outputs, building understanding of appropriate prompting technique, including how human-like conversation technique is not needed and may pose risk in interactions • Provide link to self-exploration where users construct their own prompts without prompt parameter controls • Option to apply learnings with link to counter narrative-recommendation module

8.4.2 Counter-narrative Recommendation Module

User Story #1	Title: Claim verification
Type of user	Media Professional
Action	As media professional, I want to verify if a claim published online is true with respect of reliable information source so that I can more easily confirm or debunk the claim without reading in all their extension the related documents or reports (sometime very long). I also want to have a suggestion on how to answer for social correction activities
Use Cases	• Insert claim and supposed source document

D3.1 User-centred requirements definition

• Display resulting truth assessment for claim and document along with counter-narrative recommendation • Assess the correctness of the answer based on relevant passages extracted from the document • Edit the counter-narrative as needed before using the text	
User Story #2	Title: Claim verification based on source document
Type of user	Journalists, Fact-checkers
Action	As a journalist, before trusting a claim I want to verify if it is true for a given source document, without reading in all their extension the related documents or reports (sometime very long).
Use Cases	
• Insert claim and supposed source document • Display resulting truth assessment for claim and document along with counter-narrative recommendation • Assess the correctness of the answer based on relevant passages extracted from the document	

8.5 Early Requirements

#	Early high-level functional requirements	How will the requirement be addressed in the service design?
1	The system shall provide an interactive visual illustration of the next-word prediction mechanism as a foundational principle for understanding text generation by LLMs.	The system will illustrate how LLMs generate a sentence through next word prediction on a realistic examples. It will do so by showing different words with corresponding prediction scores at different steps, allow the user to explore alternative choices at different steps and observe their effects on the sentence composition.
2	The system shall provide the possibility to interactively choose the setting for main factors in prompt construction and observe the influence of these choices on prompt variations and the resulting output, for a pre-defined set of examples.	The system will provide users with a set number of controls for interactively changing/influencing the composition of pre-defined example prompts and display resulting LLM outputs. The effects of the prompt control changes on corresponding LLM outputs will be highlighted in the user interface so they can be easily spotted and compared by users.
3	The system shall enable users to compare and analyse differences in LLM outputs for different prompt variations, including factual inaccuracies.	Main differences in LLM outputs for different prompt variations will be highlighted in the user interface so they can be easily spotted and compared by the users. Factual inaccuracies in LLM outputs for user-defined prompts will be identified via the API of the AI-CODE Personal Companion for Understanding Disinformation service (PCS API).
4	The system shall provide a self-exploration mode for user experimentation with prompt construction.	The system will provide users with a possibility to input own prompts and receive suggestions for possible prompt adaptation by directly manipulating a set of prompt construction controls. The system will present real-time LLM outputs for the original and for the adapted prompts in a way that allows users to compare

D3.1 User-centred requirements definition

		them to each other. This shall help users craft, test and optimize their own prompts to achieve more reliable results.
5	The system shall provide the possibility to change a selected number of parameters that influence the results produced by an LLM (such as system prompt or temperature settings) and allow the user to observe the difference in the results.	In the simulation, the system will allow users to change a few main types of system parameters (e.g. temperature) and observe how that effects the (pre-generated) LLM output for pre-defined prompt examples .
6	The system shall provide explanations for the parameters the user can change interactively.	The system will provide an explanation, such as a pop-up or hover, explaining the different parameters.
7	The system shall provide a guidance checklist for analyzing reliability of LLM outputs and supporting functionalities for identifying potential inaccuracies by integrating selected functions from AI CODE claim veracity verification services.	The system will provide a checklist for evaluating the quality/reliability of LLM outputs for ensuring trustworthy content. Explanations will be provided for the main checklist items.
8	The system shall provide an explanation of the LLM supervised learning process, showcase examples of LLM tasks and their outputs, and implement a guided exploration of few-shot prompting technique.	The system will present a series of examples of given LLM outputs and input-output pairs that would have been used in the supervised training output, along with an explanation of how the LLM is trained and how users can utilise this principle themselves in few-shot-prompting
9	The system shall demonstrate differences between LLMs and traditional search engines, allow users to define a query and select documents for a simulated RAG retrieval, and highlight and compare differences in the obtained results.	The system will be able to select a number of documents to be inputted with a selected prompt for a pre-generated example. The output will change according to the user selection and differences in the output will be marked to help users identify and compare the changes.
10	The system shall allow users to compare outputs from the same prompt at different points in the interaction sequence to illustrate how LLM outputs can change over time	The system will highlight differences between the corresponding LLM outputs to signal to the user parts of the output that have changed and their connection to the evaluation checklist.
11	The system shall allow users to input claims and source documents and display truth assessments for these claims and documents	The system will display a suggested counter-narrative for the claim grounded on the documents provided by the user. The user can generate multiple recommendations.
12	The system shall provide counter-narrative recommendations for the assessed claims.	
13	The system shall enable users to edit suggested counter-narratives before use.	The system will provide the counternarrative in an editable box to allow the user to change the text as needed
14	The system shall enable the user to compare the generated counter-narrative with the provided background information (to assess its consistency)	The system will provide a list of the most relevant sentences for the counternarrative, and will highlight them in the original article

#	Early high-level non-functional requirements	How will the requirement be addressed in the service design?
2	The application shall be accessible and useable for non-technical audiences	The user interface will follow intuitive design principles with clear instructions, tooltips, and guided walkthroughs for ease of navigation.
3	The application shall be easy to use	A user-friendly interface design will be implemented allowing users to quickly access

D3.1 User-centred requirements definition

		different functionalities. It will implement visual consistency, a clear logical flow and a clean design. The interface design will consider structuring functionalities from simpler to more complex (e.g. guided simulation with pre-generated examples separated from self-exploration on user-defined prompts).
4	The application shall be easy to learn	Guided walkthroughs and contextual tooltips will be embedded within the system to facilitate use without prior expertise.
5	The application shall be used in a browser	The service will implement a web-based user-interface running in a browser, requiring no installation.
6	The application shall provide differentiated support for beginner and expert users as two main types of user cohorts within the target user group of the application.	The structuring of the user interface will provide more basic, core functionalities more readily visible and accessible so that beginner users are not overwhelmed. More advanced functionalities will be provided further in the user flow and/or with navigation cues (e.g. visual shortcuts) allowing more expert users to skip ahead i.e. invoke them as needed.

8.6 Ethical Requirements

#	Ethical requirements	How will the requirement be addressed in the service design?
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system	The system will clearly indicate when a LLM is used to generate responses.
HAO-02	The capabilities of the application shall be communicated clearly to the user.	The system will provide an explanation of its functionalities and how they should be used appropriately.
PDG-01	The application shall minimize the amount of data collected on users.	The system will follow a privacy-by-design approach, collecting only necessary data with user consent and anonymization where possible.
TRA-01	The system input and output of the application shall be traceable through documentation.	The system structure and data processing workflow will be documented in the relevant deliverables.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI component.	The system will provide an explanation of the purpose of the truthfulness assessment of claims including an explanation of the obtained result of this assessment. It will inform the users that they should use the latter to critically assess the validity of this AI-generated result.
TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated	The system will allow users to observe the differences in the quality of LLM results for different prompt variations and include guidance for verifying the accuracy LLM-generated outputs.
ENV-02	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects	Risk assessment procedures will be incorporated to evaluate unintended consequences for user groups.

HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case	The tool is not allowed to automatically post the results on any communication channels. The system outputs can be edited and must be manually copy-pasted by the user in another application if they wish to use them for a specific purpose.
---------------	--	--

8.7 Early Success Criteria

System Effectiveness benchmarking

- The Counter-Narrative Recommendation Module will effectively verify and explain false claims.

User evaluation

- The Interactive Coaching and Simulator Module will help media professionals to develop a better understanding of main principles of how generative AI works (LLMs) and of the associated risks of using generative AI.
- The Interactive Coaching and Simulator Module will help media professionals to gain a better understanding of how to critically evaluate the output and the use of LLM systems in trusted media production.
- The Interactive Coaching and Simulator Module will help media professionals to gain an effective understanding of how the modality of use of LLMs influences the reliability of generated outputs.
- The Counter-Narrative Recommendation module will enable media professionals to perform claim verification and respond to disinformation posts more effectively
- Media professionals will trust the Counter-Narrative Recommendation module's claim verification.

8.8 Early Mock-ups of the Service

8.8.1 Interactive Coaching and Simulation Module

Shown below are a series of mock ups that will graphically illustrate the different functionalities of the interactive simulation.

D3.1 User-centred requirements definition

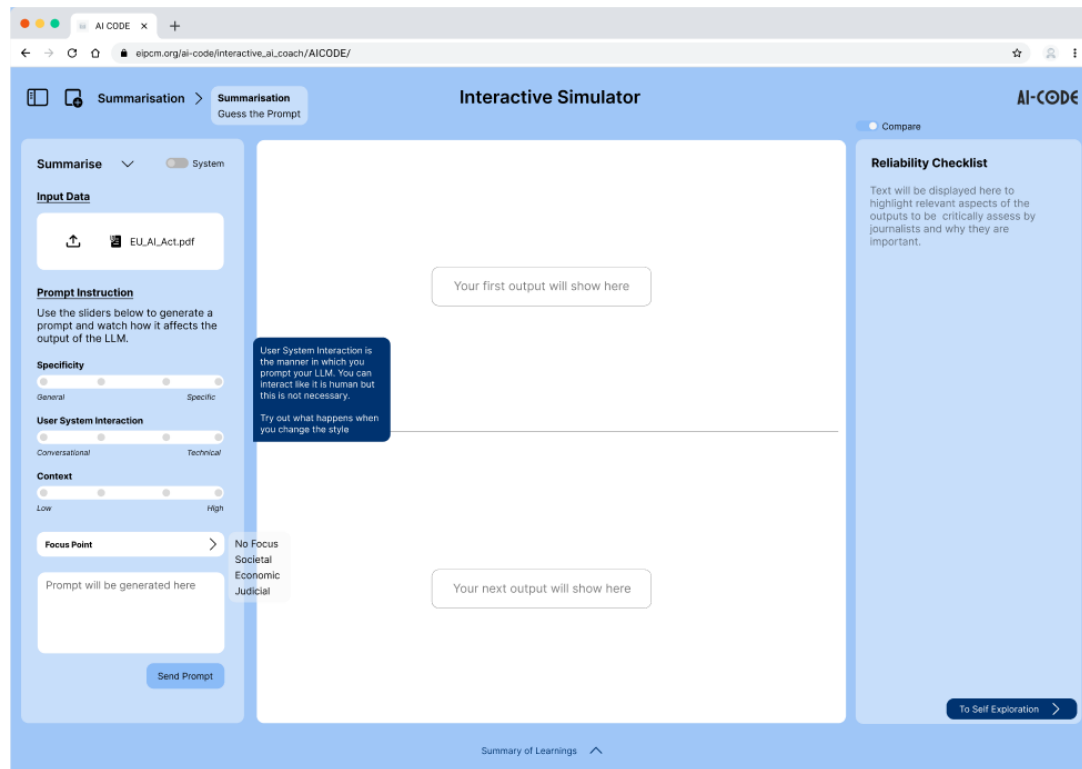


Figure 23: Initial orientation screen for the simulator teaching prompt instructions through a summarisation task. A prompt control panel on the left, Outputs panel in the centre and on the right a section for guidance. Users can hover over terms to learn about its effect.

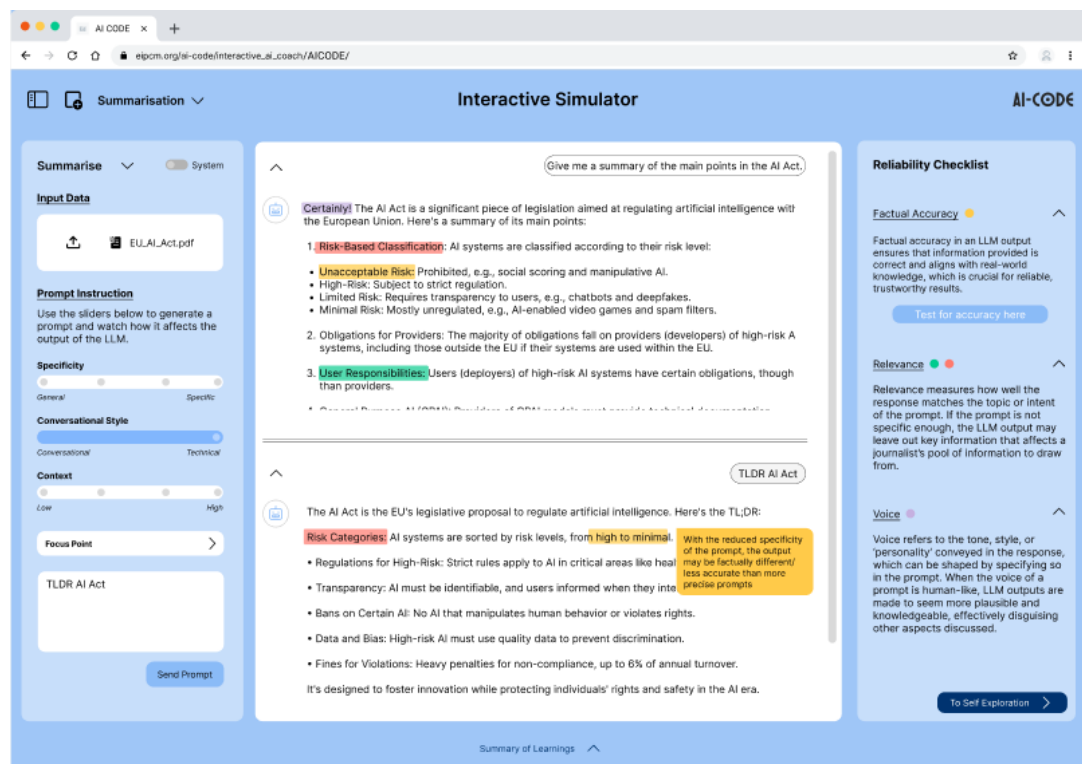


Figure 24: Simulator with two different prompts generating outputs and differences highlighted with colour, referenced to the corresponding factor on the reliability checklist. The previous prompt output settings reflect the most recent prompt. Users are encourage

D3.1 User-centred requirements definition

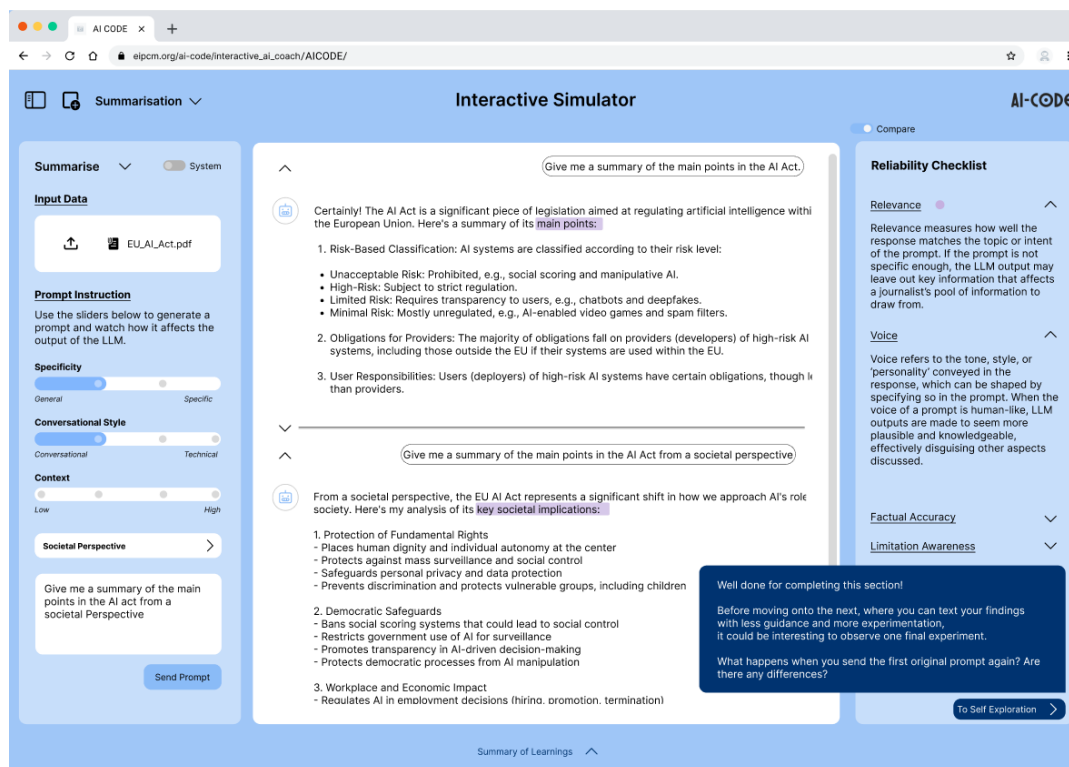


Figure 25: Users can change the focus point of a prompt instruction and pop-ups can help suggest particular topics to explore, such as to test the same prompt twice.

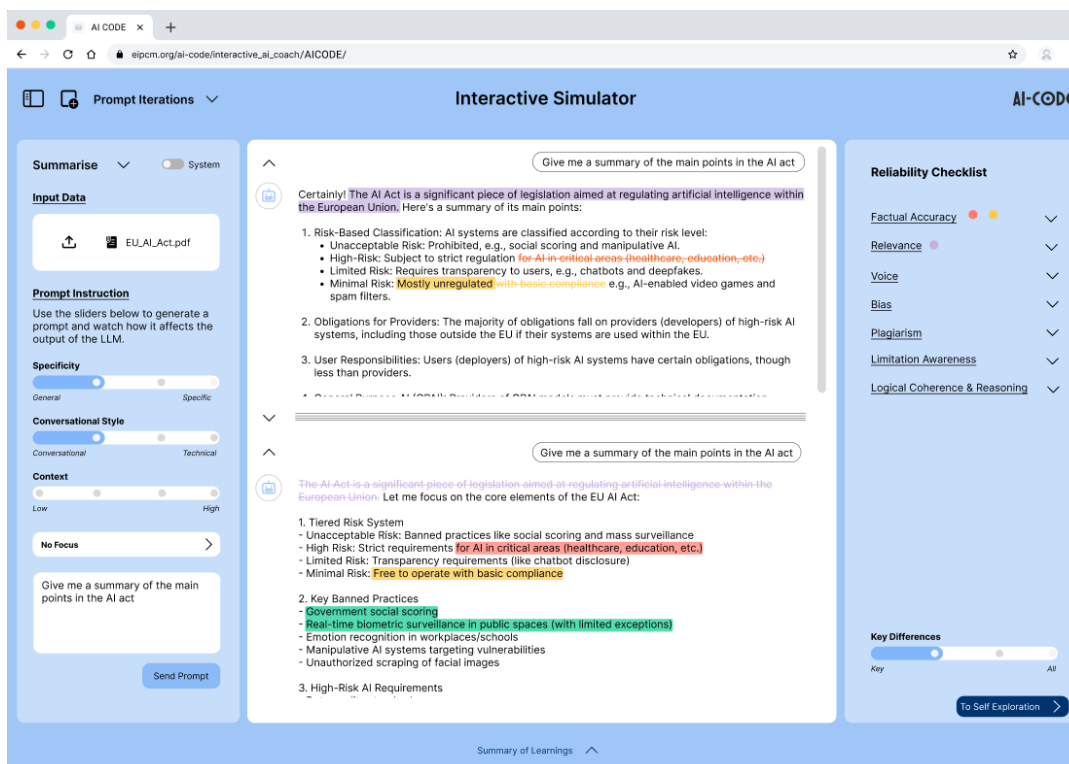


Figure 26: Simulator screen showing how each of the outputs generated can be compared, including by viewing how the same prompt produces different outputs. Users can select how many differences are highlighted with slider. Checklist items can be expanded for more

D3.1 User-centred requirements definition

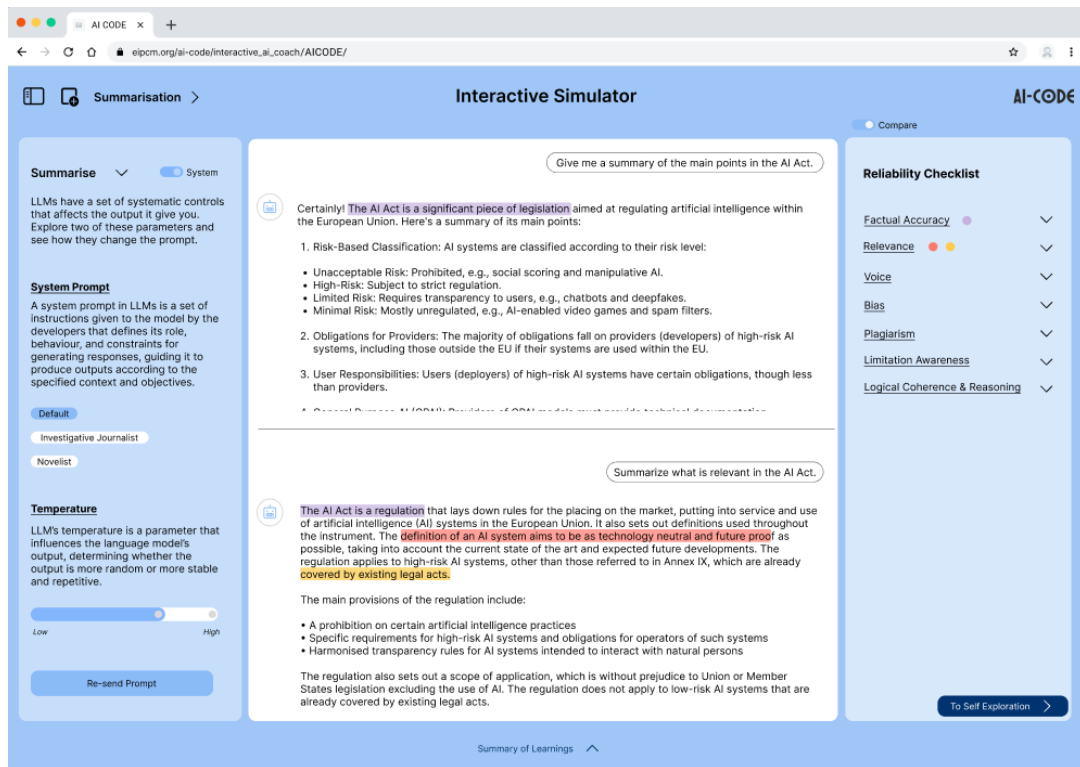


Figure 27: Simulator with the system controls toggled. Users can explore how system prompts, and temperature can affect outputs and read more to understand what they are. Differences in texts are highlighted and mapped to the checklist

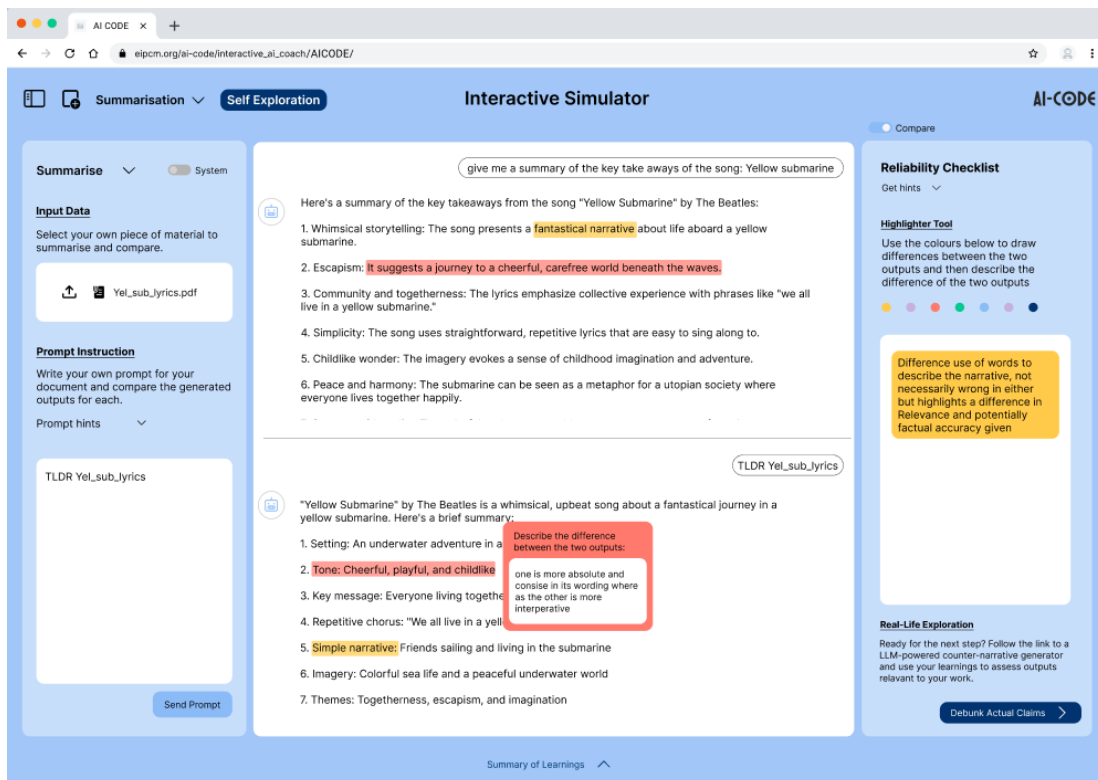


Figure 28: Self exploration of the prompt instruction task is shown, with users able to select a different example and analyse their own outputs using a highlighter tool. Users can write what the differences are between outputs. Hints are available for help and us

D3.1 User-centred requirements definition

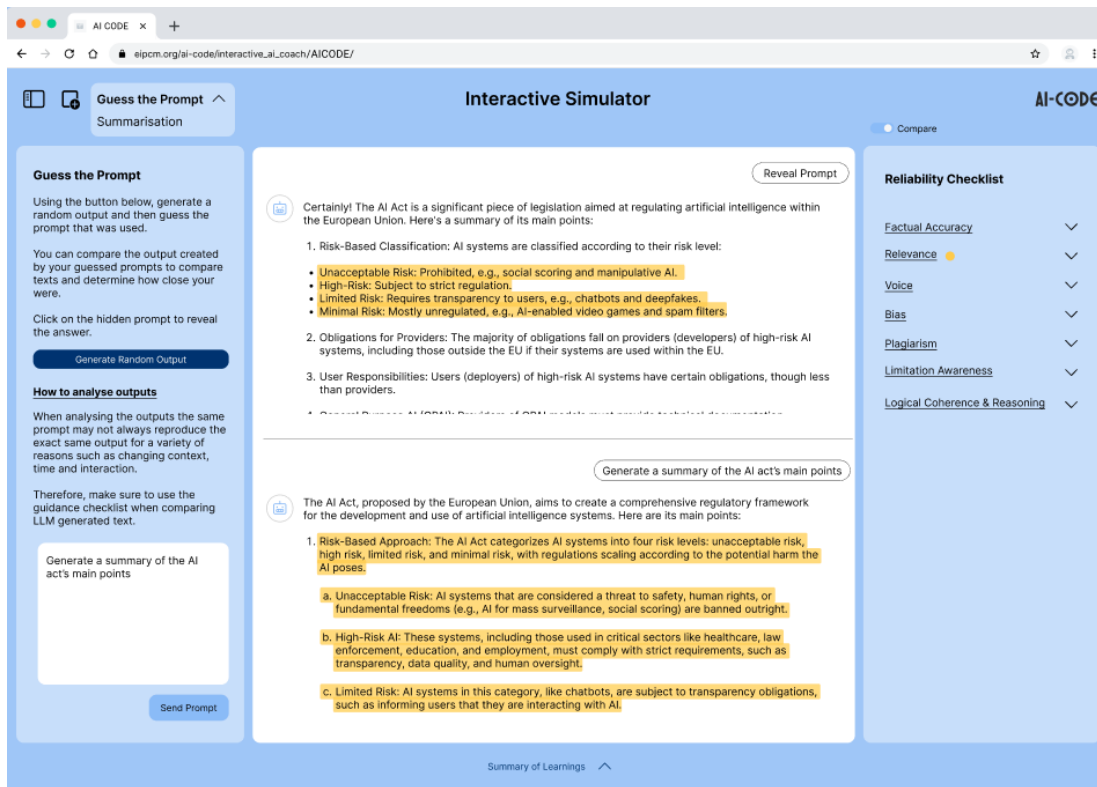


Figure 29: Guess the prompt simulation encourages users to guess what prompt created a given output and thus serve as an exercise to evaluate and compare output critically and with understanding of the prompt instruction. Through inputting their own guess prompts

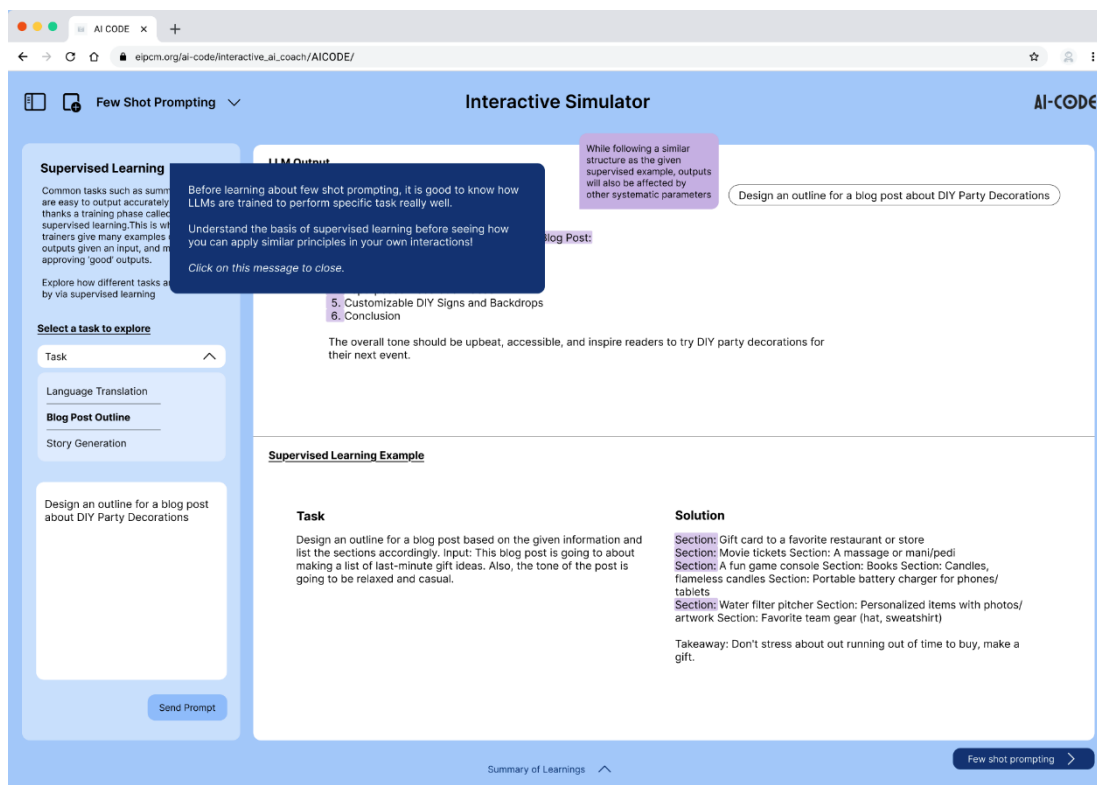


Figure 30: Interface where users learn about supervised learning and spot similarities and differences between generated outputs and example supervised learning examples for a variety of tasks.

D3.1 User-centred requirements definition

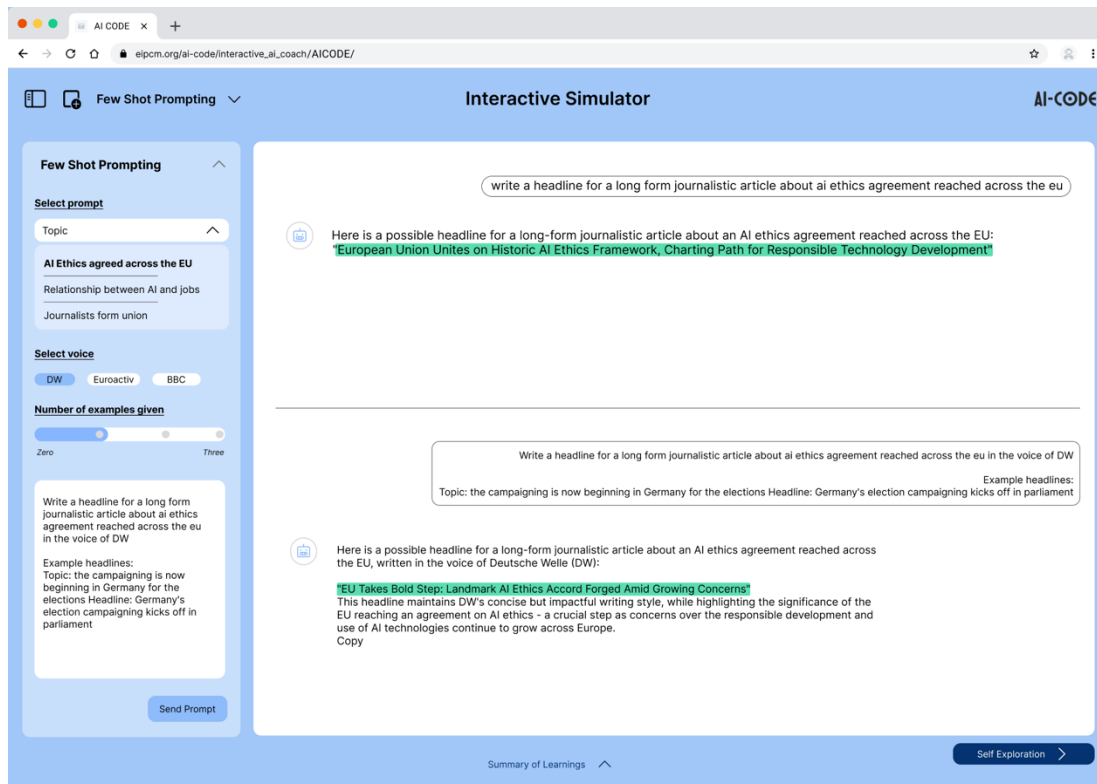


Figure 31: Interface following the supervised learning interface, where users discover how to include their own input/output examples to improve the responses they get for less common tasks. Users can control the number of examples given, a specific voice and topic, to compare outputs.

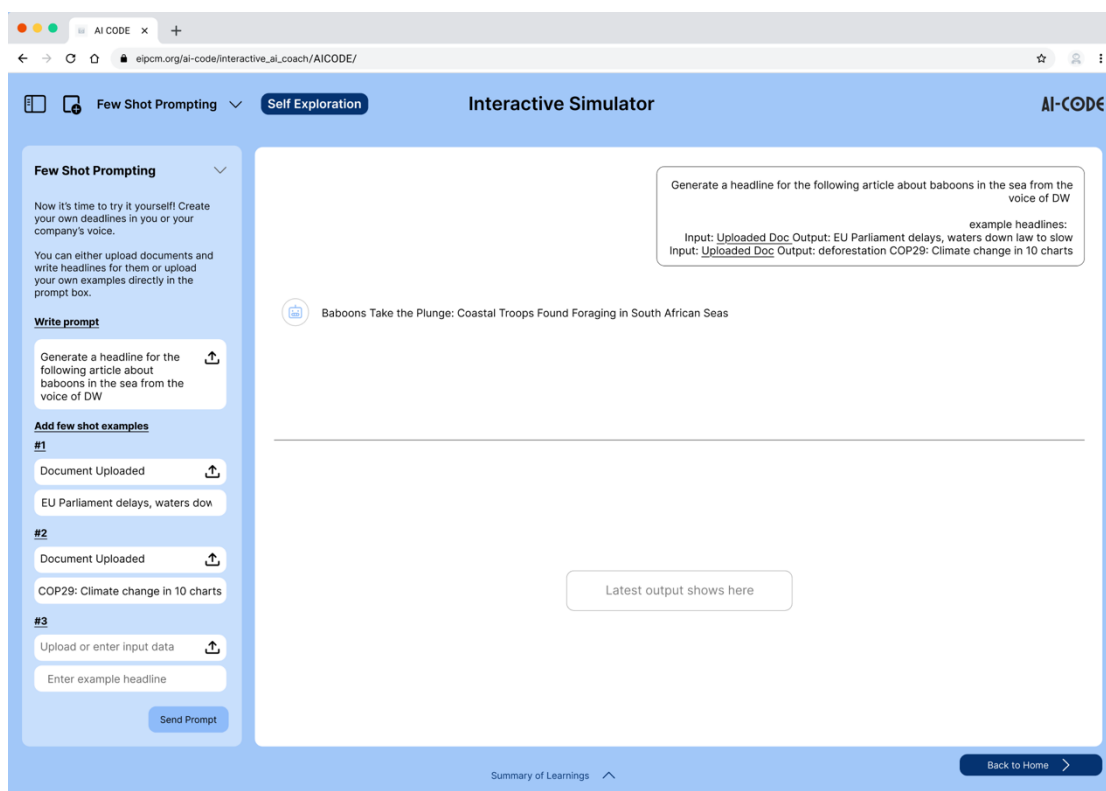


Figure 32: Self Exploration page where users can upload their own prompts and examples and see further differences

D3.1 User-centred requirements definition

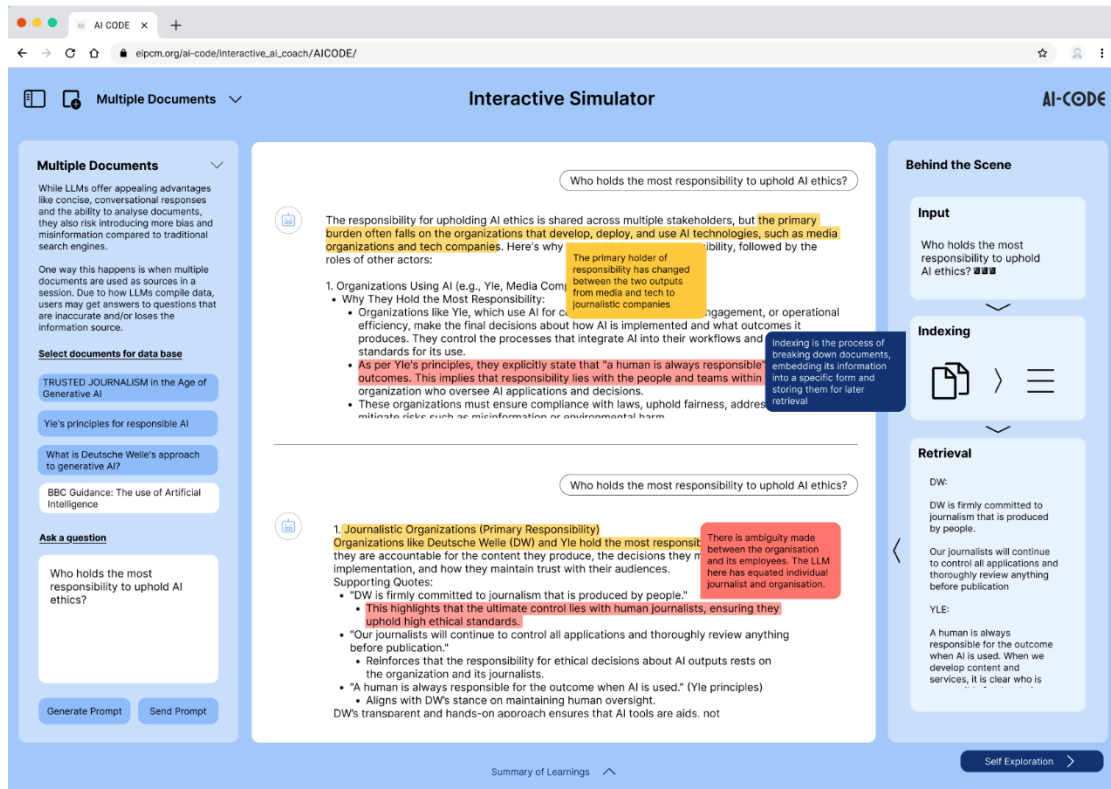


Figure 33: Users explore how combining multiple documents can introduce bias/misinformation by selecting different documents and asking questions to compare outputs given. A 'behind the scene' look at a simple RAG diagram is shown with the relevant document information shown. Hover to explore definitions of key technical terms.

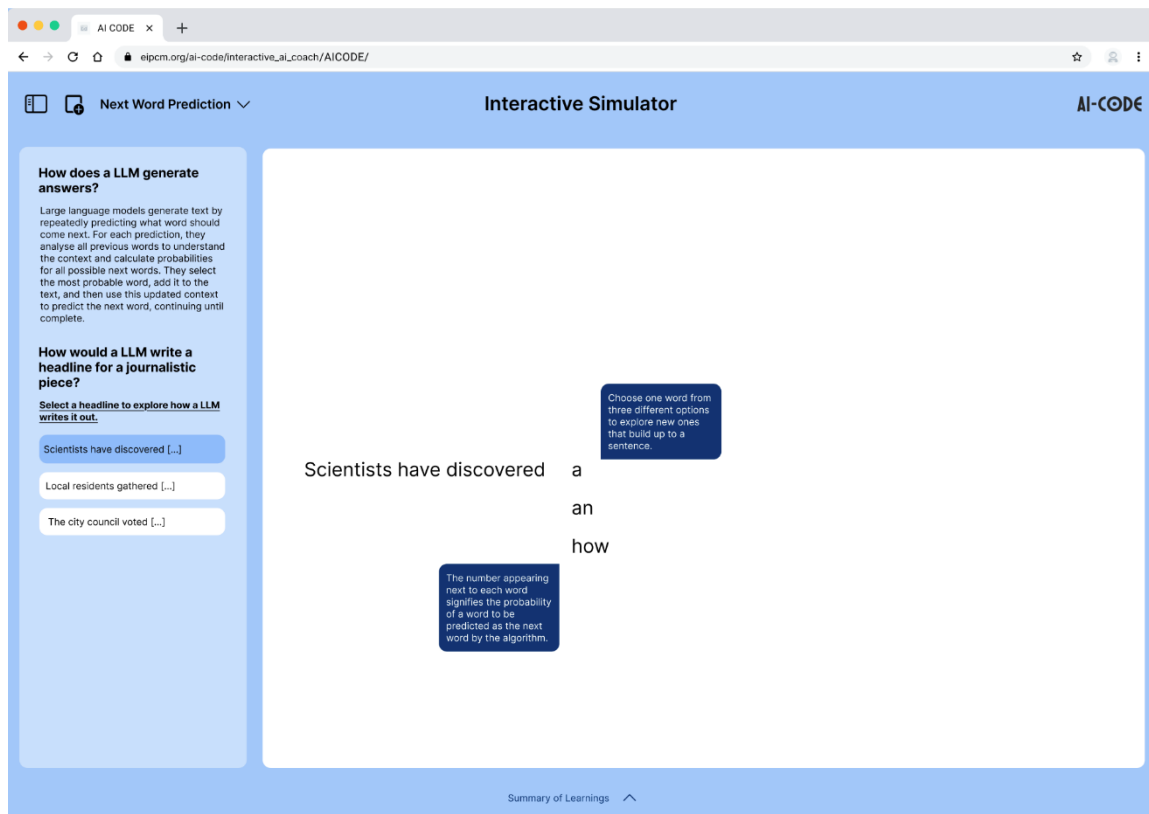


Figure 34: Interface explaining next word prediction via a interactive word chain for different example. Users can select different words with different likelihoods of following the previous word to build up a sentence.

D3.1 User-centred requirements definition

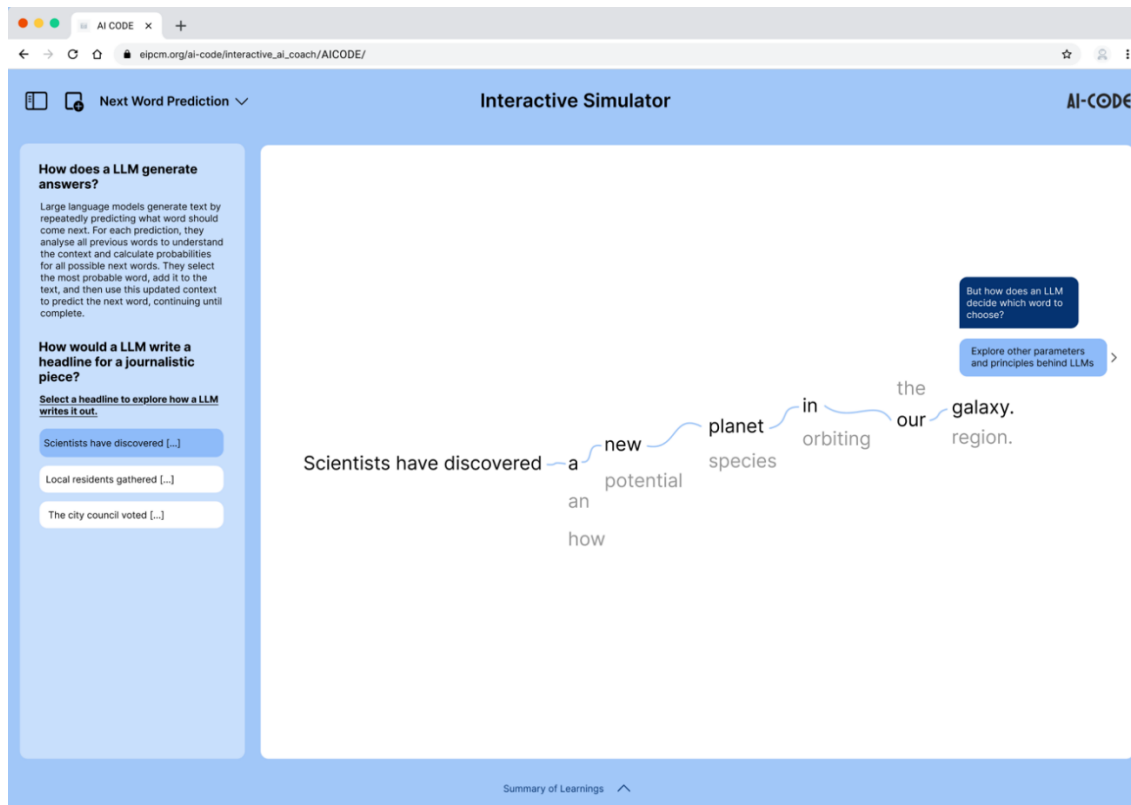


Figure 35: Screen showing completed sentence made by next word prediction principles and can then explore how LLMs choose it's words and explore other parameters.

8.8.2 Counter-narrative Recommendation Module

Counter Narrative recommendation

Insert a Claim:

We did it! All the work done up to now is enough to reach climate neutrality. We don't need to give up on combustion cars!

Insert an Article:

In addition, the Energy Efficiency Directive and the Energy Taxation Directive also have an impact on the road transport sector. For example, historically, it has been demonstrated that high taxes have contributed to more fuel-efficient cars and thus to lower average CO2 emissions per vehicle-kilometre of the car stock in the EU than the car stock in the United States, which has lower prevailing fuel taxes (Eberhard et al., 2000). This set of directives and regulations, their targets for 2030, their relevance for road transport and proposed changes in the Fit for 55 package are summarised in Table 2.2. The EU has adopted rules to ensure monitoring of progress

Number of relevant sentences to display:

5

Select the language of the text:

English

Select the writing style:

journalistic

Generate

Truthfulness explanation:

Current policies are projected to reverse this trend, but not sufficiently to reach the EU's 2050 climate neutrality target. This is a long way from the 90 % reduction by 2050 that is needed from transport to achieve the overall 2050 climate neutrality target.

Relevant Sentences:

This is a long way from the 90 % reduction by 2050 that is needed from transport to achieve the overall 2050 climate neutrality target.

Until 2019, the electrification of the vehicle fleet did not yet play a significant role in reducing CO2 emissions from road transport.

The EU climate-energy policy framework relevant for road transport greenhouse gas emissions 2.2.1 The climate challenge and the global and EU climate targets Global GHG concentrations in the atmosphere have already reached levels that the Intergovernmental Panel on Climate Change (IPCC) states should not be exceeded in order to limit the global temperature increase to 1.5°C above pre-industrial levels by 2100.

Climate neutrality means that by 2050, the EU aims to achieve net zero GHG emissions.

IPCC, 2021, 'Summary for policymakers', in: Climate change 2021: The physical science basis.

Article:

EEA Report No 02/2022 Transport and environment report 2021 Decarbonising road transport — the role of vehicles, fuels and transport demand Transport and environment report 2021 Decarbonising road transport — the role of vehicles, fuels and transport demand EEA Report No 02/2022 Legal notice The contents of this publication do not necessarily reflect the official opinions of the European Commission or other institutions of the European Union. Neither the European Environment Agency nor any person or company acting on behalf of the Agency is responsible for the use that may be made of the information contained in this report. Brexit notice EEA products, websites and services may refer to research carried out prior to the UK's withdrawal from the EU. Research and data relating to the UK will generally be explained by using terminology such as: 'EU-27 and the UK' or 'EEA-32 and the UK'. Exceptions to this approach will be clarified in the context of their use. Copyright notice © European Environment Agency, 2022 Reproduction is authorised provided the source is acknowledged. More information on the European Union is

Figure 36: The service interface consists of two sections. On the left there is the input section where the user can provide the claim and the article, as well as selecting the language (currently only English is supported) and writing style. On the right side the system provides the counter narrative with the truthfulness explanation, a list of sentences that are relevant for the counter-narrative and the article with the relevant sentence highlighted to see them in their context.

9 Transparency services for AI-model cards

9.1 Service Description

9.1.1 Scope and objectives

The transparency service for AI model cards will be developed to define, create, and manage model cards for AI models. It will establish a standardized structure for model cards, specifying necessary fields such as performance metrics, algorithm descriptions, and ethical considerations, ensuring consistency across models and third-party applications that integrate these cards.

The service will offer an intuitive user interface (UI) to make model card creation accessible to users with varying technical backgrounds. It will also provide a software development kit (SDK) for developers to annotate their code, automating the generation of performance metrics and relevant information directly from the AI models.

Beyond creation, the service will include management functionalities like version control, updates, and model card storage, enabling developers and teams to maintain a history of model improvements, compare versions, and ensure that the latest information is readily available. Additionally, the service will feature varying levels of detail within model cards to cater to both technical experts and non-technical stakeholders. It will also integrate advanced chatbot assistants to guide engineers through filling out necessary fields, offering suggestions and explanations, especially for complex model aspects.

9.1.2 Supporting critical thinking

The transparency service stimulates critical thinking among AI developers by making fields related to ethical risks, intended use, and misuse mandatory when creating a model card. By embedding these fields into the workflow and offering guided assistance through AI-driven support, the service encourages engineers to reflect deeply on the broader implications of their work. This process, going beyond technical documentation, invites model creators to critically evaluate potential harms, misuse scenarios, societal impact, and fairness issues.

On the user/consumer side, the service promotes critical thinking by providing model card information at multiple levels of detail, tailored to different user backgrounds and expertise. Non-expert stakeholders, such as journalists or media professionals, gain insights into how an AI model functions, what its limitations are, and what risks it may pose. For example, in critical domains like synthetic content detection, transparency allows users to grasp how AI systems might be used to manipulate media, the current technological limitations in detecting such cases, and how to interpret detection results. As a result, users are empowered to make informed judgments about the trustworthiness and use of AI technologies in real-world scenarios. Also, offering valuable insights into a model's training process and architecture, helps to explain model behaviour and implications. For example, imbalances in the training data can affect performance across different scenarios. Understanding these factors is crucial, as it encourages users to approach AI outputs with critical thinking rather than unquestioned acceptance.

9.2 Target Users and Expected Benefits

Target users	Expected Benefits
Primary target users	
Model Card Creators (AI / ML Engineers, Researchers, Developers, etc.)	• Streamlined Documentation Process: Simplifies the creation of comprehensive and standardised model cards, reducing the time and effort needed for documentation.

D3.1 User-centred requirements definition

	• Ethical and Responsible AI Focus: Ensures critical issues such as bias, fairness, and ethical implications are highlighted and systematically documented. • Centralised Management: Offers a lightweight yet centralized solution for easy updates and version control of model cards, ensuring that documentation remains up to date as models evolve, without adding unnecessary complexity or overhead to the development process. • Collaboration-Friendly: facilitates collaboration across teams by enabling the export of model documentation to various platforms (e.g. HuggingFace), ensuring all stakeholders have easy access to the most up-to-date information.
Data Scientists (Media Industry)	• Model behaviour: Data scientists in media industry are responsible for tasks such as analysing user behaviour, segment users into groups, identifying trends and more. An AI model that assists in such tasks should be transparent enough so the user can understand its behaviour and decision making. • Biases and pitfalls: The user should also be aware of potential biases and pitfalls that he may encounter during its use. Those must be addressed in the model card along with the training set that was used which can provide valuable information about intended and unintended uses.
Content Moderators	• Explainability & Interpretability: Providing explainability & interpretability to users of AI models that perform content moderation or identifying inappropriate content is crucial. Understanding what features or tokens influence the outcome enables transparent decision-making allowing validation of the system's judgment and targeted improvements to the model. • Performance: Moderators can use AI models much more responsibly if they understand its performance. This can provide useful information such as false positive and false negative rates or biases like underperforming on certain content. This can prevent wrongful takedowns or missed harmful content.
Model Consumers (Journalists, Fact-checkers, Developers, etc.)	• Tailored Information and Guidance: Provides different levels of detail in model cards, allowing users of varying expertise to access relevant and understandable information. The system helps guide users to the appropriate level of detail based on their specific needs, ensuring they receive the most relevant insights for their task or problem at hand. • Improved Interpretability: Helps users better understand the results of AI models, increasing trust in the model's output. • Transparency in AI: Enables users to assess key aspects such as data sources, model limitations, and ethical considerations, promoting a more informed use of AI models. • Accessibility: Makes complex technical details digestible for non-experts while still offering deep dives for those with technical background.
Publication Editors	• Auditing AI Accuracy in Editorial Workflows: Publication editors are responsible for collecting, organizing, and approving content to ensure it meets professional standards of quality, relevance, and fairness. As AI tools are increasingly being used, tasks such as what content gets promoted (e.g., based on user behaviour on a platform), how content is categorized (politics, health, science, culture, etc.), and whether it complies with editorial policy (sensitive content, offensive language, etc.) are automated. Understanding the accuracy and limitations of these models is essential. For example, analysing how content is proportioned under different conditions, reveals whether certain topics, formats, or voices are systematically promoted or downranked. Performance metrics in classification show the accuracy, including false positives and false negatives in labelling content by topic.

	Similarly in policy enforcement, it highlights whether models fail to detect policy violations or wrongly flag legitimate content. Information like this, along with insights into biases in training set and clearly defined use cases, promote responsible and informed decision-making around AI. • Reliability in AI-Assisted Breaking News and Source Selection: During breaking news coverage, editors face the challenge of publishing quickly while maintaining the accuracy of the information. AI systems that assist in retrieving, ranking, or summarizing external information can accelerate this process, but they also introduce new risks. Model cards can reveal how these systems identify and prioritize sources such as favouring government websites, academic publications, or popular media. By providing such insight, journalists can understand the sourcing and reliability of its information, helping to limit the risk of spreading misinformation or fake news.
Technical Editors	• AI Transparency in Technical Editing: Technical editors are responsible for reviewing content on sensitive or scientific subjects like health, technology, climate, and more. Transparent information in model cards, such as token-level saliency, data recency, source breakdown, and labelling behaviour, allow editors to evaluate how AI systems summarize studies, classify topics, or flag controversial content. This helps identify outdated information, misrepresentation of technical topics, or reliance on limited source types, ensuring that the final content remains accurate and suitable for the audience.

9.3 Scenarios of Use

Scenario of use #1: Interpreting Conflicting AI Results for fact-checking

Evangelia, a fact-checker at a news organization, is reviewing an image flagged by multiple AI detection models through AI-CODE's Media Asset Annotation and Management platform (T4.3) that uses AI to automatically organize media assets. One model has returned a positive result, indicating the image is likely synthetic, while two others have resulted in no detection. To resolve this uncertainty, Evangelia accesses the platform's model results dashboard, where she can directly view the detailed model cards for each detection model. These model cards, integrated within the platform, provide insight into each model's algorithm, training data, and performance metrics, helping Evangelia better understand the discrepancies between the results.

The model card for the first, positive-result model explains that its detection of synthetic content is based on the presence of patterns consistent with Latent Diffusion generative models. It also provides a high confidence score, but notes potential limitations, such as struggles with highly complex or mixed-content images. Evangelia then checks the model cards of the two models that produced negative results. One of them indicates it is optimised for detecting Generative Adversarial Networks (GANs), not diffusion models, which could explain why it missed this detection. The other model card highlights its tendency to return more conservative results, prioritising avoiding false positives over flagging borderline cases. By interpreting the specific strengths, limitations, and focus areas of each model as outlined in their cards, Evangelia can make sense of the results. She concludes that the image is most likely synthetic, produced by a Latent Diffusion method, and uses this well-rounded understanding to inform her final report.

Scenario of use #2: Providing details of AI models training dataset at different levels of detail

Leah, an expert in AI-powered fact-checking, uses a deepfake detection system to analyse a suspicious video. The AI model flags the video as likely manipulated, and Evangelia refers to the model card for detailed insights. The deepfake detection system has already integrated the transparency service and can therefore automatically provide a simplified version of the model card. Using this simplified version, she can effectively communicate the results to a colleague with limited technical knowledge.

D3.1 User-centred requirements definition

This simplified version highlights the key outcome—confirmation of the deepfake—without overwhelming technical jargon. The simplified card explains that the AI model identified manipulation patterns with high confidence and clarifies that the dataset used to train the model was balanced across various demographic groups. Leah pays special attention to the section on potential biases. The model card describes how the training dataset was carefully curated, ensuring diverse representation across different phenotypes. This reduces the risk of biased detection, particularly in relation to facial features tied to ethnicity or other demographic factors. Leah includes this information in her report to the journalist, emphasising that the AI's decision is trustworthy not only due to its technical accuracy but also because of the diverse dataset used for its training.

Scenario of use #3: Assessing AI Models for LLM-Generated Text Detection

Manu, a developer working on a new application, wants to integrate an AI model to detect text generated by large language models (LLMs). After finding a service with a provided model that offers this functionality, he decides to thoroughly assess its performance to ensure it meets his application's needs. The service, which has already integrated model cards produced by the transparency service, provides a detailed model card with comprehensive evaluation metrics. Manu consults the model card to gain insights into the model's performance and determine its suitability for his use case. The model card allows him to review how the AI performs across different types of LLMs, languages, and text categories. He's particularly interested in long news articles, a critical type of text for his application. As he explores the model card, Manu notices that while the AI performs well for short, general-purpose texts and across several languages, its detection accuracy drops significantly when applied to long articles. The model card provides specific evaluation metrics that highlight this shortcoming, offering insight into why the model struggles with this type of content. Given this detailed information, Manu is able to make an informed decision. The transparency provided in the model card helps him avoid potential issues before integrating the service, saving time and resources.

Scenario of use #4: Documenting a model using the transparency service

Sophia, an AI engineer, has trained a model to assess the trustworthiness of user profiles in social media based on various signals such as behaviour patterns, account history, and engagement metrics. Now, she needs to document her model appropriately to ensure transparency and usability for others. To do this, she turns to the AI-CODE transparency service designed for creating detailed model cards.

Sophia starts by using the service's functionality that allows her to annotate the model's source code directly. This helps her automatically generate a comprehensive set of evaluation metrics, tailored to the trustworthiness assessment task. The transparency service also guides her through the selection of appropriate metrics. By selecting the correct task category, the framework suggests which metrics are most relevant, explaining their advantages and limitations. For example, it highlights why certain fairness or accuracy metrics might be crucial, particularly in the context of trust assessment, and warns about potential shortcomings in interpreting them. Once she has a clear set of metrics, Sophia uses the service's intuitive UI to fill in the structured document corresponding to the model card. Most fields, such as model type, dataset details, and evaluation results, are easily filled through drop-down menus that adapt to the specific task of user profile analysis. This saves her time and ensures consistency across the documentation. When it comes to fields such as ethical concerns, the transparency service offers a chat-based functionality that prompts Sophia with targeted questions. The system helps her think through potential biases or unintended consequences of her model, such as over-flagging certain demographic groups. With the help of the transparency service, Sophia can create a complete and well-structured model card that reflects both the technical performance and ethical implications of her model, making it accessible and trustworthy for future users.

Scenario of use #5: Machine-readable model cards for Integration with third party services

Alex, a software engineer, is developing a face identification and deepfake analysis application for a media company. The application uses several AI models to detect manipulated faces and verify the authenticity of user-uploaded content. While the application provides real-time results, Alex wants to offer additional contextual information to help users better understand the outputs, such as why a certain image was flagged as a deepfake or the confidence level of the identification process. To achieve this, Alex turns to model cards produced by the transparency service, that not only have versions tailored for human readability at different

D3.1 User-centred requirements definition

levels of detail but also include machine-readable versions. These structured model cards allow the application to automatically pull relevant information for users based on the AI model's outputs. Under the hood, the model card is formatted in a standardised, machine-readable way, such as JSON, making it easy to integrate into Alex's application. For instance, when the face identification model flags an image, the machine-readable model card provides a snippet of information explaining the detection threshold and any biases the model may have. Similarly, in the case of a deepfake detection output, the model card offers real-time context on the type of manipulation detected and the algorithm used. By embedding the machine-readable model card into the service's backend, Alex can ensure that the application always displays the most accurate and up-to-date information about the AI models, improving transparency and helping users trust the system's outputs.

Scenario of use #6: Collaborative Management of Model Cards

Daphni and her team of AI engineers are developing a deep learning model based on a well-known architecture. This model is trained on multiple datasets using different training techniques, resulting in various model variants. Each engineer is responsible for documenting and managing specific sections of the model card related to their variant, which reports on aspects such as the dataset used, training approach, and performance results. Using the transparency service, the team collaborates on creating and managing these interrelated sections. The service provides a centralized platform where each variant of the model card is versioned and updated as needed. When Sarah, one of the engineers, fine-tunes her model variant based on user feedback, she easily updates the corresponding section of the model card, focusing on improvements such as better generalization or reduced overfitting. The service ensures that all changes are documented clearly, allowing the team to keep track of the model's evolution and maintain transparency across the different variants. This streamlined versioning and collaboration process helps the team work efficiently while ensuring that the model documentation remains accurate and up to date.

9.4 User Stories

User Story #1	Documenting an AI Model Using the Transparency Service
Type of user	AI Engineer (Creator)
Action	As an AI engineer , I want to use a transparency service to document my model with relevant evaluation metrics and ethical considerations so that others can easily understand its performance and implications.
List of Use Cases	• #1.1: Documenting a Model with Evaluation Metrics. • #1.2: Facilitate Addition of Ethical Considerations to a Model Card. • #1.3: Sharing Model Documentation in various formats.
User Story #2	Interpreting AI Results Through Detailed Model Cards
Type of user	Fact-checker
Action	As a fact-checker , I want to review detailed sections of model cards of several detection models so that I can understand the reasons behind conflicting results and make an informed decision. While model cards provide different levels of detail, I specifically need access to the in-depth information relevant to my analysis.
List of Use Cases	• #2.1: Navigating to the AI Model and Model Card Section. • #2.2: Selecting and Viewing a Detailed Model Card at multiple levels of detail. • #2.3: Comparing AI Models through their Model Cards. • #2.4 Reviewing an AI Model Limitations and Intended Use.
User Story #3	Generating Simplified Model Cards for Effective Communication of AI Insights
Type of user	AI Expert
Action	As a professional familiar with AI , such as a journalist or researcher, I want to use the transparency service to take an existing model card and, with the guidance of the

D3.1 User-centred requirements definition

	service and my own expertise, generate a simplified version . This summary will be tailored for non-expert users, allowing me to effectively communicate key insights to colleagues with varying levels of technical knowledge.
List of Use Cases	
	• #3.1: Suggesting a Simplified Model Cards for non-experts. • #3.2: Customizing Model Card Content by selecting model card fields or using the suggested ones. • #3.3: Generating a summarized version of each selected field. • #3.4: Sharing Simplified Model Cards in various formats, such as PDFs.
User Story #4	Evaluating AI Model Performance Using Model Cards
Type of user	Software engineer
Action	As a software engineer , I want to review detailed model cards for AI models so that I can assess their performance across different datasets and aspects to ensure they meet the specific needs of my application.
List of Use Cases	
	• #4.1: Evaluate AI model performance across datasets and task-relevant subsets defined by input types, categories, or other factors specified by the AI engineer or suggested by the transparency service. • #4.2: Analysing Model Performance Across Different Scenarios by inspecting performance across all relevant factors.
User Story #5	Integrating Machine-Readable Model Cards for Enhanced AI Output Transparency
Type of user	Software engineer
Action	As a software engineer , I want to integrate machine-readable model cards into my application so that I can automatically provide real-time, detailed information about AI model outputs to improve user understanding and trust.
List of Use Cases	
	• #5.1: Integrating Machine-Readable Model Cards in structured formats such as JSON, to 3rd party applications that use the associated Models. • #5.2: Automatic Display of Detailed Information for the results of AI Models through the information included in the Machine-Readable Model Cards.
User Story #6	Collaborative Management and Versioning of Model Cards for Deep Learning Model Development
Type of user	AI Engineer
Action	As an AI engineer , I want to manage and update different instances of a model card using the transparency service so that I can document changes and improvements across different model versions efficiently and transparently.
List of Use Cases	
	• #6.1: <i>Updating Model Cards with model improvements and changes</i> by documenting new performance metrics, architecture changes, and additional information whenever the model is improved or modified.. • #6.2: <i>Tracking Model Evolution and Versions</i> by logging changes across model versions, enabling users to view a detailed history of modifications, performance change and updates in the model card repository.

9.5 Early Requirements

#	Early high-level functional requirements	How will the requirement be addressed in the service design?
1	Provide standardized model card templates with fields for performance metrics, model descriptions, ethical considerations, etc.	We have defined and will continue to revise a flexible model card schema that includes all the necessary fields. The schema is designed to be modular and extensible, allowing for

D3.1 User-centred requirements definition

		easy updates and the addition of new fields as new needs and regulation requirements emerge.
2	Offer an SDK that integrates into developer environments to allow code annotations that automatically generate model performance metrics, evaluation results, and documentation.	The SDK will be a Python library compatible with popular AI frameworks, designed using best practices for easy integration into existing pipelines. It will provide annotation utilities to automatically capture performance metrics, evaluation results, and relevant metadata during model development.
3	Automatically suggest appropriate evaluation metrics based on the AI model type, dataset, or architecture.	We will implement this feature by defining an extensive and extensible mapping between common AI/ML tasks and appropriate evaluation metrics, based on established practices in scientific literature.
4	Provide different levels of granularity (technical and non-technical) in model cards, catering to developers, engineers, researchers, and non-expert stakeholders. The service will allow users to toggle between simplified summaries and full technical details.	To determine the appropriate level of detail for each user group, we will collect user feedback through workshops and demos, ensuring that the model card content is both relevant and accessible.
5	Provide a versioning system to track changes to model cards over time, including updates to performance metrics, algorithm details, and ethical considerations.	To support this functionality and ensure traceability of changes, we will design the system similarly to Git, using hash keys and Merkle Directed Acyclic Graphs (DAGs) to track updates and version history of model card content while providing a method to compare differences between them.
6	Support the ability to update model cards as models evolve (e.g., new datasets, fine-tuning) while maintaining previous versions for reference.	-
7	Provide advanced chatbot assistants to guide users in filling out model cards. Provide explanations for complex sections, such as ethical considerations or performance metrics, and offer suggestions to the user.	To ensure accuracy and reliability, the underlying models will be evaluated and fine-tuned to minimize hallucinations and operate within a narrow scope. A human-in-the-loop approach will ensure that users review and confirm all generated content before finalization.
8	Offer detailed guidance on how to evaluate and document ethical aspects, such as bias, fairness, or potential harm.	We will work closely with CEPS and incorporate feedback from the Ethics Advisory Board to map the landscape of relevant ethical aspects such as bias, fairness, and potential harm. This will inform the development of detailed guidance and structured fields within model card templates.
9	Support exporting model cards with different levels of detail to different formats.	We will design an export mechanism that leverages existing rendering engines. The system will allow users to select the desired format and detail level, enabling the generation of tailored outputs.
#	Early high-level non-functional requirements	How will the requirement be addressed in the service design?

D3.1 User-centred requirements definition

1	Support for various programming languages and machine learning frameworks (e.g., Python, TensorFlow, PyTorch).	We will design the SDK to be framework-agnostic, with native support for popular AI/ML frameworks.
2	The UI must be user-friendly and intuitive, with clear instructions and tooltips to guide users through creating and managing model cards.	The UI will be designed with a focus on usability, incorporating clear instructions, tooltips, and step-by-step guidance to help users easily create and manage model cards.
3	Support standard file formats for export (e.g., JSON, PDF, HTML) to allow model cards to be used across various environments.	The underlying data model will be format-agnostic, allowing model card content to be easily transformed into standard formats.
4	The application shall provide differentiated support for beginner and expert users as two main types of user cohorts within the target user group of the application.	The service by design will support a wide spectrum of consumers by providing several versions of a model card with varying complexities. As such a beginner can choose the appropriate version that will offer simplified explanations of important aspects such as ethical risks and limitations. On the other hand, the expert user can choose the advanced version which offers all the insights that s/he needs like model architecture, hardware used during training etc.

9.6 Ethical Requirements

#	Ethical requirements	How will the requirement be addressed in the service design?
HAO-02	The capabilities of the application shall be communicated clearly to the user.	The service will include clear documentation, an onboarding guidance, and in-app explanations and tooltips.
PDG-01	The application shall minimize the amount of data collected on users.	The application will collect minimal user data, limited to basic account information needed for registration and authentication.
TRA-01	The system input and output of the application shall be traceable through documentation.	The system will automatically log and document inputs and outputs, linking them to specific model card versions and user actions, ensuring full traceability.
TRA-02	The decisions of the application shall be understandable by the user (XAI) as far as reasonably achievable.	All functionalities of the service will be extensively analysed in a documentation including pure algorithmic and AI decisions. Furthermore, The Model Card is designed to offer transparent model decision making through fields such as usage, limitations, performance under different scenarios, train/test dataset insights and more.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system.	The application will clearly indicate when suggestions or summaries are generated by AI assistants, and for AI-generated fields, it will provide transparency into the inputs used and explain how the outputs were derived.
ACC-	The application is designed so that it is auditable to	Model Cards fields such as training set,

D3.1 User-centred requirements definition

01	third parties i.e., the development process and sourcing of training data are traceable and the AI system's process, outcomes and impacts are logged consistently.	considerations, ethical risks, usage, and limitations aim to provide comprehensive model transparency during development process.
ENV-02	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects .	Since the transparency service documents models used in broader AI systems, it may indirectly affect individuals impacted by those systems. To address this, each model card will include a section identifying affected subjects, ensuring transparency around potential impacts.

9.7 Early Success Criteria

System Effectiveness benchmarking

- The model cards will have low error rates and minimal omissions in the automatically generated fields.

User evaluation

- Users will be able to efficiently create model cards.
- The model cards will effectively communicate the ethical implications (including the trustworthiness) of the models and services that they describe.
- Media professionals will trust the information provided through the service's model cards.

9.8 Early Mock-ups of the Service

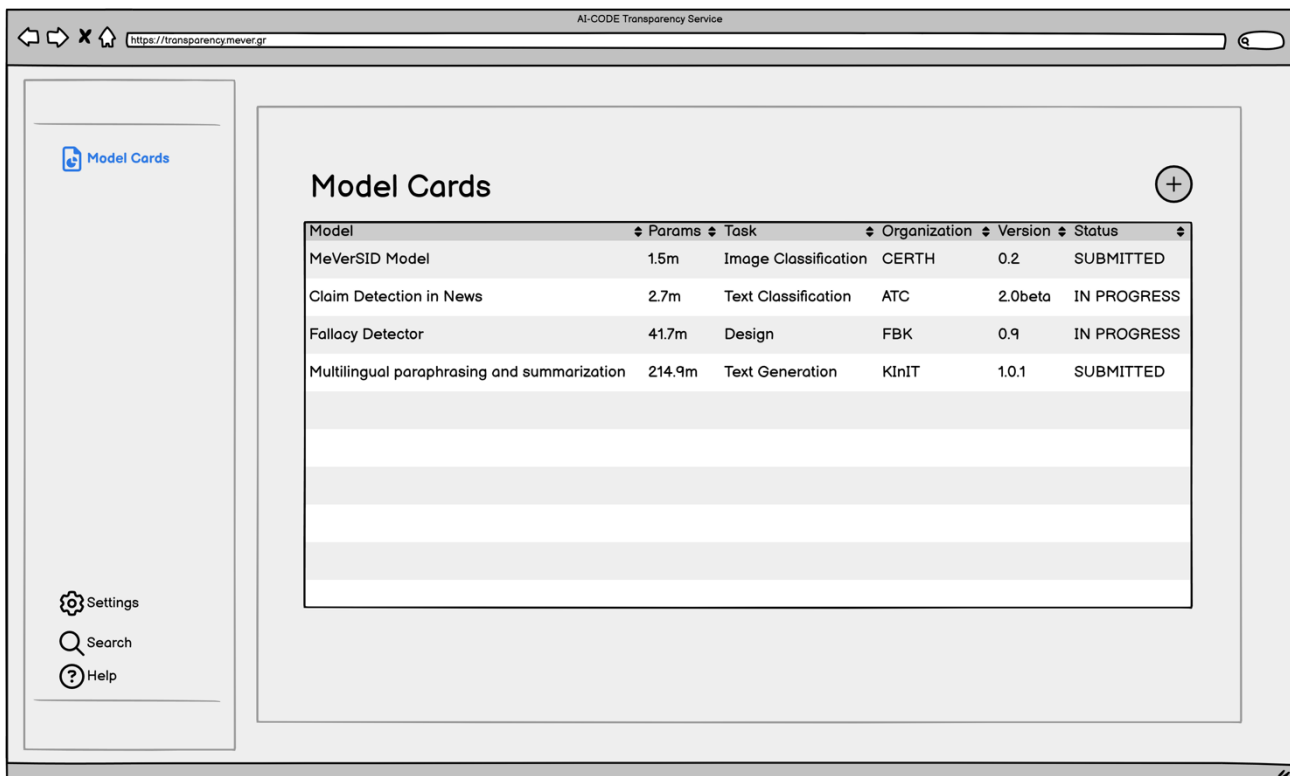


Figure 37: Model Cards view, with public model cards stored in the transparency service and shared between organizations.

D3.1 User-centred requirements definition

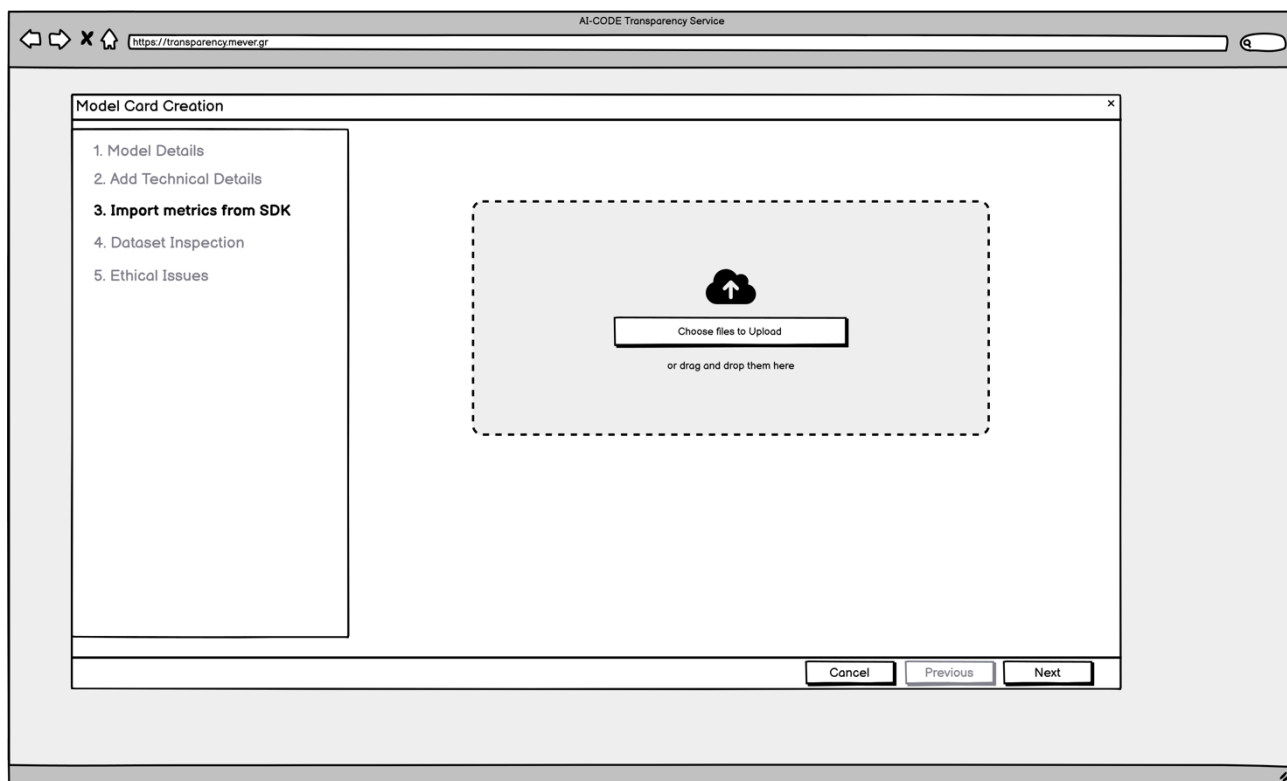


Figure 38: Add model performance metrics by uploading metrics file

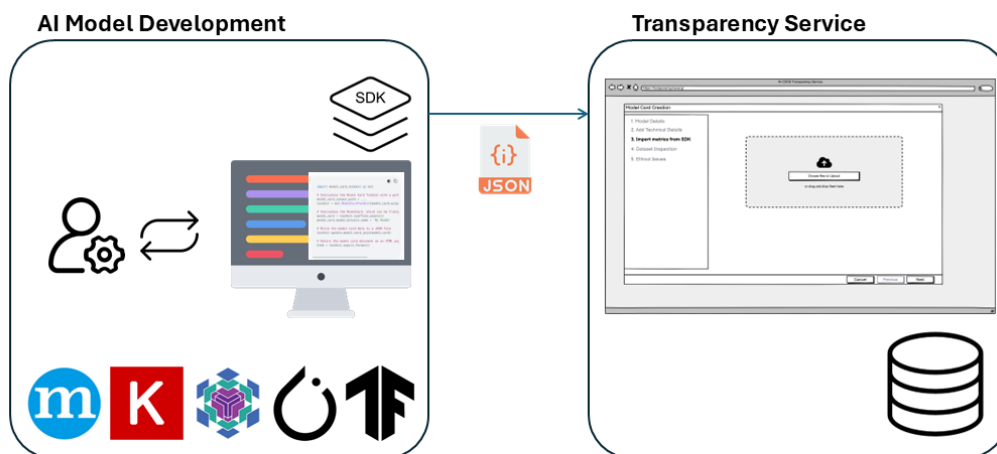


Figure 39: Interaction between the AI model development SDK and the transparency service. The SDK supports several AI frameworks and provides the results as a structured JSON file in the form required by the transparency service.

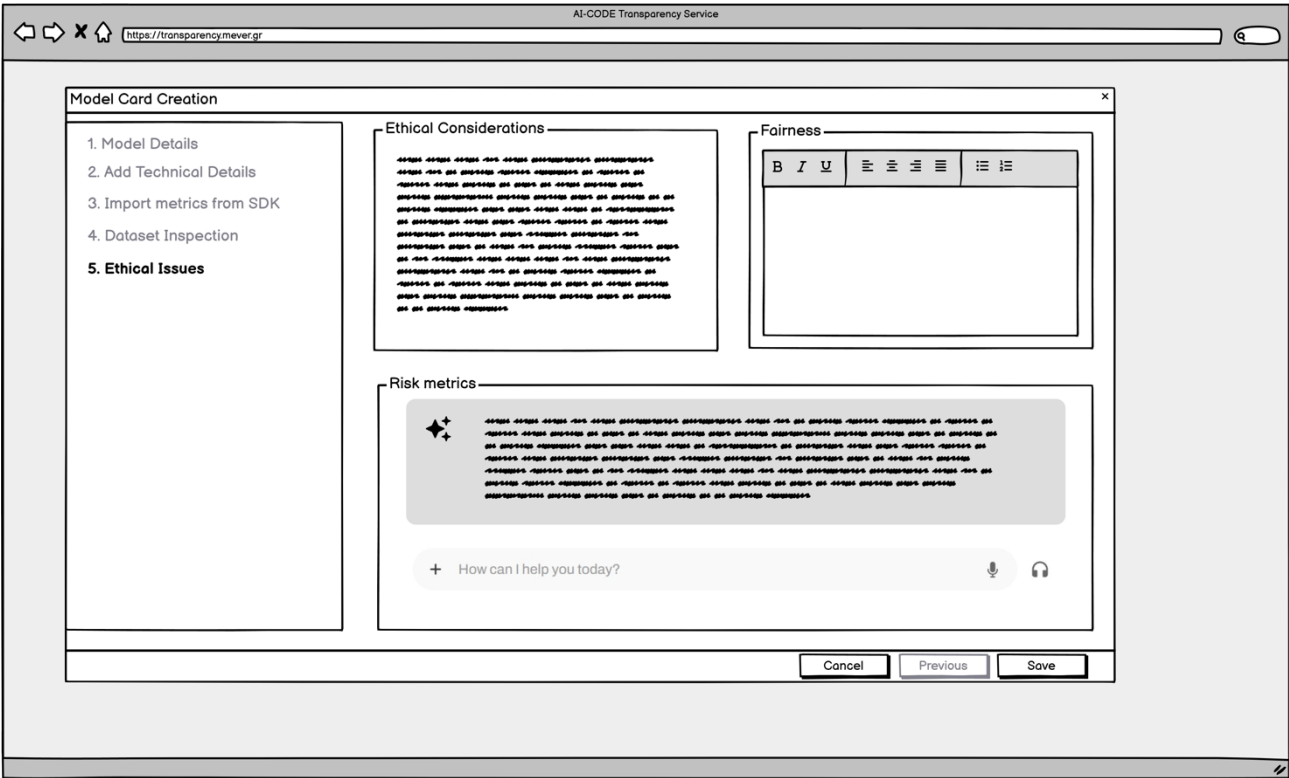


Figure 40: During the creation of a new model card, the user can provide an ethical assessment of the model. The transparency service facilitates this process with guidance provided by a chatbot.

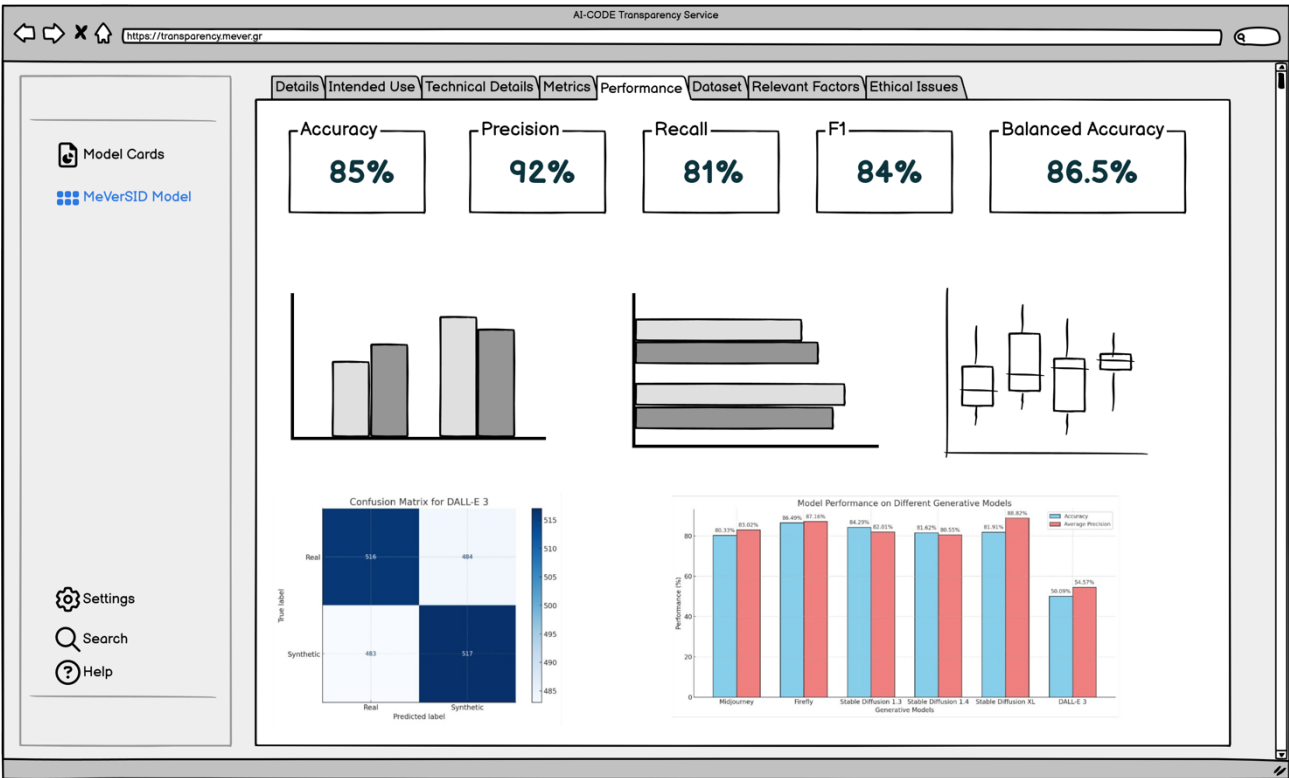


Figure 41: The Performance metrics tab of a model card.

10 Personal Companion for Understanding Disinformation for Media Professionals

10.1 Service Description

10.1.1 Scope and objectives

The Personal Companion for Understanding Disinformation is an LLM-based service that aims to assist junior, experienced journalists and fact-checkers to combat the spread of disinformation and misinformation, and, hence, improve the quality of media content. Specifically, it aims to: i) assist journalist students/in training to write articles without any logical fallacies, hate speech, and to not spread disinformation, ii) assist young journalists and media professionals to check the validity and credibility of the articles they consume, iii) produce reports on specific claims based on the websites/debunking reports they appear, iv) obtain a rough estimation of the validity of claims that have not been debunked based on the type of the media outlets they appear in, v) assist fact-checkers to search, maintain, and share their local debunking database, vi) understand the spread level of disinformation.

10.1.2 Supporting critical thinking

A service that automatically alerts users when claims in an article have already been debunked empowers individuals to question and reassess information that might otherwise be accepted uncritically. By surfacing alternative sources that discuss the same claim, the tool encourages comparison, source evaluation, and contextual understanding—core elements of critical thinking. Rather than passively consuming information, users are prompted to engage more actively, asking: *Is this true? What do others say? Why does this matter?*

Moreover, by detecting logical fallacies and instances of hate speech, the tool sharpens users' awareness of manipulative or harmful rhetorical tactics. Identifying fallacies helps users recognize flawed reasoning and avoid being misled by emotionally charged but logically weak arguments. Highlighting specific types of hate speech not only fosters awareness of biased language and discriminatory framing, but also encourages reflection on the ethical and social impact of the content being consumed. Over time, this guided exposure builds analytical habits and resilience against manipulation, reinforcing users' ability to think critically, question assumptions, and make more informed judgments in both digital and real-world contexts.

10.2 Target Users and Expected Benefits

Target users	Expected Benefits
Primary target users	
Journalist students / journalists in training	• Help them to write articles without logical fallacies, subtle hate speech, and spread disinformation • Help them assess the credibility of website articles • Have a rough estimation of the credibility of a claim based on the type of websites that mostly appears in. • Train critical reading skills • Supports editorial decision making • As a learning tool to identify problematic content and understand editorial standards • To guide them toward reputable materials and avoid overwhelming them with choices
Journalists	• Help them assess the credibility of website articles or claims

D3.1 User-centred requirements definition

	• To automatically produce reports on specific claims based on the websites/debunking reports they appear • Have a rough estimation of the credibility of a claim based on the type of websites that mostly appears in. • Useful for moderation or compliance
Editors	• Contribute to the evolving integrity of the database—fostering professional accountability and transparency • To prioritize sources and integrate results into editorial review. • To assign or flag content
Fact-checkers	• Search and maintain their local debunking database • Integrate external debunking databases into local database • Have a first rough estimation of the credibility of a claim based on the type of websites that mostly appears in and other credibility signals (e.g., hate speech) • Observe the spread level of false claims • Contribute to the evolving integrity of the database—fostering professional accountability and transparency
Media asset managers	• Will allow to quickly assess content risk (e.g., presence of hate speech or logical inconsistencies) before publication or archival
Social Media Managers	• Real-time claim verification can be used as a fast tool for checking viral or trending content. • Hate speech detection could be useful for moderating user-generated content
Main stakeholders	
Media organizations with fact-checking departments	• Improved efficiency in producing and publishing trustworthy content by better informed personnel
Fact-checking organizations	• Improved efficiency in identifying false claims

10.3 Scenarios of Use

Scenario of use #1

Rachel, a new journalist, is interested in exploring AI solutions in writing credible articles. Rachel logs in to the system and begins by copying the article from the editor she is using to the textbox of the Personal Companion service. Then, she selects from the list of the functionalities on the left of her screen the functionalities: “Matching to debunked claims”, “Logical Fallacy Detection”, and “Hate Speech Detection”. After this, the system highlights various parts of the text. The highlighted parts that include claims that have already been debunked by her organization have a yellow clickable dot on the right side of the text. The highlighted parts that include claims with logical fallacies have an orange clickable dot, and the ones that include some type of hate speech have a red clickable dot. A highlighted text can be annotated with up to three clickable dots (yellow, orange, red) depending on the type of detection (debunked claim, logical fallacy, hate speech).

Rachel seeks further information on why a particular claim is detected as debunked. Therefore, she proceeds with clicking on the yellow dot appearing on the right of the claim. She sees when this claim first appeared and clicks on the URL in the pop-up window to read the debunking report and learn why the claim is false. After this, she returns to the previous page, where she removes the debunked claim, and she repeats this process until she has checked and removed all claims that have been matched to claims debunked from her organization.

Then, in the same way, she proceeds with the logical fallacies detected in the text. She clicks on the orange dot to learn about a particular logical fallacy. A pop-up window appears indicating the type of fallacy and providing a URL to an external resource (e.g., Wikipedia) providing further information about this type of

D3.1 User-centred requirements definition

fallacy. Based on the information that she reads, she corrects, saves, and reassesses her article. She iterates this process until all claims with logical fallacies have been processed and reassessed. Finally, she continues in the same way with the parts of the text that have red dots, i.e., that contain some form of hate speech: she learns about the type of hate speech by clicking on the red dot appearing next to that part of the text and corrects it accordingly.

Scenario of use #2

Mike is an experienced journalist who wants to assess the credibility of an article he reads and produce a report on a specific claim appearing in the article based on selected resources that contain similar claims. Mike logs in to the Personal Companion service and inserts the URL of the article that he reads into the textbox appearing at the top of the page. Then he sees the inserted article in the textbox below. He selects from the list of the functionalities on the left of the textbox the functionalities: “Matching to debunked claims” and “Matching claims to other articles”. After this, the system highlights various parts of the text. The highlighted parts that include claims that have already been debunked by his organization have a yellow clickable dot on the right side of the text, while the claims that are matched to other claims that appear in articles on the web are annotated with a blue clickable dot.

He observes that the article contains few claims that have already been debunked by fact-checkers in his organization. Also, he observes that there are claims that have not been debunked but are annotated with a blue dot. He selects one of these claims and proceeds by clicking the dot where he sees in the pop-up window that similar claims have appeared in media outlets that he does not consider credible. In this way, he concludes that there is a high chance that the claim is fake. He selects the functionality “Produce AI-Generated Report” and some of URLs that appear in the pop-up window. The automatically generated report based on the selected resources is presented in a new window. Mike adjusts the report to his needs (e.g., to comply with IPR rules) and saves it. He continues with the rest of the annotated claims appearing in the article.

Scenario of use #3

Peter is an editor of a media outlet, has a hint about a specific topic, and wants to check if it is a trend. Peter writes in the textbox of the main page of the service the content that is interested in investigating and selects the functionality “Matching claims to other articles”. He observes in the pop-up window that similar claims have appeared in numerous website articles and understands that this content is trending. Then he selects the functionality “Matching claims to other articles” to automatically generate a report on the various opinions on the trending topic based on selected resources.

Scenario of use #4

Andy is a fact-checker and wants to assess the impact of a false claim and identify disinformation spreaders. Andy is using the Personal Companion service for the first time. He initially uploads all the fact-checks that he has done by clicking on the “Debunked Claims” at the top menu bar and then selecting the “Load database”. The user then uploads his .csv file. The new claims are instantly inserted into the shared database with the debunked claims (which also contains debunked claims from other fact-checkers in his organization). Then, he copies a false claim in the textbox of the main page of the service and selects the functionality “Matching claims to other articles”. He observes in the pop-up window that similar claims have appeared in numerous website articles and understands that this claim has a high impact. He copies the domain of one of these websites and pastes it to the search engine of the “Debunked Claims” database. The searcher returns a significant number of results. Andy understands that he has found a disinformation spreader.

10.4 User Stories

User Story #1	Title: Fallacy detection for own articles
Type of user	Journalist students / journalists in training
Action	As a journalist student/in training, I want to check if the article that I write contains any logical fallacies so that I can ensure my arguments are sound and persuasive.

D3.1 User-centred requirements definition

	As a media asset manager/editor I want quickly assess content risk (e.g., presence of hate speech or logical inconsistencies) of an article written by the organization I work in before publication or archival.
List of Use Cases	
	• #1.1: Enter own text for assessment • #1.2: Navigate checklist with the various functionalities • #1.3 Select “Logical Fallacy Detection” and explore the results
User Story #2	Title: Fallacy detection for website articles
Type of user	Journalist students / journalists in training, journalists, fact-checkers, editors, media asset manager
Action	As a journalist student/in training, I want to check if the article that I read contains any logical fallacies so that I can assess its veracity and enhance my media literacy and critical thinking. As a journalist/fact checker, I want to quickly check the veracity of the article that I read.
List of Use Cases	
	• #2.1: Enter external text for assessment • #2.2: Navigate checklist with the various functionalities • #2.3: Select “Logical Fallacy Detection” and explore the results
User Story #3	Title: Hate speech detection for own articles
Type of user	Journalist students / journalists in training, media asset manager, editor
Action	As a journalist student/in training, I want to check if the article that I write contains any unintended hate speech so I do not spread hate speech. As a media asset manager/editor I want quickly assess content risk (e.g., presence of hate speech or logical inconsistencies) of an article written by the organization I work in before publication or archival.
List of Use Cases	
	• #3.1: Enter own text for assessment • #3.2: Navigate checklist with the various functionalities • #3.3: Select “Hate Speech Detection” and explore the results
User Story #4	Title: Hate speech detection for website articles
Type of user	Journalist students / journalists in training
Action	As a journalist student/in training, I want to check if the article that read contains any hate speech so that I can evaluate instantly the credibility of the authors.
Use Cases	
	• #4.1: Enter external text for assessment • #4.2: Navigate checklist with the various functionalities • #4.3: Select “Hate Speech Detection” and explore the results
User Story #5	Title: Credibility assessment of claims included in own article
Type of user	Journalist students / journalists in training, journalist
Action	As a journalist student / journalist in training/ journalist, I want to check if the article that I write contains any claims that have been debunked so that I do not accidentally spread disinformation. As a media asset manager/editor I want quickly assess content risk (e.g., presence of hate speech or logical inconsistencies) of an article written by the organization I work in before publication or archival.
List of Use Cases	

D3.1 User-centred requirements definition

• #5.1: Enter own text for assessment • #5.2: Navigate checklist with the various functionalities • #5.3: Select “Matching to debunked claims” and explore the results	
User Story #6	Title: Spread assessment of claims included in own article
Type of user	Journalist/Fact-checker
Action	As a journalist, I want to check if the claims contained in the article that I write are popular on the Web and in which articles have already appeared so that I can assess their novelty level and have a rough estimation of their veracity
List of Use Cases	
• #6.1: Enter external text for assessment • #6.2: Navigate check list with the various functionalities • #6.3: Select “Matching claims to other articles” and explore the results	
User Story #7	Title: Spreading assessment of claims
Type of user	Fact-checker
Action	As a fact-checker I want to assess the spread of a false claim so to understand its impact level, to communicate it to policymakers and regulators and to track trends more effectively.
List of Use Cases	
• #7.1: Enter own text for assessment • #7.2: Navigate check list with the various functionalities • #7.3: Select “Matching claims to other articles” and explore the results	
User Story #8	Title: Credibility assessment of claims included in website article
Type of user	Journalist students / journalists in training, journalist/fact checker
Action	As a journalist/fact checker, I want to check if the article that I read contains any claims that have been debunked so to quickly assess its veracity. As journalist students / journalists in training , I want to check if the article that I read contains any claims that have been debunked so increase my media literacy and critical thinking.
List of Use Cases	
• #8.1: Enter external text for assessment • #8.2: Navigate checklist with the various functionalities • #8.3: Select “Matching to debunked claims” and explore the results	
User Story #9	Title: Spreading assessment of false claims included in external articles
Type of user	Journalist
Action	As a journalist, I want to check the spread of the false claims that appear in a web article that I read so that I can better assess the credibility of the author of the article.
List of Use Cases	
• #9.1: Enter external text for assessment • #9.2: Navigate checklist with the various functionalities • #9.3: Select “Matching to debunked claims” • #9.4: Select “Matching claims to other articles” and explore the results	
User Story #10	Title: Automated generation of report on own claims
Type of user	Journalist/ Fact checker
Action	As a journalist or fact checker, I want to automatically produce a report for a claim that I have written based on selected website articles and selected debunking reports that it may appear so to improve my efficiency in generating such reports
List of Use Cases	
• #10.1: Enter own text for assessment • #10.2: Navigate checklist with the various functionalities	

D3.1 User-centred requirements definition

• #10.3: Select “Matching to debunked claims” • #10.4: Select “Matching claims to other articles” • #10.5: Select “Produce AI-Generated report”	
User Story #11	Title: Automated generation of report on claims appearing in external website articles
Type of user	Journalist/ Fact checker
Action	As a journalist or fact checker, I want to automatically produce a report for a claim that I have read in an article based on selected website articles and selected debunking reports that it may appear, so to improve my efficiency in generating such reports.
List of Use Cases	
• #11.1: Enter external text for assessment • #11.2: Navigate checklist with the various functionalities • #11.3: Select “Matching to debunked claims” • #11.4: Select “Matching claims to other articles” • #11.5: Select “Produce AI-Generated report”	
User Story #12	Title: Processing of database with debunked articles
Type of user	Editor / Fact checker
Action	As a journalist or fact checker, I want to be able to process my database with debunked articles so to be able to remove claims that accidentally labelled as “false”, add new claims and add own debunking reports.
List of Use Cases	
• #12.1: Navigate top-bar menu • #12.2: Explore “Debunked claims” menu	

10.5 Early Requirements

#	Early high-level functional requirements	How will the requirement be addressed in the service design?
1	The service should support also non-english languages. The more languages the better.	The services of Personal companion are LLM-based. LLMs already support multilingual contexts ⁹ . Additionally, the user will be able to select between Romanian and English for the service’s language.
2	The database with the debunked claims should be shared between people of the same organization	Role-based access control mechanism will be set.
3	The database should not be accessible by externals	Role-based access control mechanism will be set.
4	The system shall provide a toolbox of all functionalities at the initial page allowing users to select the ones they need.	The user will be able to select from four different functionalities, even multiple ones at the same time, if needed.
5	The system shall provide the possibility to the user to select as many functionalities as they need	Multiple functionality selections will be facilitated through multiple external API calls
6	The system shall provide a textbox where the users can copy-paste the text they want to process	The user will be able to insert an article they read or write by copying and pasting it into the textbox of the main page of the service
7	The system shall provide a textbox where the users can insert the URL of the external article they want to process	When a user uploads or inputs a link to an online article, the system retrieves the content of the page in real time. This is done via a secure HTTP request, after which the HTML

⁹ e.g., [Llama 3.3](#), supports 7 different languages

D3.1 User-centred requirements definition

		structure of the page is parsed using a content extraction library. The module isolates the main textual content (title, body, and date, if available), removing non-essential elements such as ads, navigation menus, and scripts to ensure a clean input for downstream processing.
8	The system shall provide a logical fallacy detection functionality which will detect logical fallacies in the text and provide the type of fallacy along with an informative link.	The fallacy detection tool will be integrated. The fallacy detection tool will: a) highlight the fallacious sentence(s), b) provide the type of fallacy, c) provide definition of the type of fallacy upon request.
9	The system shall provide a hate speech detection functionality which will detect hate speech in the text and provide the type of hate speech along.	The hate speech detection tool will be integrated. The hate speech detection tool will: a) highlight the sentence(s) with hate speech, b) provide the type of hate speech.
10	The system shall provide a claim detection and matching functionality that matches claims appearing in text with debunked claims stored in the database “Debunked Claims”, and provide a similarity score along with further information (date, origin) about the first similar claim inserted in the database.	The claim detection & matching to debunked claims tool will be integrated. The claim detection tool will: a) highlight the claims in the text that have already been debunked, b) provide the URL with the debunking report and a summary of the report, c) provide further metadata
11	The system shall provide a claim detection and matching functionality that matches claims appearing in text with claims appearing on external webpages.	The claim detection & matching tool to other web sources will be integrated. The claim detection tool will highlight the claims in the text that match to claims appearing in other resources, b) provides the relevant URLs
12	The system shall provide the matching claim results by providing external links and a similarity score.	The claim detection & matching tool to other web sources will be integrated provide the URLs of the matching resources. To provide the similarity score, we will calculate the cosine similarity of the embeddings.
13	The system shall generate automatically a report providing selected resources.	This will be facilitated by taking as input the provided resource, their content and prompting in suitable manner the LLM to generate a report.
14	Based on role, the system shall allow the user to provide specific external resources for similarity checking	When the user selects the “Matching claims to other articles” functionality, then s/he will be provided the option to insert (prespecified number of) resources for searching. Once these resources are provided, the “Matching claims to other articles” functionality will search only these resources.
15	Based on role, the system shall allow the user to navigate and process the database with the debunked claims	The database will be visible to the user through a link which will direct them to the page with the debunked claims. A textbox will be in visible place, for text insertion and searching.

#	Early high-level non-functional requirements	How will the requirement be addressed in the service design?
1	The application shall be well documented and described	The service will be documented, with instructions of use, description of the various functionalities and their limitations. Also

D3.1 User-centred requirements definition

		transparency will be facilitated through the model cards structured for T6.2.
2	The application shall be easy to use	The application will be responsive, basic accessibility is ensured aligned with Web Content Accessibility Guidelines (WCAG) 2.1. Best practices in UI/UX design will be followed to ensure an intuitive, user-friendly interface.
3	The application shall be easy to learn	Instructions of use will be added
4	The application shall be used in a browser	Single-page application will be accessed through browser.
5	The application shall provide differentiated support for beginner and expert users as two main types of user cohorts within the target user group of the application.	For beginner users, the application will include built-in tutorials, FAQs, and contextual help to guide them through key features. It will also provide clear definitions of relevant terms (e.g., "What is an appeal to emotion?"). For expert users, the contextual help will be more concise, with access to research papers and technical documentation for deeper insights. Additionally, the application will offer comprehensive API documentation to support automation and system integration.

10.6 Ethical Requirements

#	Ethical requirements	How will the requirement be addressed in the service design?
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system	This will be facilitated through the UI, where the service and its limitation will be described in detail
HAO-02	The capabilities of the application shall be communicated clearly to the user.	This will be facilitated through the UI, where the service and its limitation will be described in detail
PDG-01	The application shall minimize the amount of data collected on users.	The collected data will be limited to username and password in the sign-up. Additionally, based on FAI-03 requirement the service will also collect the feedback provided by each user.
TRA-02	The decisions of the application shall be understandable by the user (XAI) as far as reasonably achievable.	All functionalities will be explained to the user in the instructions of use. Also, the user will be enabled to cross check by providing the output in an analytic way, i.e. by highlighting the claim and providing the type of fallacy/hate speech/ the matching claim. Also, when the user the definition of the detected type of hate speech/fallacy will provided further understanding of the results.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system.	A description card will be added to describe all mentioned requirements
FAI-03	The application shall have a mechanism that allows for flagging issues related to bias, discrimination or poor performance.	Feedback mechanism will be incorporated collected input text, output of service and feedback from user.

D3.1 User-centred requirements definition

HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case	The human has full oversight as each output is presented in detail. Also the service is limited to detecting disinformation/misinformation, hate speech, fallacies. It does not proceed to change anything in the assessed article.
---------------	--	---

10.7 Early Success Criteria

System Effectiveness benchmarking

- The service will identify debunked claims, certain types of logical fallacies, and hate speech in submitted sources with enhanced accuracy.

User evaluation

- Media professionals will trust the service's performance in identifying debunked claims, certain types of logical fallacies, and hate speech in submitted sources.
- The reports will help media professionals to efficiently assess/analyse what has been written in external sources in relation to/about a claim
- Media professionals will comprehend the output of the service with respect to debunked claims, logical fallacies, and hate speech and be satisfied with that output.
- Media professionals will easily utilize the service to improve the efficiency of trusted information production.

10.8 Early Mock-ups of the Service

The screenshot shows a web application interface. At the top, there's a navigation bar with 'Home', 'View', 'Insert', 'Debunked Claims', and a 'Sign out' button. The main content area has two input sections: 'Insert article URL:' with a text box containing 'http://some-url.com' and an 'Upload' button, and 'Copy-paste text:' with a larger text box containing a news article snippet about a conflict in Gaza. To the left of the text input is a 'Functionality' sidebar with five options, each with a colored circle and a checkbox: 'Matching to debunked claims' (yellow), 'Matching claims to other articles' (blue), 'Logical Fallacy Detection' (orange), 'Hate speech Detection' (red), and 'Produce AI-Generated Report' (purple). Below the text input are 'Save' and 'Copy' buttons. The footer contains the European Union logo, funding information, and social media icons.

Figure 42: Welcome screen with the various functionalities to choose from and the text boxes to either insert a URL (top textbox) or a text (bottom textbox)

D3.1 User-centred requirements definition

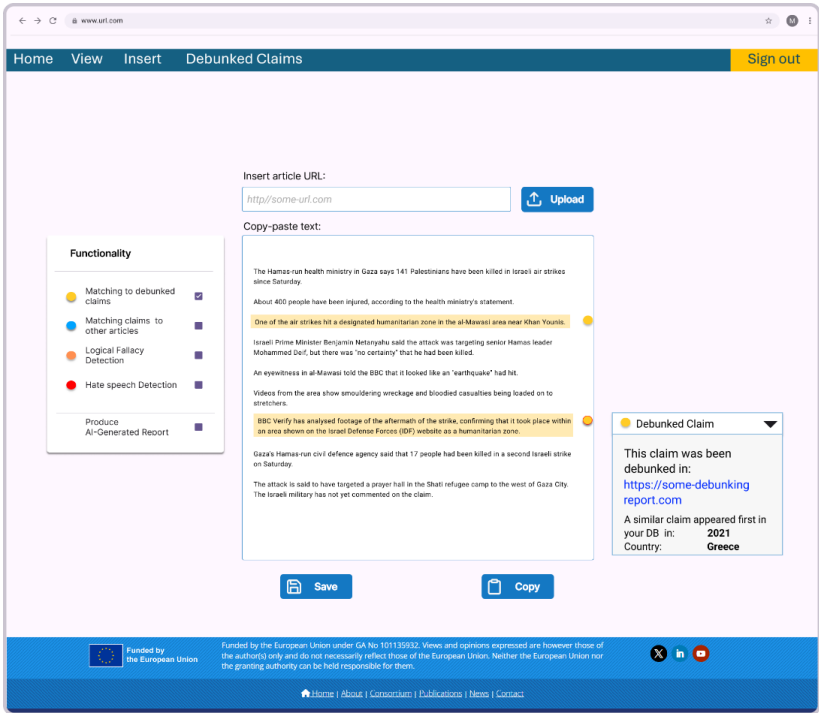


Figure 43: The sentences containing debunked claims are highlighted when the "Matching to debunked claims" functionality is selected. Upon selection, further information appears in a textbox.

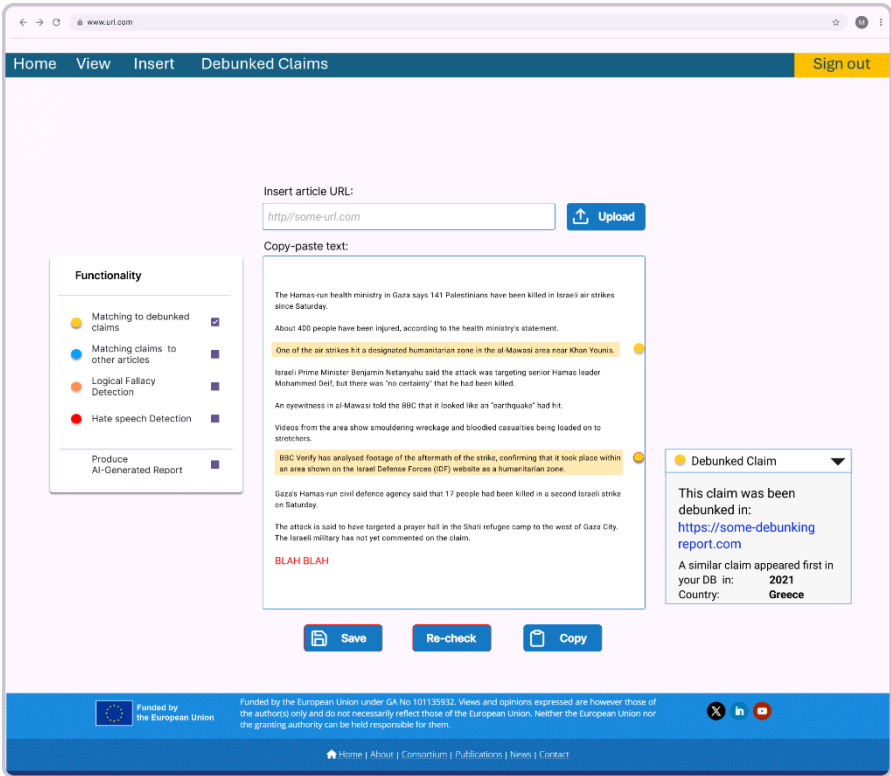


Figure 44: The user can edit the text in the textbox and select "Re-check" to recheck the updated text

D3.1 User-centred requirements definition

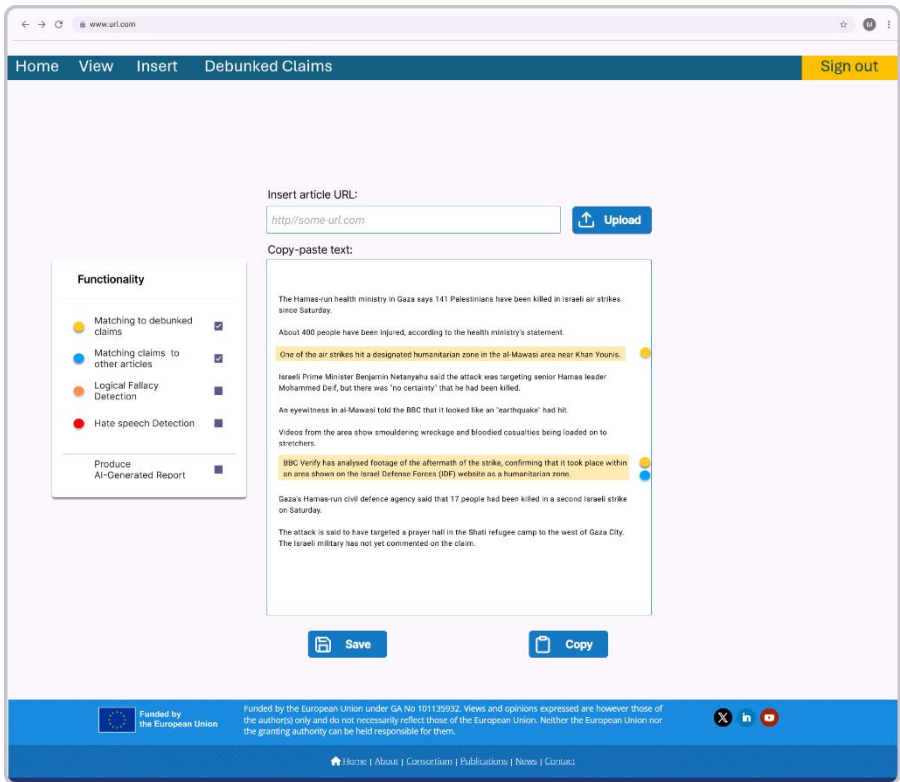


Figure 45: The sentences containing debunked claims (yellow dot) or similar claims mentioned in other articles (blue dot) on the web are highlighted when the functionalities "Matching to debunked claims" and "Matching claims to other articles" are selected.

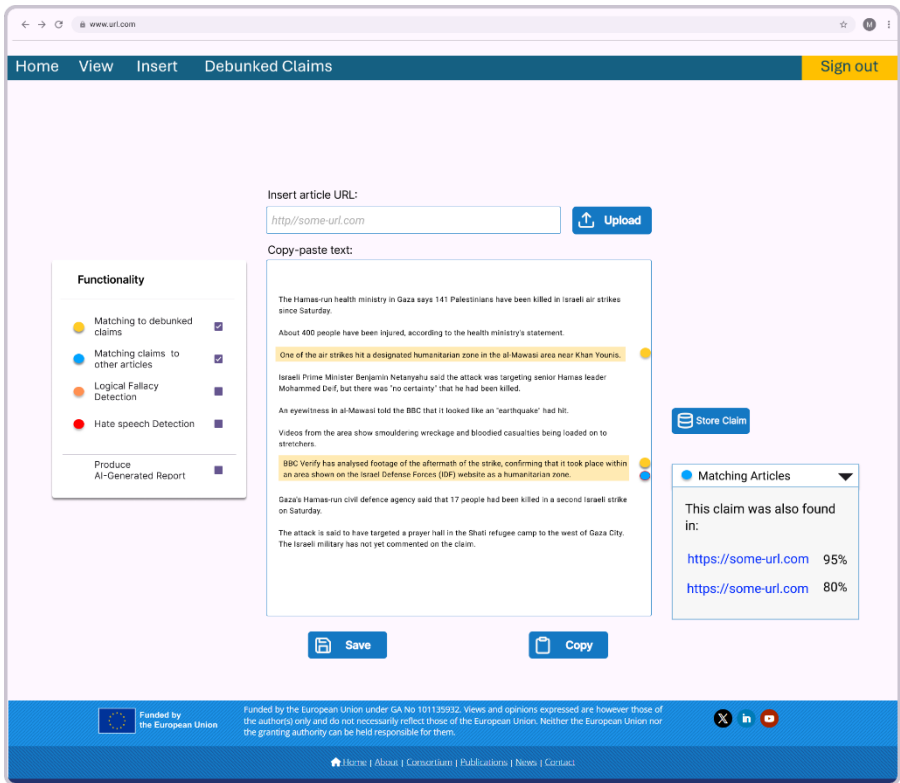


Figure 46: The user selects the blue dot to obtain further information about in which other articles the claim appears along with the similarity score. Also, they can store the claim to the DB with debunked claims by clicking the "Store claim" button.

D3.1 User-centred requirements definition

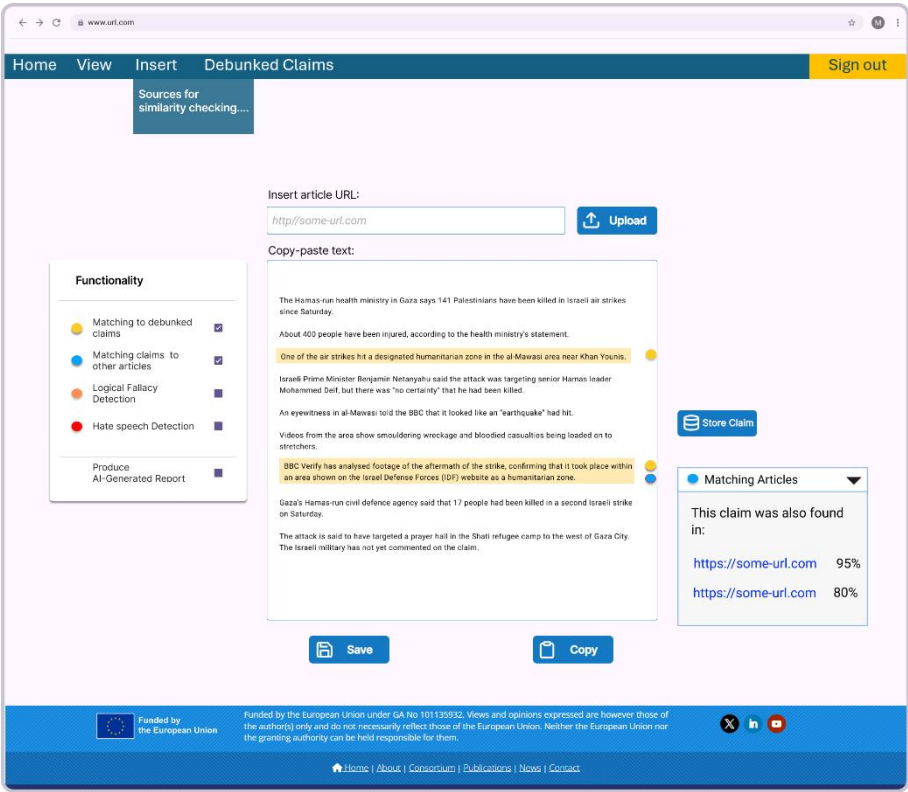


Figure 47: The user can select from the top menu bar to provide specific resources for similarity checking.

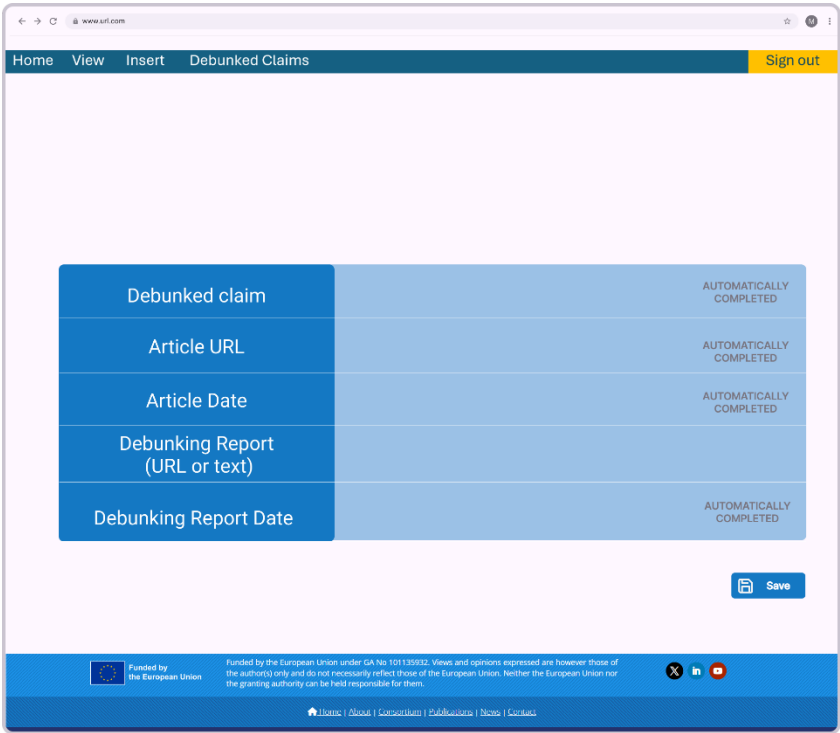


Figure 48: Once the "Store claim" button has been clicked, the table of the figure appears, where some of the fields are automatically completed while the user has to write a debunking report or provide a URL of a debunking report for the respective claim.

D3.1 User-centred requirements definition

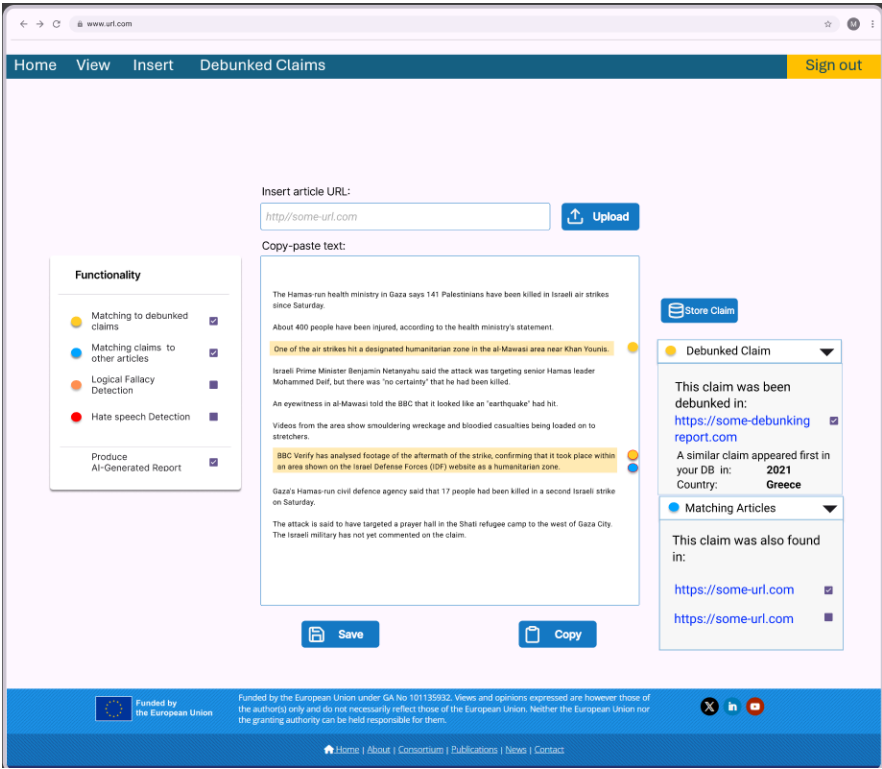


Figure 49: Once the "Produce AI-Generated Report" has been selected, checkboxes appear next to the matching resources, which the user can select.

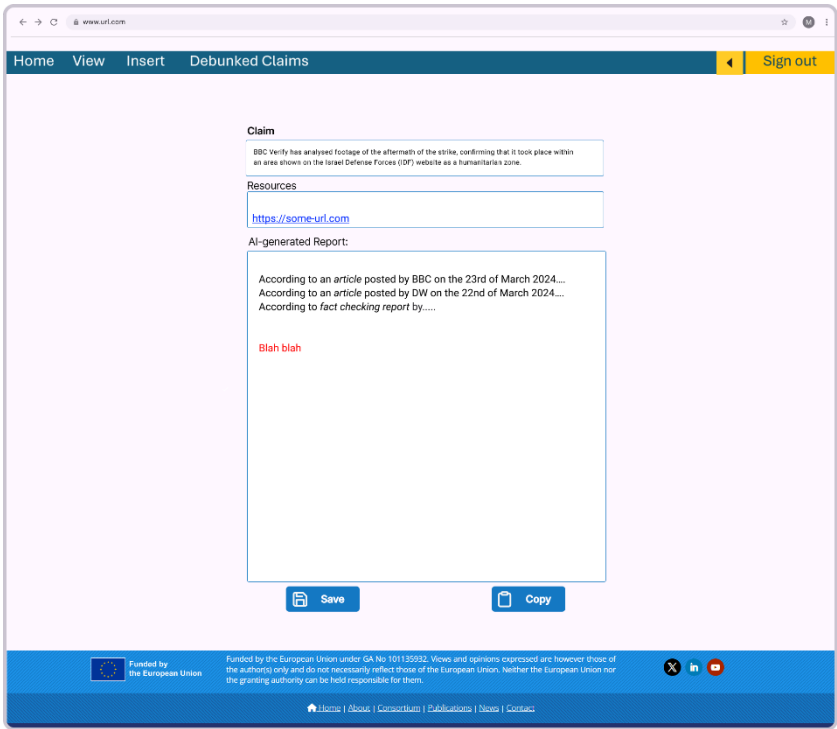


Figure 50: Based on the selected resources (Figure 8) and the selected claim, the system automatically generates a report which is editable.

D3.1 User-centred requirements definition

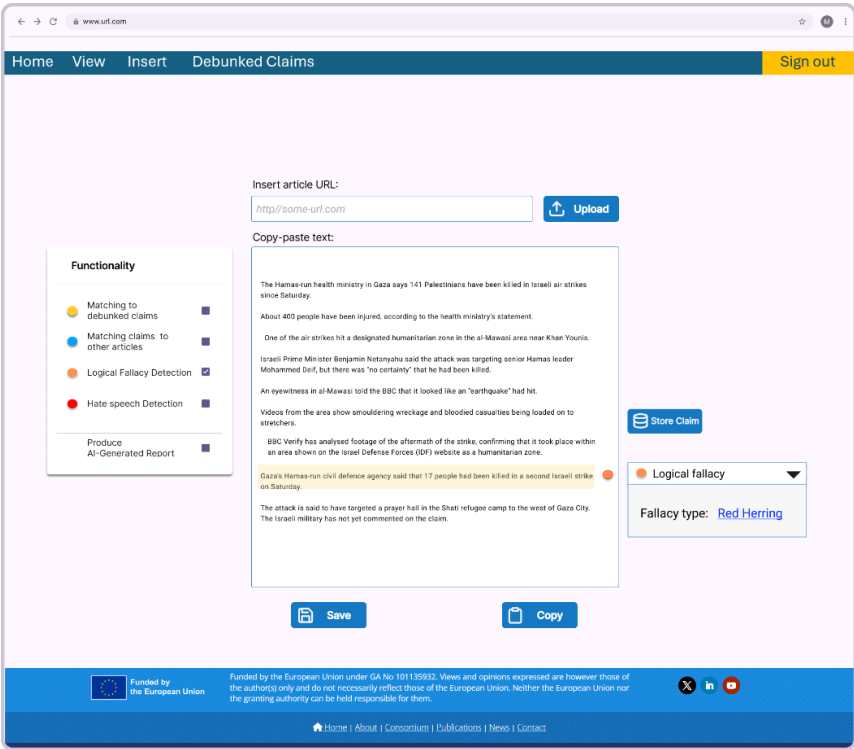


Figure 51: The sentences containing logical fallacies are highlighted when the "Logical Fallacy Detection" functionality is selected. Upon selection, further information about the type of fallacy appears in a textbox.

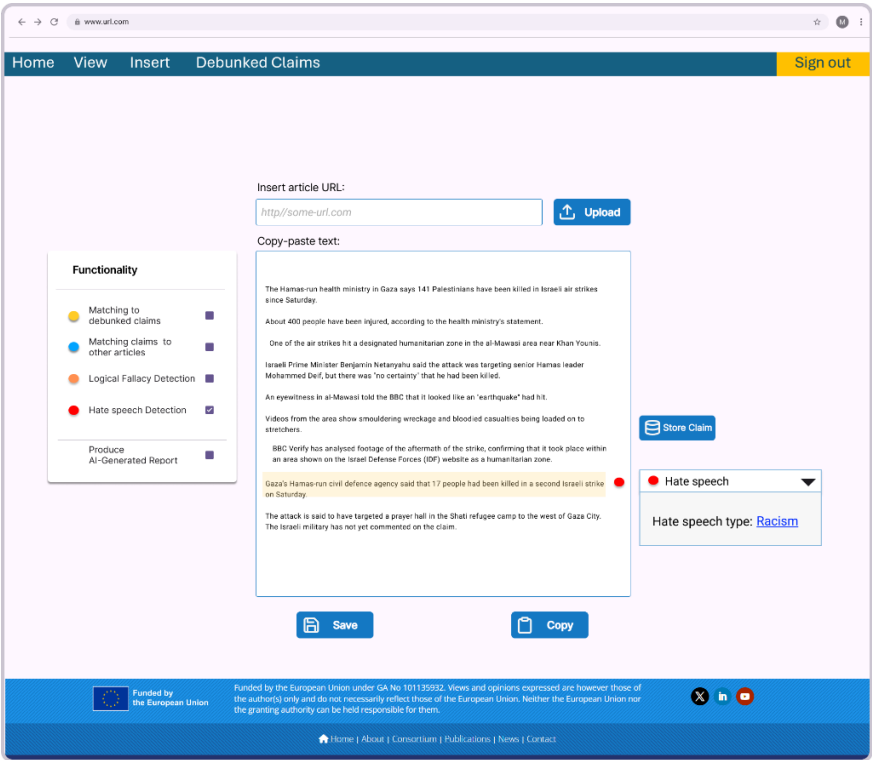


Figure 52: The sentences containing hate speech are highlighted when the "Hate Speech Detection" functionality is selected. Upon selection, further information about the type of hate speech appears in a textbox.

D3.1 User-centred requirements definition



Figure 53: The user can navigate and process the database with the debunked claims.

11 Conclusions and Next Steps

The user-centred requirements analysis for the AI-CODE project has provided valuable insights into the needs and expectations of media professionals regarding AI-powered tools for combating disinformation. The iterative process, combining exploratory interviews, literature analysis, and co-creation workshops, has yielded a comprehensive set of preliminary requirements for each of the six AI-CODE services.

Key outcomes of this initial phase include:

1. Defined target user groups for each service, validated through user feedback
2. Initial scenarios of use and user stories, offering context-rich descriptions of service applications
3. Early functional and non-functional requirements, outlining the core capabilities and performance expectations for each service
4. Ethical requirements, based on established frameworks and tailored to the specific challenges of AI in media contexts

The mock-ups and prototypes developed during this phase have proven effective in eliciting user feedback and refining the proposed functionalities. The co-creation workshops have been particularly valuable in ensuring that the services align with real-world needs and workflows of media professionals.

While these results provide a solid foundation for the project's next phases, it is important to note their preliminary nature. Further refinement and specification are planned, particularly in the upcoming deliverable D3.2 “Functional specifications and technical requirements” (M16), which will provide more detailed technical requirements and formal use case specifications. Notably, another round of co-creation workshops will be conducted with more mature iterations of service prototypes in Q1 of Y2. This will allow for more in-depth user feedback on more developed versions of the services.

The technical requirements to be reported in D3.2 will build upon the detailed use case descriptions established in this initial phase. Based on these use cases, the technical specifications for each service will be elaborated, ensuring a direct link between user needs and technical implementation.

The user-centred approach adopted by the AI-CODE project demonstrates a commitment to developing tools that are not only technologically advanced but also practical and ethical in their application. This balance is crucial in addressing the complex challenges of disinformation in the digital age.

Moving forward, continued engagement with end-users and ethical considerations will be essential to ensure that the final AI-CODE services meet the evolving needs of media professionals while upholding the highest standards of responsible AI development and deployment. The iterative process, including the upcoming co-creation workshops and detailed technical specifications, will further refine and enhance the services to better serve their intended purpose in combating disinformation.

12 References

- Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centred design. In W. Bainbridge (Ed.), *Encyclopedia of Human-Computer Interaction* (pp. 763-768). Thousand Oaks: Sage Publications.
- BBC (2024): *BBC AI Principles*. Available at: <https://www.bbc.co.uk/supplying/working-with-us/ai-principles/> (accessed 12.11.2024).
- Beckett, C., and M. Yaseen. 2023. *Generating Change. A Global Survey of What News Organisations are Doing with AI*. The London School of Economics and Political Science.
- Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. In *Proceedings of Machine Learning Research* 81:77-91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Bordot, F. (2022). ‘Artificial Intelligence, Robots and Unemployment: Evidence from OECD Countries’, *Journal of Innovation Economics & Management* N° 37, no. 1 (7 January 2022): 117–38, <https://doi.org/10.3917/jie.037.0117>;
- Cockburn, A. (2000). *Writing Effective Use Cases*. Addison-Wesley Professional.
- Coeckelbergh, M. (2020). ‘Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability’, *Science and Engineering Ethics* 26, no. 4 (August 2020): 2051–68, <https://doi.org/10.1007/s11948-019-00146-8>.
- Cools, H., & Diakopoulos, N. (2024). Uses of Generative AI in the Newsroom: Mapping Journalists’ Perceptions of Perils and Possibilities. *Journalism Practice*, 1–19. <https://doi.org/10.1080/17512786.2024.2394558>
- Cohn, M. (2004). *User Stories Applied: For Agile Software Development*. Addison-Wesley Professional.
- Corrêa, Nicholas Kluge et al. (2023). ‘Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance’, *Patterns* 4, no. 10 (October 2023): 100857, <https://doi.org/10.1016/j.patter.2023.100857>.
- Czech Radio (2024): *Czech Radio Rules for Using Artificial Intelligence*. Available via <https://www.ebu.ch/resources> (accessed 12.11.2024).
- Diakopoulos, N., Cools, H., Li, C., Helberger, N., Kung, E., Rinehart, A., and Gibbs, L. (2024). *Generative AI in Journalism: The Evolution of Newswork and Ethics in a Generative Information Ecosystem*. The Associated Press. <https://a-mcc.eu/en/library/studies-and-reports/generative-ai-in-journalism-the-evolution-of-newswork-and-ethics-in-a-generative-information-ecosystem/>
- EBU (2023): *AI Ethics Group Guidelines Repository*. Available at: <https://www.ebu.ch/groups/ai-ethics> (accessed 12.11.2024).
- European Commission (2019). European High-Level Expert Group on AI, ‘Ethics Guidelines for Trustworthy AI’, <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Frey, C.B. and Osborne, M. A. (2017). ‘The Future of Employment: How Susceptible Are Jobs to Computerisation?’, *Technological Forecasting and Social Change* 114 (January 2017): 254–80, <https://doi.org/10.1016/j.techfore.2016.08.019>.
- Hrckova, A., Moro, R., Srba, I., Simko, J., & Bielikova, M. (2022, November 22). *Autonomation, not Automation: Activities and Needs of Fact-checkers as a Basis for Designing Human-Centered AI Systems*. arXiv.org. <https://arxiv.org/abs/2211.12143>
- Jones, B., Luger, E., and Jones, R. (2023). *Generative AI & Journalism: A Rapid Risk-Based Review*. The University of Edinburgh.
- LRT (2024): *Public Enterprise Lithuanian National Radio and Television Guidelines for the Use of Artificial*

D3.1 User-centred requirements definition

Intelligence. Available via <https://www.ebu.ch/resources> (accessed 12.11.2024).

Kalluri, P. R. et al. (2023). 'The Surveillance AI Pipeline' (arXiv, 2023), <https://doi.org/10.48550/ARXIV.2309.15084>.

NRK (2023): *NRK's Guidelines for Editorial Use of Artificial Intelligence*. Available via <https://www.ebu.ch/resources> (accessed 12.11.2024).

RFE/RL (2023): *Radio Free Europe/Radio Liberty/ Standards of Ethical Journalism/ Use of Generative Artificial Intelligence*. Available at: <https://about.rferl.org/wp-content/uploads/2024/01/Editorial-Standards-EN.pdf> (accessed 18.11.2024).

Rosson & Carrol (2022). Scenario-based Design. Chapter 53 in J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. Lawrence Erlbaum Associates, 2002, pp. 1032-1050.

Simon, F. M. (2024). Artificial intelligence in the News: How AI retools, rationalizes, and reshapes journalism and the public arena. In *Columbia Journalism Review*. https://www.cjr.org/tow_center_reports/artificial-intelligence-in-the-news.php

SVT: *General policy for AI usage at SVT*. Available via <https://www.ebu.ch/resources> (accessed 12.11.2024).

SWR (2023): *SWR Editorial guidelines for "artificial intelligence"*. Available via <https://www.ebu.ch/resources> (accessed 12.11.2024).

ZDF (2023) *ZDF Guiding principles for the use of Generative AI* Available at: <https://www.zdf.de/zdfunternehmen/guiding-principles-generative-ai-100.html> (accessed 18.11.2024).

13 Appendix

Appendix 1: AI-CODE Pre-Interview Questionnaire

Thanks for taking a few minutes to answer this short questionnaire to prepare for your interview about your use of generative AI in your work as a media professional.

Your answers will help us guide the interview process to better fit your background and professional activities.

1. **Email ***

2. **Your name**

3. **Name of your organization and your professional role in the organization**

4. **In the course of your professional activities, do you use the following categories of generative AI tools?**

Tick all that apply.

- ☐ Large language models such as ChatGPT
- ☐ Image generation tools such as DALL-E
- ☐ Video generation tools such as Lumiere
- ☐ Audio generation tools (for example voice over generation tools)
- ☐ Other:

5. **In what ways do you encounter the problem of disinformation in your work?**

Tick all that apply.

- ☐ I need to identify disinformation
- ☐ I have to debunk disinformation
- ☐ I work on assessing trustworthiness and credibility
- ☐ Questions related to disinformation are rarely relevant to my work
- ☐ Other:

D3.1 User-centred requirements definition

6. To what extent are you experiencing problems with the following types of AI generated disinformation in your work?

Mark only one oval per row.

	Not at all	Very little	Somewhat	To a large extent	To a great extent
AI generated text content	☐	☐	☐	☐	☐
AI generated images	☐	☐	☐	☐	☐
AI generated video	☐	☐	☐	☐	☐
AI generated audio	☐	☐	☐	☐	☐

7. What social media platforms are relevant for you in your work related to identifying and countering disinformation?

8. To what extent are the following platforms relevant for your work related to identifying and countering disinformation?

Mark only one oval per row.

	Most relevant	Not relevant	Slightly relevant	Moderately relevant
Threads	☐	☐	☐	
BlueSky	☐	☐	☐	
Metta Horizon Worlds	☐	☐	☐	

D3.1 User-centred requirements definition

Very relevant	E		
	x	☐	
	t		
	r	☐	
	e	☐	
	m	☐	
	e		
	l	☐	
	y		
	r		
	e		
	l		
	e		
	v		
	a		
	n		
	t		
			☐
			☐
			☐
			☐

D3.1 User-centred requirements definition

9. Please indicate if you are currently using tools for any of the following purposes?

Tick all that apply.

- ☐ To understand how generative AI works
- ☐ To learn prompting techniques for achieving high-quality results from generative AI tools
- ☐ To understand a specific AI models' capabilities, limits, potential risks, and repercussions
- ☐ To detect disinformation in social media
- ☐ To learn strategies for how to identify disinformation
- ☐ To generate content that debunks disinformation
- ☐ To assess the trustability and credibility of content and sources
- ☐ To annotate content and detect deepfakes in the media asset collection of your organization

D3.1 User-centred requirements definition

10. Would it be relevant to your work if you had a tool designed for media professionals to help you:

Mark only one oval per row.

	Not relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
Understand how generative AI works	☐	☐	☐	☐	☐
Learn prompting techniques for achieving high-quality results from generative AI tools	☐	☐	☐	☐	☐
Understand a specific AI models' capabilities, limits, potential risks, and repercussions	☐	☐	☐	☐	☐
Detect disinformation in social media	☐	☐	☐	☐	☐
To generate content that debunks disinformation	☐	☐	☐	☐	☐
Learn strategies for how to identify disinformation	☐	☐	☐	☐	☐
Assess the trustability and credibility of content and sources	☐	☐	☐	☐	☐
Annotate content and detect deepfakes in your media asset collection	☐	☐	☐	☐	☐

11. **Would you be willing to share with us, during the interview, an example of something that you've created with the assistance of generative AI? If so, please have that available during our interview so we can discuss your workflow in creating it.**

Mark only one oval.

☐ Yes

☐ No

12. **Are there any particular concerns related to the use of generative AI that you would like to discuss with us in the interview?**
