

AI-CODE

AI services for COntinuous trust in emerging Digital Environments

D8.2 Validation Methodology

Grant Agreement no.

101135437

Project Title

AI-CODE - AI services for COntinuous trust in
emerging Digital Environments

Contractual Due date

30 November 2025 (M24)

Actual submission date:

28 November 2025

Lead beneficiary:

RU

Deliverable number

8.2

Deliverable Title:

Validation Methodology

Description:

This deliverable contains the methodology designed and developed to evaluate the AI-CODE services.

Type/Nature

- ☒ R - Document, report
- ☐ DEM - *Demonstrator, pilot, prototype*
- ☐ DEC - *Websites, patent filings, videos, etc.*
- ☐ OTHER
- ☐ ETHICS – *Ethics requirement*
- ☐ ORDP – *Open Research Data Pilot*

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

8

Work Package leader

ATC

Author(s)*in alphabetical order*

Bahia Albrecht
 Mathias Becker
 Randi Cecchine
 Viktoras Daukšas
 Paula Gurtler
 Riccardo Leoni
 Jasminko Novak
 Laura Seryte
 Annie Tang
 Laima Venclauskienė
 Ilona Wilmont

Internal Reviewers
in alphabetical order

Mathias Becker (EIPCM)
Mihaela Caramizaru (DEB)
Jasminko Novak (EIPCM)

AUTHORS

Name	Partner	E-Mail
Bahia Albrecht	DW	bahia.albrecht@dw.com
Mathias Becker	EIPCM	m.becker@eipcm.org
Randi Cecchine	RU	randi.cecchine@ru.nl
Viktoras Daukšas	Debunk	viktoras@debunk.org
Paula Gurtler	CEPS	paula.gurtler@ceps.eu
Riccardo Leoni	DST	r.leoni@dstech.it
Jasminko Novak	EIPCM	j.novak@eipcm.org
Laura Seryte	Debunk	laura.seryte@debunk.org
Annie Tang	EIPCM	a.tang@eipcm.org
Laima Venclauskienė	Debunk	laima@debunk.org
Ilona Wilmont	RU	ilona.wilmont@ru.nl

HISTORY OF CHANGES

Version	Date	Changes
0.1	26.9.2025	Table of Contents submitted
0.2	10.11.2025	Collated inputs received from authors in sections 3,4,5, and 6
0.3	13.11.2025	First draft submitted for internal review
0.4	24.11.2025	Revised draft based on all partner feedback
1.0	27.11.2025	Final version submitted to Quality Control

TABLE OF CONTENTS

1. Executive summary.....	6
2. Introduction.....	7
3. Foundational Principles of the Validation Methodology	8
3.1 User-Centred Design Approach	8
3.2 Ethical Framework for Trustworthy AI.....	9
3.3 Ethical Protocols: Anonymisation, Privacy, and GDPR Compliance	11
4. The Validation Methodology	12
4.1 Overview of the Validation Lifecycle	12
4.2 The Think Aloud Protocol (TAP)	13
4.3 The System Usability Scale (SUS)	14
4.4 Evaluation Criteria.....	16
4.4.1 Qualitative Evaluation Criteria.....	16
4.4.2 SUS Scores.....	17
4.5 User Testing Protocol.....	18
4.6 Validation Timeline and Milestones	19
5. Methodology for Phase 1: Formative Evaluation (User Testing + Co-creation Sessions)	21
5.1 Participants	21
5.2 Data Collection Methods	23
5.3 Data Analysis Methods	24
5.3.1 Qualitative Analysis: Think Aloud Protocol (TAP) and Co-creation Discussion	25
5.3.2 Quantitative Analysis (SUS).....	27
5.3.3 Mixed-Methods Findings Analysis	28
5.3.4 Findings Synthesis & Feedback Integration	30
6. Methodology for Phase 2: Summative Evaluation	30
6.1 Participants	31
6.2 Data Collection Methods	31
6.3 Data Analysis Methods	32
7. Conclusion.....	33
8. References	34
9. Appendix A: User Groups.....	35
10. Appendix B: Formative user test protocol.....	35

TABLE OF FIGURES

Figure 1: Validation timeline and milestones	21
Figure 2: Target user group prioritization per AI-CODE service	22
Figure 3: Mixed methods analysis.....	24

TABLE OF TABLES

Table 1: The System Usability Scale Standard Version (Lewis & Sauro, 2018) with notes on AI-CODE Modification	14
Table 2: Parameters utilized for each protocol variation	16
Table 3: SUS Testing phase and corresponding SUS scores.....	18
Table 4: Validation milestones.....	20
Table 5: Illustrative Codes for Each Qualitative Theme*	26
Table 6: Mapping framework for triangulating qualitative themes with SUS dimensions	28
Table 7: Conceptual structure of triangulation across analytical layers.....	29
Table 8: Summative Triangulation Structure	32

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, which includes:
 - a. Interactive Coaching and Simulation Module (ICSM)
 - b. Counter-narrative Recommendation Module (CNRM)
5. Transparency Service for AI Model Cards
6. Personal Companion for Understanding Disinformation

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

This deliverable outlines the user-centred validation methodology for the AI-CODE services. It covers the foundational principles of the validation methodology, including user-centred design, the ethical framework for Trustworthy AI and protocols for ensuring GDPR compliance. It describes the validation methodology, including an overview of the validation lifecycle which is performed in two phases of user testing: a formative phase (targeting five participants per service) aimed at gathering detailed feedback to improve the services, and a summative phase (targeting ten participants per service, totalling 70+ interactions) aimed at assessing overall usability, user acceptance and ethics. The test procedure in both the formative and summative phases is explained. The formative phase includes a co-creation discussion, while the summative phase includes an additional questionnaire aimed at assessing perceived effectiveness and efficiency and a short reflective conversation on user acceptance and ethics.

The deliverable introduces the research methods and instruments used including: the Think Aloud Protocol (TAP) and a modified System Usability Scale (SUS) tailored to the specific needs of the AI-CODE services. Then, the evaluation criteria are discussed, and an overview of the user test protocol is given, followed by the validation timeline, which spans from October 2025 (M24) to mid-2026 (M30). The methodology for the formative and summative phases is described in detail, including methods for participation selection and recruitment, data collection methods, and qualitative, quantitative, and mixed-method analysis methods.

The primary output of this work is a standardized testing protocol ready for immediate implementation. The user-centred service validation methodology presented in this deliverable was iteratively developed in close collaboration with the service leaders and media partners. It will be implemented by the service leaders, with media partners supporting participant recruitment under T8.3.

2. Introduction

The AI-CODE project addresses critical challenges facing media professionals in combating disinformation and supporting trustworthy content production in emerging digital environments. The project follows an iterative design and development process structured in three phases. The first project phase (M1-M12) established user requirements and initial prototypes through extensive co-creation with end users. The current second phase (M13-M24) is focusing on designing and prototyping the AI-CODE services to counter disinformation spread and adapt emerging technologies to the media-user relationship. The third project phase (M24-M36) will then concentrate on experimentation and field validation of developed tools and methodologies within specific use cases.

This deliverable presents the user-centred validation methodology designed to systematically evaluate the AI-CODE services from an end-user perspective thus contributing to Objective 4 of the project: to co-design, demonstrate and evaluate innovative approaches and technologies with end users. By bringing media organizations together with technology providers and research institutions, the validation process will further strengthen alignment of the project results with real user needs, from service design through deployment and evaluation.

The validation methodology builds upon foundational work established in earlier project deliverables. D3.1 User-centred Requirements Definition provides the user requirements, target user groups, user stories, and functional requirements that inform the validation criteria and testing protocols. D3.2 Functional Specification establishes the technical specifications that define service capabilities. Together with the prototype deliverables (D4.1, D4.2, D4.3 for content-driven services and D6.1, D6.2, D6.3 for user-driven services), D3.1 and D3.2 inform what will be tested for each service.

In the user-centred validation methodology, quality and functionality of the AI-CODE services are tested from the perspective of the target groups. A protocol for a user test has been designed including both qualitative and quantitative components. It includes assessment of usability, user acceptance and the extent to which both the ethical requirements for Trustworthy AI and the user needs are met. A qualitative protocol was chosen to allow for a deeper investigation of user perception. The methodology described in this deliverable focuses on user testing. All technological tests are covered in the corresponding service deliverables (D 5.1, D5.2, D5.3, D7.1, D7.2, and D7.3.)

The user tests are performed in two phases. The first phase is a formative phase aimed at gathering feedback for the improvement of the services; the second phase is a summative phase aimed at evaluating the services. A detailed overview is given in Sections [5](#) and [6](#).

The deliverable is structured as follows:

[Section 3](#) describes the foundational principles of the validation methodology including user-centered design, the ethical framework for Trustworthy AI, and GDPR-compliant data handling protocols.

[Section 4](#) explains the validation methodology, introducing research methods, evaluation criteria, and the validation lifecycle.

Sections [5](#) and [6](#) detail the methods used for the formative and summative phases of the user test, respectively. The procedure for gathering participants is described, followed by the methods for data collection and analysis.

3. Foundational Principles of the Validation Methodology

This section describes the foundations underlying the validation methodology. The user-centred design approach initiated in deliverable D3.1 is revisited and the relations between the design outcomes and the validation methodology are justified. Furthermore, the Ethical Framework for Trustworthy AI and the ethical protocols that are employed in the user test are discussed.

3.1 User-Centred Design Approach

An iterative user-centred design approach (Abras et al., 2004), where user needs and context of use guide key design decisions, has been well established throughout the project. By centring on the users throughout the project, AI-CODE services aim to address professional needs in combatting disinformation and supporting trustworthy Generative AI use.

The process of the requirements analysis has been as follows:

1. A literature review including a survey of 105 media organizations across 46 countries (Beckett et al., 2023) and a study of 135 interviews from 35 news organizations (Simon, 2024) served to understand wider context including media specific considerations. Ethical considerations based on journalistic ethics guidelines, as well as AI ethics principles (see D3.1) were reviewed and integrated into the design process of the services.
2. Exploratory interviews with 18 media professionals across the EU were concurrently conducted and triangulated with insights from the broader literature review and large-scale media studies (see #1), followed by a qualitative analysis of attitudes, needs and pain points towards generative AI use in media and journalism. These interviews established a baseline understanding of users and their professional contexts, essential for designing suitable services and interpreting user feedback during validation.
3. A series of service specific co-creation workshops, at the time of writing totalling 144 participants, 81 of whom were external to the AI-CODE project, were facilitated to refine the design of the six services and test existing functionalities. These workshops validated the interview and literature review findings, showing user needs had been met and how they could be improved further.

From this process, 21 target user groups (provided as [Appendix A](#)) were identified as well as a series of user stories and scenarios of use to contextualize the services. Functional requirements for each service were also set, describing all the functionalities services could provide to users. Lastly, a set of prototypes were created with integrated feedback from co-creation and validation methods carried out in workshops. The integration of an explorative process with systematic validation allows for an iterative process where feedback could be easily harnessed and translated into concrete service improvements.

These outcomes provide the essential groundwork for the validation lifecycle, detailed in [Section 4](#), and are as follows:

- *Target user groups*: 21 different professional roles were identified to be suitable beneficiaries of the AI-CODE services. These users represent a diverse range of media contexts and countries, with their understanding of generative AI and how they currently use AI in their professional practice informing both the design of the services and their validation.
- *User stories and requirements*: Drawing on interviews and the literature review, a set of user stories and scenarios of use were developed to establish realistic contexts for each service. In parallel, functional requirements were defined to specify the key features to be implemented and evaluated within the validation lifecycle. Together, these user stories and requirements both inform service design and provide measurable criteria for testing protocols.
- *Non-functional requirements* (e.g. related to usability and performance) provide the basis for evaluation criteria used to assess the services.

The user-centred validation methodology laid out in the following sections is the next step in the implementation of the user-centered design process.

3.2 Ethical Framework for Trustworthy AI

As mentioned above, AI ethics principles were also considered in the development of the services. As described in more detail in D9.5 Ethical Design Requirements, AI-CODE draws on the Ethics Guidelines for Trustworthy Design written by the European High-Level Expert Group on AI (AI-HLEG). The AI-HLEG defines an AI system as ‘trustworthy’ when it is lawful, ethical, and robust (European Commission, 2019, p. 2). The requirement for ‘ethical’ AI systems is further elaborated in seven key principles that need to be met:

1. human agency and oversight: AI systems should support the user’s agency, foster fundamental rights, and allow for human oversight (European Commission, 2019, p. 15)
2. technical robustness and safety: AI systems should be developed “with a preventative approach to risks” to avert unintentional and unacceptable harm by ensuring that AI systems “reliably behave as intended” (ibid. p. 16)
3. privacy and data governance: AI systems should be developed and deployed in a manner that follows adequate data governance protocols that address the quality and integrity of the data used, ensures the data’s relevance in light of the deployment domain, and protects privacy in access and processing protocols (ibid. p. 17)
4. transparency: The AI system developer should facilitate transparency on data, the system’s functionality and performance, and the business model behind the AI system (ibid. p. 18).
5. diversity, non-discrimination and fairness: AI systems should “enable inclusion and diversity throughout the entire AI system’s life cycle” ranging from stakeholder considerations and accessible “involvement through inclusive design processes” as well as equal treatment by the AI systems’ outputs (ibid. p. 18).
6. environmental and societal well-being: AI systems should be “used to benefit all human beings, including future generations”, thus “broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system’s life cycle.” (ibid. p. 19).

7. accountability: AI systems should have mechanisms in place “to ensure responsibility and accountability for AI systems and their outcomes” covering development, deployment and use phases (ibid. p. 19).

These principles are the foundation of the ethics-by-design (EbD) approach in the project, as explained in detail in D9.5.

For this purpose, the 7 Principles for Trustworthy AI have been translated into ethical design requirements in WP3 (see D3.1). Furthermore, in T9.3 the progress towards their implementation has been evaluated with the help of the Assessment List for Trustworthy AI (ALTAI). Tentative results from the ALTAI evaluation are included in D9.5. The results are tentative, on one hand, because the project is not concluded and service development is still in progress. On the other hand, they are tentative because some aspects of trustworthy design can be assessed only from the user's perspective.

The user-centered service validation is complementary to the T9.3 ethical evaluation, building on the same principles of trustworthiness. While in some definitions of ‘trustworthiness’ it is not definitive that users actually trust AI systems (Ryan, 2020), user perception matters for the uptake of technology and for the overall ethical assessment. Pragmatically, if AI systems are not perceived to be trustworthy by users, they will not be used (Coeckelbergh, 2020). For the success of the AI-CODE project, that implies that if our target user groups do not trust the AI-CODE services, they may not use them. For this purpose, the user validation methodology includes a number of questions related to the users’ perception of ethical principles, such as:

- Do you trust the result of the system?
 - If not, what type of information would allow you to trust that this result is correct?
- Do you have any security concerns while using the system?
- Do you have any privacy concerns while using the system?

However, user perception does not matter only for the pragmatic reason of user uptake. Instead, the implementation of some ethical principles, such as human agency and oversight and transparency depend on the position of the user.

For example, whether the principle for transparency is well implemented depends on whether the user is actually aware that they are interacting with an AI system, and not whether the developer believes this should be apparent to the user. Furthermore, the human agency and oversight principle requires that the user understands the decisions of the AI system in order to spot mistakes and potentially override the AI recommendation, e.g. when detecting logical fallacies or assessing hate speech. Thus, the requirement for human agency and oversight is dependent on measures that prevent overreliance. It is important to confirm how users understand and rely on the AI recommendations to determine whether these measures are successful. Transparency of the services can also help prevent overreliance. If it is apparent to the user that the AI system does not have an accuracy rate of 100%, they are likely to stay more vigilant in their oversight. Again, whether sufficient transparency is created for the user to exercise oversight, should be answered from the specific epistemic position of the AI-CODE users.

The following questions relating to the trustworthy AI requirements ‘human agency and oversight’ and ‘transparency’ are included in the user-centred service validation methodology as part of the

user testing script:

- Has it been apparent to you that you’ve been interacting with an AI system?
 - If so, which tasks were completed by the AI system?
- Do you understand the results of the system?
 - If not, what type of information would help you to understand system outputs?

These questions aim to reveal that ‘Trustworthiness’ is not only in the design of the services but is also dependent on the users’ epistemic faculties.

3.3 Ethical Protocols: Anonymisation, Privacy, and GDPR Compliance

Ensuring ethical integrity and compliance with data protection regulations is a central pillar of the AI-CODE project’s validation methodology. The user testing and co-creation sessions conducted under WP8 involve the collection and processing of personal data, including audio-visual recordings, transcriptions, surveys, and qualitative feedback. Therefore, all procedures are designed to fully comply with the EU General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679), with particular attention to participants’ privacy, consent, and data minimisation principles.

Informed Consent and Participation Agreements

Before any data collection takes place, all participants are required to sign an individual Participation Agreement accompanied by a Data Protection Notice. These documents inform participants about:

- The voluntary nature of their participation;
- The purpose and scope of data collection (testing and co-creation within the AI-CODE project);
- The categories of data collected (e.g., name, organisation, voice, image, transcripts and qualitative feedback);
- Data recipients, including the AI-CODE consortium partners and the European Commission for reporting purposes;
- Retention periods (until the end of the Project and for an additional two years thereafter);
- Participants’ rights under the GDPR, including rights of access, rectification, erasure, restriction of processing, portability and the right to lodge a complaint with a supervisory authority.

Each session is recorded (video and audio) for transcription and analysis. Participants are clearly informed about this prior to the session, and their explicit consent is obtained through the signed Participation Agreement.

Data Minimisation and Pseudonymisation Procedures

To ensure privacy, the project adopts a strict data minimisation approach — only data necessary for the defined research purposes is collected. All qualitative data (transcripts, notes, and coded responses) are pseudonymised before being analysed or included in project deliverables. Personal identifiers (names) will be replaced with unique participant codes, stored separately in a secure key file. Any excerpts used in reporting are paraphrased or quoted using these pseudonyms to prevent re-identification.

Raw audio and video recordings are stored securely and accessible only to authorised researchers

involved in data analysis. Once transcription and pseudonymisation are complete, the original recordings are archived under restricted access for the legally required retention period and subsequently deleted.

Data Storage, Access Control, and Security

Personal data is stored on secure, access-controlled servers of the consortium partners, in compliance with GDPR Article 32 (Security of Processing). The collected data will also be uploaded to the shared dedicated Microsoft Teams folders where all AI-CODE project materials are stored.

Technical and organisational safeguards include:

- Encrypted storage and password-protected access;
- Restricted access rights to authorised project staff only;
- Regular internal checks to ensure ongoing compliance and data integrity.

GDPR Roles and Responsibilities

Each partner acts as an independent Data Controller for the processing activities conducted within its domain, in line with the Data Protection Notice. Consortium partners cooperate under clearly defined responsibilities for safeguarding participant data and ensuring compliance across jurisdictions.

Transparency, Accountability, and Ethical Oversight

Transparency is maintained by providing participants with detailed information regarding data use, retention, and sharing. Ethical and privacy concerns identified during user testing are initially reviewed within the consortium and then, if necessary, referred to the project's independent ethics advisor.

4. The Validation Methodology

The user-centred validation methodology was iteratively developed in close collaboration with the service leaders, ensuring alignment with both technical capabilities and user needs. Literature on usability testing and Agile user testing was consulted to find the most appropriate methods for the test, which were integrated into a comprehensive user-centred test protocol. The shared protocol was then tailored to the specific needs of the services by the service teams. The user test is performed in two phases: a formative phase aimed at gathering as much feedback as possible for prototype improvement, and a summative phase aimed at evaluating the final level of usability.

This section starts with an overview of the iterative validation lifecycle. Second, the research methods and instruments are introduced: the Think Aloud Protocol and the System Usability Scale. Then, the evaluation criteria are discussed. Finally, an overview of the user test protocol is given, followed by the validation timeline.

4.1 Overview of the Validation Lifecycle

The validation lifecycle for the AI-CODE services consists of a formative and a summative user test. Formative and summative evaluation both originated in education to measure learning progress and achievement, respectively (Glazer, 2014):

- **Phase 1: Formative Evaluation (Iterative Improvement):** In the software development

context, formative assessment assesses which aspects of a software design work and why, whereas summative assessment evaluates how well the overall design performs. Formative assessments use qualitative and quantitative testing protocols to gather detailed feedback on prototype usability and interpretation of interface aspects. The findings help to uncover issues in the prototype, which are used to improve future versions of the prototype. This leads to another round of testing, thus allowing for an iterative testing approach.

- **Phase 2: Summative Evaluation (Performance Assessment):** Summative evaluation assesses how well the final design performs as compared to earlier versions of the prototypes. Both qualitative and quantitative approaches like questionnaires can be used for this phase, to triangulate different sources of data. The goal is to evaluate the overall quality and effectiveness of the completed product. Any new issues discovered may be registered for future updates.

For our user test, we choose to include both a formative and a summative phase. The formative strategy is applied to the first phase of user testing. A Think Aloud Protocol (TAP) is conducted (described in [Section 4.2](#)), during which a user is asked to perform predefined tasks with the prototypes while voicing their thoughts. The facilitator may prompt the user to explain more. A co-creation discussion is included at the end of the test session, allowing the user to share their ideas for possible improvements based on their experiences. This provides detailed insights for the service teams to further develop the prototypes, and how to better meet user needs. On top of the formative assessment methods, the System Usability Scale (SUS) questionnaire is administered (described in [Section 4.3](#)).

The summative strategy is applied to the second phase of user testing. Again, a TAP is conducted with a larger sample of users, with the SUS being administered immediately afterwards. This time, however, no co-creation discussion is held. Instead, an additional questionnaire is administered aimed at assessing perceived effectiveness and efficiency, and a short reflective conversation on user acceptance and ethics. This protocol gives the service teams an overview of the extent to which usability goals and user needs have been met.

4.2 The Think Aloud Protocol (TAP)

To obtain detailed insights into participants' thoughts and feelings while working with a new tool, a Think Aloud Protocol (TAP) is employed in this user test. This method requires users to voice their thoughts while performing a task. TAP uncovers why users take specific actions, what they expect to find, where they experience confusion or uncertainty, and how they interpret interface elements. This allows designers and developers to identify why users behave the way they do and adapt the tools accordingly. Since TAPs can also be flexibly applied in different stages of product development, and be conducted either in person or remotely, they are particularly suited to providing rich, detailed insights to complement quantitative metrics like the System Usability Scale.

TAP has its origins in cognitive psychology. TAPs were used to elicit participants' thought processes in real time, as opposed to retrospective reporting. In their seminal work, Ericsson and Simon (1980) theorized that task performance would not suffer if participants just vocalized their immediate thoughts without further processing. A meta-analysis conducted 31 years later showed that this was indeed the case (Fox et al., 2011). This meta-analysis validated that think-aloud procedures did not affect accuracy of task performance as opposed to a silent control condition, across various task

types and study designs. In contrast, when verbalization procedures required more cognitive processing, such as explanatory procedures or responding to directed task instructions, performance differed significantly from silent controls. More demanding verbalizations did increase task performance, which is consistent with the notion that requiring participants to come up with explanations or descriptions induces additional cognitive processing that can enhance problem-solving. Fox et al. (2011) note that time to task completion increases with the use of TAP regardless of task type. This shows that verbalizing thoughts takes additional time, but not at the cost of disrupting underlying cognitive processes.

4.3 The System Usability Scale (SUS)

To evaluate the usability of the services, the System Usability Scale (SUS) is employed as a key performance indicator. It is administered during the sessions in two successive evaluation phases corresponding to different levels of system maturity. The SUS was introduced by Brooke (1996) as a quick and efficient way to assess usability and has since become a widely accepted method for assessing user experience and system usability (Lewis & Sauro, 2018). This standardized scale focuses on key parameters including ease of use, consistency, satisfaction, system integration, and barriers to entry of use. It is comprised of 10 questions, which can be seen in Table 1. The SUS provides a single score ranging from 0 to 100.

The SUS offers a quantitative benchmark that complements qualitative usability insights. The two-phase approach allows the project to measure the progressive refinement of user experience from early prototypes to near-operational versions, ensuring continuous improvement of design and interaction quality.

Table 1: The System Usability Scale Standard Version (Lewis & Sauro, 2018) with notes on AI-CODE Modification

Number	Question	AI-CODE Modification	1	2	3	4	5
1	I think that I would like to use this system frequently.	Removed for T7.1 ICSM module.					
2	I found the system unnecessarily complex.						
3	I thought the system was easy to use.						
4	I think that I would need the support of a technical person to be able to use this system.						
5	I found the various functions in this system were well integrated.						
6	I thought there was too much inconsistency in this system.						
7	I would imagine that most people would learn to use this system very quickly.	Removed for all services.					
8	I found the system very awkward to use.						
9	I felt very confident using the system.						
10	I needed to learn a lot of things before I could get going with this system.						

The SUS utilizes a standard five-point Likert scale, where respondents rate their level of agreement on a scale from 1 to 5 with a series of statements ranging from "Strongly Disagree" to "Strongly Agree" (Brooke, 1996). This scaling approach allows for consistent measurement and easy comparison across different users.

The SUS was tailored to the needs of the AI-CODE services. Question 7 "I would imagine that most people would learn to use this system very quickly" was deemed methodologically unsound to include in the survey. Asking participants to assess how "most people" might feel about the service is immaterial, as participants will be projecting their experiences and biases onto others. Moreover, other researchers have also pointed out the unsuitability of question 7 and have removed it following similar objections (Gronier & Baudet, 2021, Lewis & Sauro, 2017). Research has also shown that removing one question from the scale does not compromise the validity (Lewis & Sauro, 2017) and that the SUS questionnaire can also generally be reduced in the number of items, (in extreme cases, leaving as few as four items) without losing its validity (Lewis et al., 2015). Therefore, question 7 has been removed for all services.

For the Interactive Coaching and Simulation Module (ICSM), question 1 "I think that I would like to use this product frequently" was also problematic, since the ICSM is designed for training purposes and therefore not meant to be used frequently. Asking participants about frequency of use for a training module designed for periodic skill development rather than daily operations would produce misleading data that does not reflect the tool's actual purpose or effectiveness. Therefore, question 1 has also been omitted from the SUS in the ICSM protocol.

The removal of question 7 across all services, along with the additional removal of question 1 for the ICSM requires adapting the standard 10-item SUS scoring algorithm. The modifications necessitate a recalculation of the maximum possible Raw Score and the corresponding final Scaling Factor required to convert the results to the standard 0–100 SUS scale. As supported by literature (Lewis & Sauro, 2017), the removal of a limited number of items does not compromise the validity of the scale.

The standard SUS consists of 10 items, where after normalization (converting the 1-5 Likert scale responses) each item contributes a maximum of 4 points to the raw score, establishing a canonical Raw Score range of 0 to 40. The removal of items directly reduces this maximum raw score proportionally.

For services where one question (question 7) was omitted, the maximum achievable Raw Score is reduced to 36 (9 items × 4 points/item). For the ICSM protocol, where two questions (questions 7 and 1) were omitted, the maximum Raw Score is reduced to 32 (8 items × 4 points/item).

The core principle of item normalization remains consistent with the standard SUS methodology. For all remaining items, the contribution to the Raw Score is calculated using the following formulas, where R_n is the Likert response (1 to 5) for item n :

- For Odd-Numbered Items (Positive tone): $I_n = R_n - 1$
- For Even-Numbered Items (Negative tone): $I_n = 5 - R_n$

The Raw Score (Raw_S) is the sum of the normalized contributions of all remaining items. The Raw_S is then multiplied by the appropriate Scaling Factor (SF) to yield the final SUS Score (range 0–100):

- $SUS\ Score = Raw_S \times SF$

Table 2 summarizes the parameters utilized for each protocol variation.

Table 2: Parameters utilized for each protocol variation

Number of Items	Services Covered	Max Raw Score	Scaling Factor (SF)	Final Calculation Formula
Standard (10)	Reference/Baseline	40	$\frac{100}{40} = 2.5$	$Raw_S \times 2.5$
9-Item SUS	All services (excluding ICSM)	36	$\frac{100}{36} \approx 2.778$	$Raw_S \times 2.778$
8-Item SUS	Interactive Coaching and Simulation Module (ICSM)	32	$\frac{100}{32} = 3.125$	$Raw_S \times 3.125$

The application of these revised scaling factors ensures that the resulting SUS Score accurately reflects the perceived usability across the different protocol variations and remains comparable to results obtained from the standard 10-item scale.

4.4 Evaluation Criteria

The validation framework included in this deliverable provides a comprehensive methodology to assess the user acceptance and usability of the AI-CODE services. It is structured to ensure that the outcomes are measurable, comparable, and replicable, thereby supporting a transparent validation of the project's progress toward its objectives and its contribution to trustworthy media innovation in Europe. The definition of clear evaluation criteria provides a robust and transparent basis for assessing progress and validating outcomes of the project.

4.4.1 Qualitative Evaluation Criteria

For analysis of the qualitative data resulting from the Think Aloud Protocol (TAP), a set of six analytical categories have been derived from the structure of the test protocol and informed by established human-centred and trustworthy AI evaluation principles. The following categories have been formulated:

- **Usability & User Experience (UX):** focuses on navigation, clarity of guidance, interface layout, responsiveness, and perceived ease of use. It examines how intuitively users can interact with the service, whether the interface supports efficient task completion, and how users perceive the overall experience. This category is primarily assessed through TAP observations during task performance and complemented by SUS items related to ease of use and efficiency.
- **Cognitive Understanding & System Interpretability:** examines how well users understand the system's logic, functions, and outputs. It addresses whether users can interpret what the system does, comprehend the results it produces, and predict system behaviour. This

dimension is critical for enabling meaningful human oversight and preventing over-reliance on AI recommendations. Assessment occurs primarily through TAP verbalizations as users work through tasks, supplemented by ethical discussion questions about understanding system outputs.

- **Trust, Ethics & Transparency:** explores perceptions of reliability, data use, privacy, ethical comfort, and clarity of system processes. It assesses whether users are aware they are interacting with an AI system, whether they trust the results, and whether they have concerns about security or privacy. This dimension directly addresses the Trustworthy AI principles established in the project's ethical framework. Assessment combines TAP observations, SUS items related to confidence and control and dedicated ethical discussion questions derived from the ALTAI self-assessment tool.
- **Functionality & Performance:** analyses interaction with core features and system responsiveness. It examines whether the service's functions work as expected, whether features are well integrated, and whether the system performs consistently. Assessment occurs through TAP observations of task execution and is supported by SUS items addressing complexity and consistency.
- **User Value & Professional Fit:** assesses how well the service aligns with user needs and professional expectations as defined in D3.1. It evaluates practical relevance, perceived usefulness in professional workflows, and potential barriers to adoption. This dimension ensures that services address real professional needs identified during requirements analysis. Assessment draws primarily from TAP verbalizations and co-creation discussions, complemented by SUS items related to satisfaction and willingness to use.
- **Future Improvements & Ideas:** captures user suggestions, desired features, and co-creation input. It documents opportunities for enhancement that emerge from user experience and professional insight. Assessment occurs primarily through co-creation discussions in Phase 1 and acceptance discussions in Phase 2, enriched by spontaneous suggestions during TAP sessions.

These categories are explained in more detail in Sections [5.3.1](#) and [6.3](#).

4.4.2 SUS Scores

For interpreting SUS scores, we adopt the framework developed by Bangor et al. (2009), which is based on empirical adjective rating data from over 1000 SUS assessments.

This framework is particularly appropriate for developmental evaluation contexts such as Technology Readiness Levels (TRL) 6-7 prototype testing:

85–90+ → Excellent usability

71–84 → Good usability

51–70 → OK/Fair usability

26–50 → Poor usability

For AI-CODE services at TRL 6-7, a realistic benchmark is a SUS score above 51 (achieving at least OK/Fair usability), with aspirational targets in the 71-84 range, corresponding to good usability. The reason for setting expectations in this range is that at TRL 6-7, the focus is on demonstrating technical functionality, verifying system performance, and ensuring integration within relevant or operational environments. Development efforts therefore concentrate on achieving reliable operation of the core components and validating overall technical feasibility.

The System Usability Scale (SUS) is applied across two consecutive evaluation phases to measure how effectively and intuitively end users can interact with the developed services. Each phase is aligned with the system’s maturity and development progress:

- **First SUS survey (Initial user testing and co-creation phase)**

The first SUS survey is carried out during the October/November 2025 user testing and co-creation sessions. This initial evaluation aims to collect early user feedback on the usability, interface intuitiveness, and functional adequacy of the system prototypes. The data gathered helps identify usability gaps, interface inconsistencies, and areas for improvement. Insights from this first survey directly inform the design and fine-tuning of the services in preparation for the second user testing phase planned for mid-2026, ensuring that user needs and expectations are progressively integrated into the development process.

- **Second SUS survey (The second or final project validation phase)**

The second SUS survey is conducted in the final year of the project, when the services are expected to reach TRL 6-7 and be demonstrated in near-operational environments as intermediate steps toward market readiness. This evaluation assesses the improvements achieved following the first feedback cycle and involves a broader, more representative user base. The expected outcome of this second survey is to achieve a SUS score between 71 and 84, corresponding to “Acceptable” to “Good/Above Average” usability levels.

An overview of the relation between testing phase and SUS scores can be found in Table 3.

Table 3: SUS Testing phase and corresponding SUS scores

SUS Phase	System Maturity	Expected SUS Score	Action
Survey 1	System prototype demonstrated in relevant environment	51-70	Conduct the first SUS survey during Oct–Nov user testing and co-creation sessions. Analyse results to identify usability gaps and define design improvements for the summer 2026 testing round.
Survey 2	System prototype near operational in relevant environment	71-84	Address and resolve specific usability issues identified during testing.

4.5 User Testing Protocol

The protocol design is based on usability literature covered in [Section 3](#), Agile user testing methodology and best practices from the media partners. The user tests are moderated 1-on-1 sessions, conducted remotely. The protocol has been carefully scoped in terms of task scope, time constraints, technical resources, the composition of the test team, and access to participants.

To standardize the user test process, the protocol script has been created, including a generic

process description with pre-determined texts for the facilitator to read to the participants. Additionally, each service leader has added service-specific task instructions to their version of the script. Task design is based on the user stories developed in D3.1. As introduced above, after the user task, the protocol continues with the administration of the SUS and concludes with instructions for a co-creation discussion addressing ideas for improvement and additional functionality in the formative phase, as well as ethical aspects. In the summative phase, a questionnaire focused on perceived effectiveness and efficiency is administered, followed by a short reflective discussion on user acceptance and ethics. In both the formative and summative phase, the ethical discussion is based on selected ALTAI questions.

The formative user test has the following setup:

[5 mins]	Welcome participant and sign informed consent
[5 mins]	Introduce AI CODE and explain user test procedure
[20 mins]	Perform the task while participant thinks aloud
[10 mins]	Administer the System Usability Scale
[15 mins]	Co-creation discussion including ALTAI questions
[5 mins]	Thank you and follow-up

The full script is provided in [Appendix B](#).

The Summative user test has the following setup:

[5 mins]	Welcome participant and sign informed consent
[5 mins]	Introduce AI CODE and explain user test procedure
[20 mins]	Perform the task while participant thinks aloud
[10 mins]	Administer the System Usability Scale
[15 mins]	Effectiveness and Efficiency Questionnaire and ALTAI discussion
[5 mins]	Thank you and follow-up

4.6 Validation Timeline and Milestones

As described above, validation activities are divided into two main phases — formative evaluation (Phase 1) and summative evaluation (Phase 2). Each phase includes clearly defined actions, deliverables and review milestones to ensure systematic progress tracking and consistent quality assurance across all partners.

- Phase 1 – formative evaluation (user testing and co-creation workshops): This phase takes place between October and November 2025. Five participants per service are expected to take part in the workshops, resulting in a total of at least 35 participants.
- Phase 2 – summative evaluation (user testing sessions, no co-creation discussions): This phase is scheduled for mid-2026, when all service prototypes are being developed. It follows a similar methodology to Phase 1 but involves a larger participant base. Ten participants per service are expected to take part, totalling at least 70 participants. During this phase there are no co-creation discussions; however, a SUS survey is conducted along with a service-specific questionnaire and a short reflective conversation on user acceptance and ethics.

Table 4: Validation milestones

Milestone No.	Milestone Title	Timing	Description / Objective	Responsible Partner(s)
M1	Validation Framework Finalised	September 2025	Final approval of the validation methodology, testing protocols, and participant recruitment plan	RU, DEB
M2	Phase 1 Initiation – Formative Evaluation Start	October 2025	Launch of user testing and co-creation sessions (5 participants per service)	All service leads
M3	Phase 1 Completion – Preliminary Findings Report	January 2026	Completion of all formative sessions and preliminary analysis report shared for internal review	DEB
M4	Methodology Refinement and Alignment	February-April 2026	Refinement of user testing protocol and tasks based on Phase 1 insights; preparation for summative evaluation	RU, DEB
M5	Phase 2 Initiation – Summative Evaluation (User Testing)	May-June 2026	Start of the second user testing cycle (10 participants per service) using updated protocol	All service leads
M6	Phase 2 Completion – Summative Report	September 2026	Completion of summative testing; analysis and synthesis of cross-service usability and ethics metrics.	DEB
M7	Final Validation Report Delivered	October 2026	Delivery of the consolidated validation report summarizing results across formative and summative evaluation phases.	DEB

Table 4 presents an overview of the AI-CODE validation timeline and milestones. Figure 1 visualises the main phases, their duration, and the expected completion points for each key activity described above.

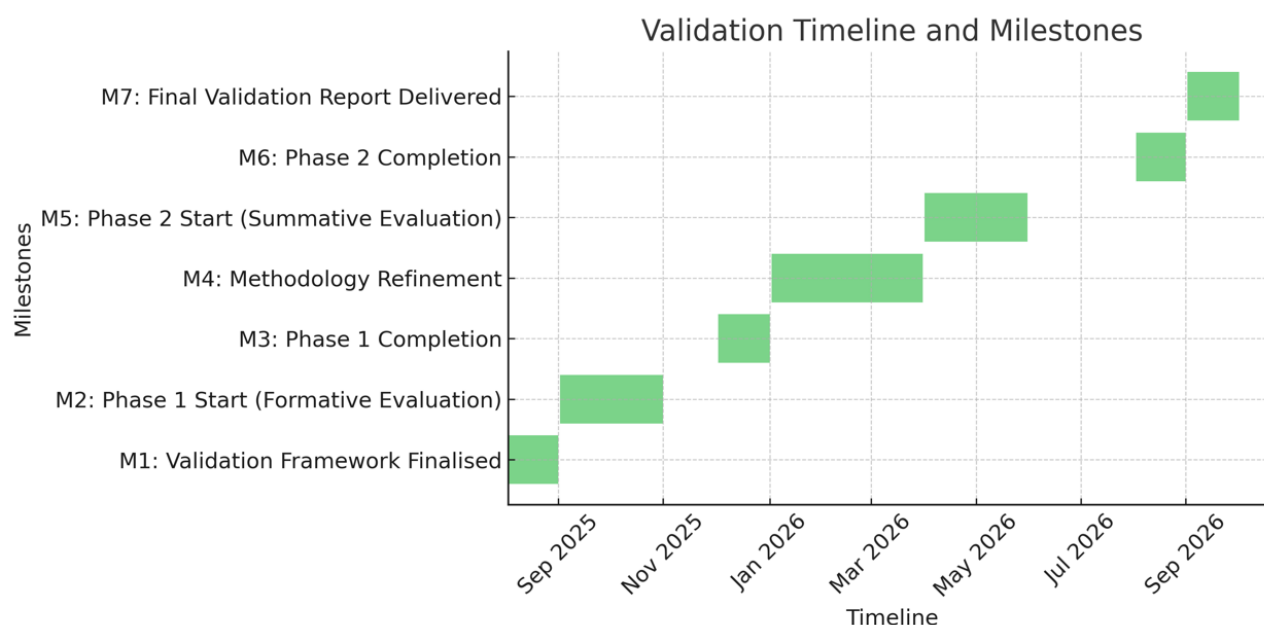


Figure 1: Validation timeline and milestones

5. Methodology for Phase 1: Formative Evaluation (User Testing + Co-creation Sessions)

The formative evaluation phase operationalizes the validation methodology presented in [Section 4](#) by conducting structured user testing with early-stage prototypes of the AI-CODE services. This phase enables systematic assessment of service usability, functionality, and alignment with Trustworthy AI principles before advancing to final validation. Scheduled for October–November 2025, this phase involves at least five participants per service (minimum 35 total participants) and aims to identify usability barriers, ethical concerns, and opportunities for improvement that will directly inform iterative design refinements.

Each user test session follows the same structure using a mixed-method approach. First, as introduced in [Section 4.2](#) a qualitative Think Aloud Protocol (TAP) is conducted. Next, a quantitative assessment using the System Usability Scale (SUS) questionnaire ([see Section 4.3](#)) is performed. The session concludes with a qualitative co-creation discussion, during which participants are asked for ideas on how to improve the services followed by a short discussion of ethical aspects based on selected ALTAI questions. This triangulated approach ensures that both immediate user reactions and considered reflections inform service development. Evaluation criteria are organized around six analytical themes: Usability & User Experience, Cognitive Understanding & System Interpretability, Trust, Ethics & Transparency, Functionality & Performance, User Value & Professional Fit, and Future Improvements & Ideas.

5.1 Participants

The formative user testing phase involves a minimum of five participants per service, resulting in a total of at least **35** participants across the consortium. A convenience sampling strategy is applied, focusing primarily on existing AI-CODE media partners including DW, Debunk.org, Euractiv Romania, and Euractiv Poland. These participants include both users who have previously engaged in co-

creation activities and new users who bring a fresh perspective to the evaluation.

The target group consists mainly of different types of media professionals, like journalists, editors, and fact-checkers who work with information production, verification, and editorial workflows. A list of AI-CODE target users is identified in [Appendix A](#). The aim of the approach to participant recruitment is to capture diverse user perspectives on the usability, user acceptance, and trustworthiness of the services, while acknowledging that strict demographic representativeness is not feasible in the context of voluntary participation.

Recruitment focuses on ensuring a balanced distribution of user types identified in the project's earlier user research (see D3.1), aiming to include both primary target users and expanded user groups. As seen in Figure 2, each service lead prioritizes their first, second, and third target user group for user testing sessions. The definition of these ranked user groups is informed by the primary and secondary target user groups defined in deliverable D3.1. This information is used to provide contact details for appropriate media professionals provided by the media partners.

Target User Group	5.1 Disinformation Detection in the Next-Generation Social Media	5.2 Trustability and Credibility Assessment	5.3 Media Asset Annotation and Management (MAAM)	7.1 Interactive Coaching and Simulation Module (ICSM)	7.1 Counter-narrative Recommendation Module (CNRM)	7.2 Transparency Service for AI Model Cards	7.3 Personal Companion for Understanding Disinformation
(Publication) Editors						2	3
Brands & Marketing teams	3						
Content Moderators	1	1				3	
Data Scientists (Media Industry)			2			2	
Digital content creators of outreach / dissemination / divulgation of present events		2					
Documentary filmmakers		3		2			
Election Monitoring Organizations			2				
Fact-checkers	2		1		1	3	1
Government Organizations / Policy makers / Regulators			3				
Journalists / Investigative Journalists	2	1	1	1	3	2	1
Media Archivists			3	2			
Media Asset Managers			2	2			2
Media educators & media students			2	1	3		3
Media Organizations				3			2
Model Card Creators (AI/ML Engineers, AI-System Designer, Researchers, etc.)						1	
Model Consumers (Journalists, Fact-checkers, Developers, etc.)						1	
Online community managers	1	3			2		
PR / communication departments	3						
Social media managers		2			2		
Social scientists and Researchers in Media Studies/Journalism			1				
Technical Editors						2	

Figure 2: Target user group prioritization per AI-CODE service

Eligibility criteria are aligned with the target user groups defined in D3.1. While many services could be tested by typical end users without specialized technical knowledge, services requiring specific expertise (e.g., technical, data-related, or domain-specific content expertise) are matched with participants who possessed the necessary professional background.

These criteria focus on professional role; involvement in editorial or verification workflows; and ability to participate remotely in English. Where possible, efforts are made to ensure diversity in terms of gender, professional background, and geographical distribution within the constraints of the available participant pool.

Given the voluntary nature of participation and the risk of no-shows or last-minute cancellations, a larger recruitment pool is necessary to secure the minimum number of completed sessions. At the same time, the unpredictable schedules and high workload of media professionals require flexible and adaptive planning to accommodate changing availability while ensuring that each service partner can still conduct the required number of user tests.

Participants are briefed on the objectives of the test, their rights under GDPR, and the voluntary nature of participation. Written consent is collected prior to any data collection activities. Participants are also informed that all responses are treated confidentially and used exclusively for research and development purposes.

To validate the technical setup and test workflow, most service partners carried out a pre-test with DW. These internal pilot sessions ensure that the tools, instructions, and data collection procedures function as intended, allowing minor adjustments to be made before involving external participants.

5.2 Data Collection Methods

Each service conducts at least five user test sessions. The facilitator is a member of the service development team. All sessions are conducted in English using Microsoft Teams, a video conferencing tool that is capable of video and audio recording and accurate on the fly transcription in English. Both audio and video are recorded. Transcripts are pseudonymised according to the GDPR-compliant procedure described in [Section 3.4](#).

A script has been developed to standardize the course of the sessions. First, the user is welcomed and informed about the data storage policy described in [Section 3.4](#). After this, the AI-CODE project and the procedure for the session are explained.

For the Think Aloud Protocol (TAP), service-specific tasks have been designed. The facilitator encourages the user to talk freely while performing the task. In case the user falls silent, the following generic open prompts are suggested in the script:

- What are you thinking?
- What are you searching for?
- What do you think you see on the screen?
- Do you know what to do with the results of this task? How sure are you?

Service-specific prompts have been added to the scripts by the service teams. The facilitator logs any issues that come up such as points of confusion, navigational breakdowns, and UI/UX issues. These are to be discussed during the co-creation discussion.

Next, the facilitator administers the System Usability Scale (SUS). The SUS is available as a Google Form. The facilitator talks the user through the questions but does not see the user's responses so as not to influence the user to be falsely optimistic. The SUS is also pseudonymised as described in [Section 3.4](#).

After SUS completion, the co-creation discussion is held. For this discussion, the following questions have been specified in the script:

1. What worked well for you while you were using the service?
2. Is there anything that did not work for you while using the service?
3. *Optional follow up on 1 and 2: Ask remaining questions about how well the service works that came up during the TAP.*
4. How could the existing features be improved?
5. What would you like to add to the service in the future?

Additionally, several ethical questions from the self-assessment tool, ALTAI, are given that can optionally be included in the co-creation discussion, as the service teams see fit. The following questions are given, as introduced in [Section 3.2](#).

1. Has it been apparent to you that you've been interacting with an AI system?
 - a. If so, which tasks were completed by the AI system?
2. Do you understand the results of the system?
 - a. If not, what type of information would help you to understand system outputs?
3. Do you trust the result of the system?
 - a. If not, what type of information would you like to see to convince you that this result is correct?
4. Do you have any security concerns while using the system?
5. Do you have any privacy concerns while using the system?

5.3 Data Analysis Methods

The analysis of data collected in the formative user testing phase follows a mixed-methods design that ensures a consistent analytical approach across all services, visualized in Figure 3. It combines:

- **qualitative analysis** of TAP sessions and co-creation discussions,
- **quantitative analysis** of SUS responses,
- **cross-method triangulation** of qualitative and quantitative findings, enriched by reflective insights from co-creation discussions, and
- **a structured synthesis and feedback process** that translates results into actionable improvements for each service.

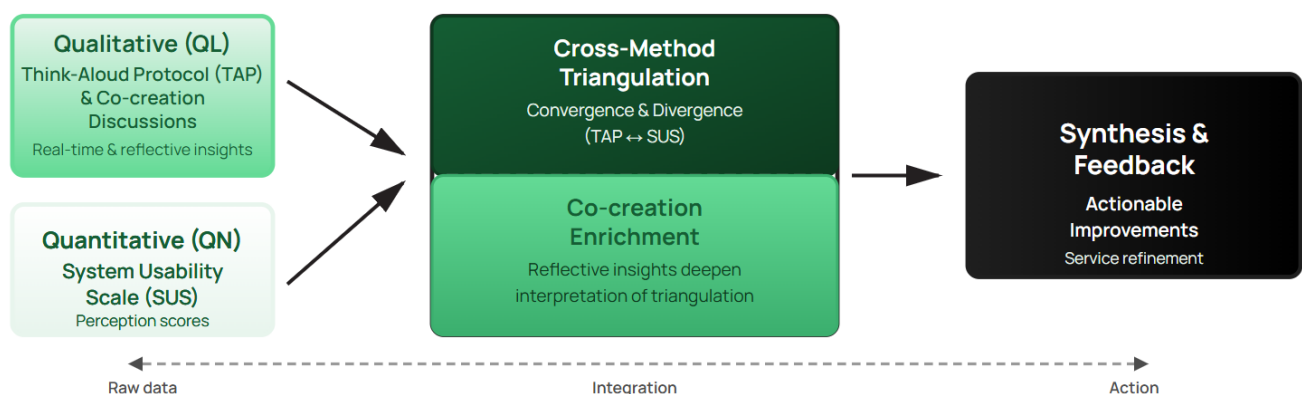


Figure 3: Mixed methods analysis

All user testing and co-creation sessions are audio- and video-recorded, transcribed verbatim, cleaned, and pseudonymised according to the GDPR-compliant protocol described in [Section 3.3](#).

5.3.1 Qualitative Analysis: Think Aloud Protocol (TAP) and Co-creation Discussion

Once the transcripts are cleaned—filler words removed where appropriate, participant identifiers standardised, and key segments (introduction, TAP session, SUS reflection, co-creation discussion, and AI trustworthiness questions) timestamped—the analysts review transcripts, revisiting relevant video segments only when clarification of tone, affect, or interaction flow is needed. This transcript-first approach ensures consistency across analysts and supports efficient, transparent interpretation of participant feedback.

Each transcript is then prepared for analysis following a structured framework aligned with the research objectives and ethical commitments of the project. This includes initial familiarisation, note-taking, and preparation for thematic coding as described below.

During the **familiarisation phase**, analysts review the cleaned transcripts and, where needed, selected video segments to clarify tone or context, noting recurring ideas, points of confusion, and expressions of engagement. These early observations do not constitute formal coding but may provide occasional paralinguistic or behavioural contextual cues (observational markers)—such as noting hesitations or shifts in tone when they meaningfully clarify user uncertainty or intent.

Following the initial familiarisation, the cleaned transcripts are prepared for analysis within **computer-assisted and/or AI-supported qualitative environments**, depending on analytical and technical requirements at the time of implementation. These may include traditional Computer-Assisted Qualitative Data Analysis (CAQDA) platforms or AI-assisted computational frameworks (e.g., large language model-based workflows or topic-modelling approaches). These tools assist with tasks such as tagging, hierarchical coding, and retrieving thematically related excerpts. All coding and interpretation remain fully human-led: analysts review and validate all outputs, make all coding decisions, and conduct the final thematic synthesis to ensure methodological rigour and alignment with the project's ethical principles.

Each transcript is prepared for coding using the project's structured analytical framework. The coding process combines a **primarily deductive approach** (coding according to predefined categories derived from the protocol) with **inductive flexibility** (allowing new codes to emerge from the data when participants raise unexpected points). Deductive coding ensures alignment with the protocol structure and project objectives, while inductive coding allows additional nuances or emerging patterns to be incorporated when they recur across participants

Inductive patterns may be detected either through manual interpretive review or with the support of AI-assisted text exploration tools (e.g., clustering, keyword extraction, or semantic grouping). All such emergent insights are examined, refined, and validated by human analysts to ensure conceptual coherence and alignment with the project's ethical and analytical framework. When recurring or substantively meaningful, these emergent codes are either integrated within existing deductive categories or formalised as new sub-themes, ensuring that the coding framework remains both structured and responsive to authentic user perspectives.

The **predefined set of six analytical categories** — or **core themes** — is derived from the structure of the test protocol (introduction, TAP, SUS reflection, co-creation discussion, and AI

trustworthiness questions) and informed by established human-centred and trustworthy AI evaluation principles. The six core themes (described in detail in [Section 4.4.1](#)) are:

- Usability & User Experience (UX)
- Cognitive Understanding & System Interpretability
- Trust, Ethics & Transparency
- Functionality & Performance
- User Value & Professional Fit
- Future Improvements & Ideas

Table 5 provides examples of codes and tags associated with the categories. This structured framework ensures analytical consistency across services and supports comparable interpretation of participant responses, while allowing flexibility for inductive coding when novel insights emerge.

*Table 5: Illustrative Codes for Each Qualitative Theme**

Theme	Description	Illustrative Codes/Tags
Usability & UX	Focuses on navigation, clarity of guidance, interface layout, responsiveness, and perceived ease of use.	e.g., navigation clarity; task guidance; interface layout; responsiveness; feedback clarity
Cognitive Understanding & System Interpretability	Examines how well users understand the system's logic, functions, and outputs.	e.g., result interpretation; explanation clarity; information sufficiency; system predictability
Trust, Ethics & Transparency	Explores perceptions of reliability, data use, privacy, ethical comfort, and clarity of system processes.	e.g., trust in results; privacy concerns; ethical comfort; transparency of process; perceived safety; AI awareness
Functionality & Performance	Analyses interaction with core features and system responsiveness.	e.g., feature integration; stability and responsiveness; consistency of feature behaviour; smooth progression through tasks
User Value & Professional Fit	Assesses how well the service aligns with user needs and professional expectations.	e.g., practical relevance; perceived usefulness; efficiency gains; barriers to adoption
Future Improvements & Ideas	Captures user suggestions, desired features, and co-creation input.	e.g., desired features; improvement suggestions; missing functionality; suggestions for alignment with professional practice; ethical recommendations; perceived limitations

**Codes/tags listed in the table are illustrative examples. Actual codes/tags will vary by service and will be refined during analysis.*

Transcripts will be coded according to these themes using descriptive tags to capture recurring

perceptions, usability barriers, interpretations of system outputs, and suggestions for improvement.

In addition to TAP data, **co-creation discussions** provide reflective insights that complement real-time observations and enrich the qualitative dataset, also strengthening subsequent triangulation.

Co-creation discussion data are analysed using the same overarching coding framework and mixed deductive–inductive logic, ensuring conceptual consistency across analytical phases. The focus and granularity of coding are expected to differ: while TAP sessions emphasise usability and task-level behaviour, co-creation discussions capture reflective reasoning, ethical considerations, and forward-looking suggestions. This shared but context-sensitive approach enables results to be compared and synthesised while respecting the distinct characteristics of each data type.

5.3.2 Quantitative Analysis (SUS)

As a validated and lightweight tool designed to measure users’ overall perception of a system’s usability, the **SUS**, [as described in Section 4.3](#), complements qualitative insights gathered through TAP and co-creation discussions by providing **quantitative evidence** of perceived ease of use, efficiency, and confidence. The SUS uses ten 5-point Likert items; scores are converted to a 0–100 index via the standard formula. In this project, services use an adapted version of the SUS: Question 7 “I would imagine that most people would learn to use this system very quickly” was removed for all services because it is methodologically unsound (see [Section 4.3](#)) and question 1 “I think that I would like to use this system frequently” was omitted for the Interactive Coaching and Simulation Module (ICSM) of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals because it is not intended to be used frequently. This resulted in a 9-item SUS for most services and an 8-item SUS for ICSM. As laid out in [Section 4.3](#) this adaptation still preserves the validity of the questionnaire.

When items are removed, the same scoring logic still applies, as each SUS item contributes equally once converted to the 0–4 scale; the total possible score is adjusted to reflect the reduced number of items, after which scores are proportionally rescaled to the 0–100 range. This proportional rescaling approach, detailed in [Section 4.3](#), is established in usability practice and ensures comparability with standard SUS interpretation thresholds.

Following the interpretation framework established in [Section 4.4.2](#) (Bangor et al., 2009), we interpret scores as follows:

- **85–90+** → Excellent usability
- **71–84** → Good usability
- **51–70** → OK/Fair usability
- **26–50** → Poor usability
- **0–25** → Worst imaginable

This framework, based on empirical adjective rating data, is appropriate for prototype systems at TRL 6-7, where resources are focused on core technical functionality rather than interface optimization.

SUS scores are analysed alongside qualitative codes to provide a holistic understanding of user experience. This mixed-methods triangulation enables interpretation of why particular usability ratings occur by linking numerical feedback with observed behaviour and verbalised reasoning; it is expanded upon in the dedicated section on triangulation.

Scores are analysed at the service level to examine distribution patterns, identify areas of strength or potential usability concerns, and support interpretation alongside qualitative findings. To facilitate mixed-methods analysis, each SUS dimension is linked to corresponding qualitative themes, enabling coherent comparison across data types.

5.3.3 Mixed-Methods Findings Analysis

Triangulation between System Usability Scale (SUS) scores and Think Aloud Protocol (TAP)-derived insights enables a structured comparison between participants' self-reported usability and the qualitative evidence that explains or contextualises these ratings. This mixed-methods integration strengthens interpretive validity by indicating where usability perceptions align with, or diverge from, participants' articulated experiences.

For analysis, SUS items are interpreted through commonly used conceptual groupings (e.g., ease of use, complexity, confidence), hereafter referred to as 'SUS dimensions,' which are then mapped to corresponding qualitative themes and example codes from the TAP transcripts. This allows analysts to examine patterns of **convergence** (e.g., high SUS ratings reinforced by clear qualitative indications of ease or clarity) and **divergence** (e.g., positive ratings despite reported confusion or uncertainty).

Both convergent and divergent cases are analytically valuable: alignment supports confidence in the findings, while discrepancies can point to unexpressed expectations, contextual influences, or usability challenges not fully captured through rating scales alone. The mapping framework below illustrates how these conceptual SUS groupings correspond to qualitative thematic categories, ensuring that numerical scores are interpreted alongside concrete user evidence.

The mapping framework in Table 6 shows how SUS dimensions link to qualitative themes, allowing each numerical score to be interpreted alongside concrete participant insights and revealing areas of convergence, partial alignment, or divergence between self-reported usability and articulated experience.

Table 6: Mapping framework for triangulating qualitative themes with SUS dimensions

SUS Dimension	Corresponding Themes	Illustrative Codes/Tags
Ease of Use / Efficiency	Usability & User Experience (UX)	e.g., navigation clarity; task guidance; interface layout; responsiveness; feedback clarity
Learnability	Usability & UX	e.g., initial orientation; clarity of guidance; intuitive starting point; self-explanatory functions; low onboarding effort
Complexity / Cognitive Load	Cognitive Understanding & System Interpretability	e.g., result interpretation; explanation clarity; information sufficiency; system predictability
Consistency &	Functionality & Performance;	e.g., feature integration; stability and

Predictability	Usability & UX	responsiveness; consistency of feature behaviour; smooth progression through tasks
Confidence & Control	Trust, Ethics & Transparency	e.g., trust in results; privacy concerns; ethical comfort; perceived safety; transparency of process
Satisfaction / Acceptance / Willingness to Adopt	User Value & Professional Fit	e.g., relevance to professional practice; perceived usefulness; efficiency gains; barriers to adoption; conditional willingness to adopt

Together with TAP and SUS data, findings from **co-creation discussions** enrich the overall analytical framework. They situate usability insights within broader dimensions of trust, ethical comfort, and professional relevance, completing the interpretive cycle of the analysis. Certain themes—such as ‘Future Improvements & Ideas’—are informed primarily through co-creation and post-use reflection, enriching the triangulation with forward-looking insights that extend beyond the scope of SUS ratings.

Table 7 summarises the conceptual structure of this **multi-layered triangulation**, showing how each analytical component contributes distinct yet complementary insights into user experience and trust in the AI-CODE services.

Table 7: Conceptual structure of triangulation across analytical layers

Layer	Purpose	Data Type	Triangulation role
TAP (Think-Aloud Protocol)	Captures real-time reasoning, confusion, discovery, and verbalised experience.	Qualitative (verbal + contextual)	Qualitative base of triangulation
SUS (System Usability Scale)	Quantifies usability perceptions.	Quantitative (scores)	Validates TAP themes numerically
Co-creation Discussions	Captures reflective reasoning, trust, and value alignment.	Qualitative (post-use reflection)	Contextual enrichment — deepens interpretation of TAP–SUS results

Key outcomes will be organised into three actionable categories:

- **Critical Usability and Ethical Issues** — high-impact findings that impede effective use or conflict with Trustworthy AI principles (e.g. lack of transparency, privacy concerns, poor interpretability). These issues require immediate action.
- **Moderate or Contextual Issues** — recurring but non-critical challenges that can be

addressed through interface optimisation or additional documentation. These issues can be monitored in subsequent iterations.

- **Enhancement Opportunities** — user-driven suggestions for new functionalities or contextual improvements. These are positive design elements that foster trust and user satisfaction.

This consolidated analysis will directly inform design refinements for the summative evaluation phase, ensuring that AI-CODE services evolve toward both functional usability and ethical robustness.

5.3.4 Findings Synthesis & Feedback Integration

When the coding and analysis phases are complete, each service leader will receive a structured feedback report prepared by the project partner Debunk.org. The reports follow a uniform format to ensure cross-service comparability and include sections on critical usability and ethical issues, moderate or contextual issues, and enhancement opportunities.

Each item is accompanied by a proposed design action, responsible partner(s), and an indicative implementation timeframe to guide the refinement process. Feedback will also be incorporated through bi-weekly coordination meetings between the service leaders and the validation team. These sessions serve to:

- Review feedback reports and agree on the prioritisation of issues;
- Define specific design modifications, usability refinements, or ethical mitigations;
- Track implementation progress between the formative and summative testing phases.

All changes resulting from user feedback are documented in a validation change log, maintained by each service partner and consolidated by Debunk.org. The log records:

- The description and impact level of the issue;
- The corrective action taken;
- The verification status (addressed, pending, or deferred).

This documentation ensures full traceability and accountability in the integration process, linking design evolution to user evidence and ethical reasoning.

By embedding feedback loops between the formative and summative phases, AI-CODE's validation methodology operationalises the project's commitment to continuous improvement. The iterative feedback-integration process ensures that the services evolve not only toward higher usability and performance but also toward stronger ethical compliance and user trust.

Ultimately, this approach guarantees that the insights gathered during Phase 1 directly inform the refinement of prototypes tested in Phase 2, creating a transparent and reproducible pathway from user input to service enhancement within the overall AI-CODE lifecycle.

6. Methodology for Phase 2: Summative Evaluation

The feedback and experiences gathered during the first phase of evaluation directly inform the

operationalisation of the second phase. The summative phase involves conducting the same usability tests as the formative phase, but the co-creation discussion is replaced by a service-specific questionnaire on perceived effectiveness and efficiency, and a short reflective conversation on user acceptance and ethics. Additionally, at least ten participants per service perform the test instead of 5. The focus of the summative phase is to assess whether the services meet the usability and user acceptance criteria.

6.1 Participants

The summative user testing phase involves a minimum of ten participants per service, resulting in a total of at least **70** participants. These participants include both users who have previously engaged in formative evaluation and co-creation activities and new users without prior exposure to the services. Special efforts are made to bring AI-CODE services to media organizations outside of the convenience sample of AI-CODE consortium partners. Efforts to introduce the AI-CODE services to additional media organizations through user testing are made by all members of the consortium, including the service developers.

The target group and recruitment process is similar to the process outlined in the formative phase in [Section 5.1](#), however recruitment is expanded beyond the convenience sampling used in the formative phase to enable a broader and more structured participant pool suitable for summative evaluation. Additional pre-tests with any partner organization may be conducted in the summative phase if needed, building on the formative pilot with DW, which had already helped validate core workflows and materials.

6.2 Data Collection Methods

Each service conducts at least ten user tests. The first part of the summative protocol is identical to the first part of the formative protocol as described in [Section 5.2](#). Participants are asked to perform the same service-specific task to conduct the user test, again using a Think Aloud Protocol (TAP) to elicit participants' experiences. The same script for the facilitator is used. After the TAP, the System Usability Scale (SUS) questionnaire is again administered.

As noted above, after administering the SUS, the summative protocol deviates from the formative protocol.

The co-creation discussion is replaced by a service-specific Effectiveness & Efficiency Questionnaire, and a User Acceptance & Ethics Discussion. Each service's questionnaire includes tailored questions to assess the user's perception of how well the service aligns with user-centred functional and non-functional requirements specified in D3.1, with a focus on perceived effectiveness and efficiency of the service. Perceived effectiveness assesses how well the service supports the user to complete the task, and perceived efficiency focuses on how efficiently the user thinks the task can be completed with the service. The session concludes with a short reflective conversation on user acceptance and ethics, using the ALTAI questions for the latter.

It should be noted that the Effectiveness & Efficiency questionnaire and the prompts for the reflective conversation will be further operationalized during milestone M4 'Methodology Refinement and Alignment', as specified in Table 3. The operationalization will be guided by the

insights from the formative user test phase.

6.3 Data Analysis Methods

While the data collection instruments differ between the formative and summative phases, the data analysis methods remain unchanged. TAP transcripts in the summative phase are analysed using the same six qualitative analytical categories as the formative phase (Usability & UX; Cognitive Understanding & System Interpretability; Trust, Ethics & Transparency; Functionality & Performance; User Value & Professional Fit; Future Improvements & Ideas). These categories ensure consistent interpretation of user experience across both evaluation phases and facilitate continuity in how findings are interpreted across phases.

These analytical categories provide the foundation for coding and interpreting TAP data across both evaluation phases. The integration of these qualitative findings with quantitative measures follows a structured triangulation framework, enabling multi-faceted understanding of user experience. Table 8 summarises the analytical layers and their role in synthesising TAP, SUS, the Effectiveness & Efficiency Questionnaire, and User Acceptance & Ethics Discussion.

Table 8: Summative Triangulation Structure

Analytical Layer	Data Source	Corresponding Themes	Purpose in the Summative Evaluation	Illustrative Indicators (Codes / Metrics)
TAP (Think-Aloud Protocol)	Real-time task execution	All six qualitative themes (Table 7)	Identify persistent usability barriers, misinterpretations, remaining friction points	Navigation clarity; task guidance; system predictability; feature behaviour
SUS (System Usability Scale)	Standardised usability questionnaire (adapted)	Usability & UX; User Value & Professional Fit	Quantify perceived improvements since formative testing; compare with established usability benchmarks	SUS score; distribution changes; improvement over Phase 1
Effectiveness & Efficiency Questionnaire	Service-specific summative survey	Functionality & Performance; User Value & Professional Fit; Trust, Ethics & Transparency;	Assess whether the service helps users complete tasks more accurately and efficiently; assesses relevant user-centred requirements from D3.1	Perceived efficiency gains; perceived task accuracy; relevance to workflow

User Acceptance & Ethics Discussion	Short reflective conversation	Trust, Ethics & Transparency	Examine user acceptance, responsible use concerns, and user confidence in final-stage prototypes	Trust in outputs; perceived safety; ethical comfort; transparency expectations
--	-------------------------------	------------------------------	--	--

Since the D8.3 Validation Report is due in Month 36 — the final month of the project — there will be no coordination meetings or separate feedback sessions with the service leaders. All findings will be integrated into this report and shared with the service leaders.

7. Conclusion

This deliverable describes the comprehensive validation methodology designed to systematically evaluate the AI-CODE services from the end-user perspective. The methodology comprises a user test conducted in two phases: the formative evaluation aims at gathering detailed feedback for prototype improvement, and the summative evaluation focuses on assessing overall performance and user acceptance of refined services. Through a mixed-methods approach combining Think Aloud Protocol (TAP) for qualitative insights, System Usability Scale (SUS) for quantitative usability assessment, and structured ethical discussions, the methodology ensures robust and transparent evaluation of service quality, functionality, and alignment with Trustworthy AI principles. A comprehensive data analysis approach is designed based on thematic analysis and previous experiences with user testing conducted by the partner organisations.

The choice for TAP results in a rich insight into the user's experience and thoughts, and the choice for triangulating this qualitative data with the validated SUS questionnaire ensures valid, reliable, and complete feedback for the services. The structured analytical framework organized around six core themes—Usability & User Experience, Cognitive Understanding & System Interpretability, Trust, Ethics & Transparency, Functionality & Performance, User Value & Professional Fit, and Future Improvements & Ideas—provides consistent evaluation criteria across all services while allowing flexibility for service-specific requirements.

The validation framework directly supports Objective 4 of the AI-CODE project by operationalizing the project's commitment to continuous improvement and ethical responsibility, ensuring that services not only demonstrate technical functionality but also earn user trust and align with professional needs and practices.

The next steps are to continue the implementation of the evaluation methodology through Task 8.3, involving the execution of the formative evaluation phase with at least five participants per service, followed by prototype refinements based on structured feedback reports. The summative evaluation phase will then assess whether services achieve targeted usability levels (SUS scores of 71-84, corresponding to good usability) and meet user acceptance criteria with at least ten participants per service. The consolidated results and recommendations from both evaluation

phases will be presented in deliverable D8.3 (M36). This iterative evaluation process will ensure that the AI-CODE project will deliver mature, user-validated services which can effectively support journalists in combating disinformation and in producing trustworthy content.

8. 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.
- Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3), 114-123.
- Beckett, C., & M. Yaseen. 2023. *Generating Change. A Global Survey of What News Organisations are Doing with AI*. The London School of Economics and Political Science.
- Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
- Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215–251. <https://doi.org/10.1037/0033-295X.87.3.215>
- European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316-44. <https://doi.org/10.1037/a0021663>
- Glazer, N. (2014). Formative Plus Summative Assessment in Large Undergraduate Courses: Why Both? *International Journal of Teaching and Learning in Higher Education*, 26(2), 276-286.
- Gronier, G., & Baudet, A. (2021). Psychometric Evaluation of the F-SUS: Creation and Validation of the French Version of the System Usability Scale. *International Journal of Human-Computer Interaction*, 37(16), 1571-1582. <https://doi.org/10.1080/10447318.2021.1898828>
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Investigating the correspondence between UMUX-LITE and SUS scores. In *Design, User Experience, and Usability: Design Discourse: 4th International Conference, DUXU 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings, Part I* (pp. 204-211). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20886-2_20
- Lewis, J. R., & Sauro, J. (2017). Can I Leave This One Out? The Effect of Dropping an Item From the SUS. *Journal of Usability Studies*, 13(1), 38-46. <https://uxpajournal.org/dropping-item-sus/>
- Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. *International Journal of*

Human–Computer Interaction, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>

Ryan, M. (2020). In *AI We Trust: Ethics, Artificial Intelligence, and Reliability*. *Science and Engineering Ethics*, 26(5), 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Morgan Kaufmann. (p. 203 and p. 204, Table 8.6)

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

9. Appendix A: User Groups

1. Brands & Marketing teams
2. Content Moderators
3. Data Scientists (Media Industry)
4. Digital content creators of outreach / dissemination / divulgation of present events
5. Documentary filmmakers
6. Election Monitoring Organizations
7. Fact-checkers
8. Government Organizations / Policy makers / Regulators
9. Journalists / Investigative Journalists
10. Media Archivists
11. Media Asset Managers
12. Media educators & media students
13. Media Organizations
14. Model Card Creators (AI/ML Engineers, AI-System Designer, Researchers, etc.)
15. Model Consumers (Journalists, Fact-checkers, Developers, etc.)
16. Online community managers
17. PR / communication departments
18. (Publication) Editors
19. Social media managers
20. Social scientists and Researchers in Media Studies/Journalism
21. Technical Editors

10. Appendix B: Formative user test protocol

This script provides a protocol for conducting a user test tailored to your own service. The script contains generic instructions for guiding the user through the session, facilitating the task (Think Aloud prompts and the questionnaire), and conducting the co-creation part. In certain places service specific instructions must be included that you need to write yourself. These places are highlighted in yellow.

Overview of the user test procedure

This is the general procedure for the user test. Please stick to these planned durations for each

part meticulously to get everything done.

[5 mins]	Welcome participant and sign informed consent
[5 mins]	Introduce AI CODE and explain user test procedure
[20 mins]	Perform the task while participant thinks aloud
[10 mins]	Administer the System Usability Scale
[15 mins]	Co-creation discussion
[5 mins]	Thank you and follow-up

Prerequisites for the session

- Informed consent form
- Service specific task script
- A tool for recording the session that can record audio and video, and transcribe
- Any task-specific materials that should be made available to the user before starting the task to ensure a smooth start. Think about a working login, collections of videos or images etc.
- A digital version of the System Usability Scale with selected questions

Welcome and informed consent [5 mins]

[00:00] Tell user: “Thank you very much for participating in this user test. Before we start this test, we ask your permission to record this session for research purposes. *{Describe data policy in a few sentences: who will see this data? What will we do with it? Where will it be stored? For how long will it be stored?}*. Do you agree to this? If so, will you sign this informed consent form?”

[Present user with form]

[Let user sign form]

[Start recording]

Introduction AI CODE and procedure explanation [5 mins]

[00:05] Tell user: “Thank you once again for participating. We have developed a collection of services to help media professionals use generative AI for trusted information production, and to counter threats such as misinformation and disinformation. These services are being developed as part of an EU-project called AI-CODE, a collaboration between academia, media partners and ethical experts.”

“In this user test, we will be testing the *{service name. Add a clear description explaining your service, provide what is minimally needed to allow the user to understand the service and required context so they can perform the task, but without being overwhelmed}*. We want to know how usable the tool is, and whether it meets your professional needs. We are testing *{features X, Y}*, and your goal is to *{specify task goal}*. Please keep in mind that the service is still a prototype, and unexpected issues may arise.”

“We will be spending the first 30 minutes of this session working with the tool. We ask you to complete a number of tasks, and we want to see how you use them. During this task, we will

ask you to think out loud. Please keep in mind that there are no right or wrong actions or answers, the goal is to test the prototype and every bit of input we get from you is valuable.”

“After the tasks, we will ask you to fill out a 10-item questionnaire. After the questionnaire we will discuss ethical aspects. We will spend the last 15 minutes of the session discussing more generically what you liked and disliked about the tool, and what we could improve and add in the future.”

Perform the task [20 mins]

Now that the prerequisites are available and the tasks have been designed, it is time to start the actual user test. First, the Think Aloud Protocol will be conducted, second, the questionnaire will be administered, and third, ethical aspects will be discussed.

[00:10] Tell user: “We will now start the first part of this test. We will be using a method called a Think Aloud Protocol. This means that while you are working, you will be asked to speak about what you are thinking. We will occasionally prompt you with further questions if needed. Again, there are no right or wrong answers.”

[Insert your task design]

[Present the user with the task and explain]

Think Aloud Protocol

Ask the following questions while the user is performing the task you designed. There are generic prompts and service specific prompts.

Generic TAP prompts

- What are you thinking?
- What are you searching for?
- What do you think you see on the screen?
- Do you know what to do with the results of this task? How sure are you?

Service-specific TAP prompts

*{To create service-specific prompts, identify specific issues you want to verify, and/or specific validation questions you want to ask, eg. UI questions. Please make sure they are not leading questions. If the issue related to the prompt shows up during the task performance, you can probe the user with these questions. If the issues do **not** show up during task performance, leave the prompts for the co-creation part after the task so as not to interrupt the user’s flow while performing the task.}*

Administer the System Usability Scale (SUS) [10 mins]

Now that the task has been performed, the SUS must be administered. Make sure you can check whether the user has completed the questionnaire.

[00:30] Tell user: “We now ask you to fill out a {x}-item questionnaire to complement the discussion we’ve just had while you were performing the tasks.”

[Provide user with SUS (digital link?) and complete it together]

Co-creation [15 mins]

Now that the task performance and SUS have been completed, a generic discussion follows about what did and what didn't work while using the service. In any case, ask questions 1-5 from the generic co-creation discussion list. You may add questions from the ALTAI self-assessment list if you deem them relevant for your task.

[00:40] Tell user: “For these last 15 minutes, we will discuss more generically what you liked about the service and what you didn't like, and how you would like to see the service improve in the future”.

Generic co-creation discussion**[Discuss the following questions with the user]**

1. What worked well for you while you were using the service?
2. Is there anything that did not work for you while using the service?
3. *{Ask any remaining validation questions from the list of service-specific TAP prompts that did not come up during the task performance}*
4. How could the existing features be improved?
5. What would you like to add to the service in the future?

Optional questions from the ALTAI self-assessment list

1. Has it been apparent to you that you've been interacting with an AI system?
 - a. If so, which tasks were completed by the AI system?
2. Do you understand the results of the system?
 - a. If not, what type of information would help you to understand system outputs?
3. Do you trust the result of the system?
 - a. If not, what type of information would you like to see to convince you that this result is correct?
4. Do you have any security concerns while using the system?
5. Do you have any privacy concerns while using the system?

Ending the session [5 mins]

[00:55] Tell user: “Thank you so much for your participation.”

{Add anything you would like the user to know now that the session is over, inform them about the potential follow-up.}