

ITHACA

AI To Enhance Civic Participation

ITHACA

artificial Intelligence To enHance Civic pArticipation

D4.3: Final Evaluation, Impact Assessment, and Recommendations Report

Work Package 4: Pilots’ implementation & Evaluation

Authors:	Katerina Toulidou, Maria Panou, Aristotelis Spiliotis (CERTH), Rares Onescu (Brasov), Michal Stupák (Martin), Adrian Dragota (SIMAVI), Evangelos Rigas (KT), Michael Bedek, Maria Zangl, Erich Weichselgartner, Alexander Nussbaumer, Jonas Seier (UniGraz), Iliana Loi, Panagiotis Zachos (UPAT), Eva De Lera (RtF), Emmanouil Dimogerontakis, Charikleia-Eleni Nikolaou (SNP)
Status:	Final
Due Date:	30.11.2025 (M35)
Version:	Final
Submission Date:	19.12.25
Dissemination Level:	PU

Disclaimer:

This document is issued within the frame and for the purpose of the ITHACA project. This project has received funding from the European Union's Horizon Europe Framework Programme under Grant Agreement No. 101094364. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

This document and its content are the property of the ITHACA Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the ITHACA Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the ITHACA Partners. Each ITHACA Partner may use this document in conformity with the ITHACA Consortium Grant Agreement provisions.

(*) Dissemination level. - Public — fully open (automatically posted online)

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

ITHACA Project Profile

Grant Agreement No.: 101094364

Acronym:	ITHACA
Title:	artificial Intelligence To enHance Civic pArticipation
URL:	https://www.ithaca-project.eu/
Start Date:	01/01/2023
Duration:	36 months

Partners

Short Name	Legal Name	Country
KT	KONNEKT ABLE TECHNOLOGIES LIMITED	IE
CERTH	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	EL
UPAT	PANEPISTIMIO PATRON	EL
RtF	RAISING THE FLOOR	BE
SnP	STAMADIANOS KAI SYNETAIROI DIKIGORIKI ETAIREIA	EL
UniGraz	UNIVERSITAET GRAZ	AT
MNLT	MNLT INNOVATIONS IKE	EL
SIMAVI	SOFTWARE IMAGINATION & VISION SRL	RO
PEDAL	PEDAL CONSULTING SRO	SK
BMA	AGENTIA METROPOLITANA PENTRU DEZVOLTARE DURABILA BRASOV ASOCIATIA	RO
MARTIN	MESTO MARTIN	SK



DOCUMENT HISTORY

VERSION	DATE	CHANGES	RESPONSIBLE PARTNER
0.1	15/10/2025	Table of Contents	CERTH
0.2	07/11/2025	Addition of Performance testing results	SIMAVI, CERTH
0.3	17/11/2025	Addition of Phase 2 Martin results	CERTH
0.4	21/11/2025	Addition of Gamification results	CERTH
0.5	25/11/2025	Addition of T4.5 related chapters	UniGraz (sections 6.1, 6.3 & 6.7), UPAT (sections 6.5 & 6.8), CERTH (sections 6.4 & 6.6), SNP (section 6.2)
0.6	01/12/2025	Addition of Phase 2 Brasov results	CERTH
0.7	01/12/2025	Additions of remaining sections and sent for internal peer review	CERTH
0.8	12/12/2025	Review of the content and feedback provided	KT and SIMAVI
Final	19/12/2025	Integrated feedback and version sent to be submitted to EC	CERTH

TABLE OF CONTENTS

1. Introduction 10

 1.1 Purpose and Scope of D4.3 10

 1.2 Relation to D4.1 and WP4 Objectives 10

2. Overview of the ITHACA Platform..... 12

 2.1 Functionalities and architecture 12

 2.2 Accessibility, inclusiveness and AI components 14

 2.3 Revisions since D4.2..... 16

3. Methodological Framework 17

 3.1 Evaluation Dimensions, Validation Criteria and KPIs..... 17

 3.2 Testing Instruments and Data Collection Tools 18

 3.3 Overall end user evaluation 20

 3.4 Moderator Evaluation 21

 3.5 Gamification module evaluation user sessions 23

4. Results 26

 4.1 Pre-pilot performance testing (Gate A) 26

 4.2 Evaluation of the platform 28

5. Algorithmic Impact Assessment (AIA) 87

 5.1 Scope and approach 87

 5.2 Fairness and coverage of AI-generated summaries 88

6. Revision of the concept..... 96

7. Recommendations 116

 7.1 Technical robustness and performance 116

 7.2 User experience, accessibility and inclusiveness 117

 7.3 AI components, fairness and explainability 118

 7.4 Gamification and meaningful engagement..... 119

 7.5 Governance, policy and capacity-building..... 120

 7.6 Recommendations for future research and transfer..... 121

8. Conclusions..... 121

References 121

Annex 1: Online mini-surveys 129

Annex 2: Focus group material 136

Annex 3: Field Protocols..... 144

Annex 4: Performance Testing Protocol..... 154

Annex 5: Logged Data Request Brief..... 159

Annex 6. Gamification Module evaluation 163

LIST OF FIGURES

Figure 1. Platform topics suggestions and comments (a), AI-summary page (b), Toxicity check page 13

Figure 2. Gamification profile..... 13

Figure 3. Moderator’s dashboard 14

Figure 4. Frequency of participation in online public discussions (Martin)..... 29

Figure 5. Perceived suitability of accessibility functions..... 30

Figure 6. Participants' frequency of reporting problematic online content note (Martin)..... 31

Figure 7. Mean baseline scores for gamification expectations and preferences 34

Figure 8. Summary of mini-survey results (Martin) 36

Figure 9. Themes in participants' responses on what slowed them down 37

Figure 10. Graphical summary of Brasov user experience indicators..... 46

Figure 11. Thematic breakdown of what slowed users down (Brasov)..... 47

Figure 12. Micro-survey results (Gamification module)..... 70

LIST OF TABLES

Table 1. Target service-level objectives for Gate A performance tests..... 26

Table 2. Participation frequency, AI trust and main goals (Martin) 29

Table 3. Use of accessibility and assistive tools in everyday digital life (Martin) 30

Table 4. Participation frequency, AI trust and main goals (Braşov) 32

Table 5. Top-3 user-prioritised fixes (Martin)..... 41

Table 6. Priorities for improvements (Martin) 42

Table 7. Top-3 user-prioritised fixes (Braşov) 50

Table 8. Priorities for improvements (Braşov) 51

Table 9. Perceptions of AI summaries (Martin; Moderators)..... 55

Table 10. Focus group with moderators in Martin (Top-3 organisationally relevant fixes)..... 61

Table 11. Perceptions of AI summaries (Braşov; Moderators) 62

Table 12. Focus group with moderators in Braşov (Top-3 organisationally relevant fixes) 65

Table 13. Priorities and recommendations (Gamification module) 72

Table 14. Summary of legal requirements and their implementation in ITHACA..... 100

Table 15. KPIs, thresholds and Phase 2 evaluation outcome..... 122

ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
AIA	Algorithmic Impact Assessment
ANOVA	Analysis of Variance
API	Application Programming Interface
ARIA	Accessible Rich Internet Applications
AT	Assistive Technology
CEP	Civic Engagement Platform
CPU	Central Processing Unit
CSAT	Customer/Participant Satisfaction
CSV	Comma-Separated Values
DEM	Demonstrator
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
GA	Grant Agreement
GAI	Generative AI
GDPR	General Data Protection Regulation
HCI	Human–Computer Interaction
HMI	Human–Machine Interface
HTTP/HTTPS	Hypertext Transfer Protocol / Secure Hypertext Transfer Protocol
IP	Internet Protocol
IQR	Interquartile Range
KPI	Key Performance Indicator
LOCUST	Load-testing framework (Python-based)
M	Mean (statistical)
MFA	Multi-Factor Authentication
N	Sample size (number of participants)
NLP	Natural Language Processing
RBAC	Role-Based Access Control
REA	European Research Executive Agency
RoPA	Record of Processing Activities
RPS	Requests Per Second
SD	Standard Deviation
SLA	Service-Level Agreement
SLO	Service-Level Objective
SRE	Site Reliability Engineering
SRL	Societal Readiness Level
SSL	Secure Sockets Layer
STT	Speech-to-Text
SUS	System Usability Scale
TLS	Transport Layer Security
TTS	Text-to-Speech
UI	User Interface
URL	Uniform Resource Locator
UUID	Universally Unique Identifier
UX	User Experience
WAI	Web Accessibility Initiative
WCAG	Web Content Accessibility Guidelines
WP	Work Package
XAI	eXplainable AI

EXECUTIVE SUMMARY

This Deliverable provides the final assessment of the ITHACA platform at project closure. Its objective is to determine to what extent the evaluation and KPI targets set in D4.1 have been achieved, using Phase 2 evidence from the pilots in Braşov and Martin, the dedicated gamification and user engagement study at UPAT, and the concept revision work in T4.5 (UniGraz and other partners). The methodology follows a mixed-methods approach that combines technical performance and load tests, structured missions and free exploration by citizens and moderators, surveys and focus groups, platform analytics and authentication logs, curated borderline-item exercises for AI-assisted moderation and targeted discussions for the Algorithmic Impact Assessment (AIA).

In Braşov and Martin, the Phase 2 evaluation focused on how citizens and municipal staff experience the platform in realistic participation scenarios over a two-week period. Participants in both sites were able to complete the main journeys (logging in, locating topics, reading and posting in threads, consulting AI-generated summaries, reacting and reporting) without major technical obstacles. Usability was generally positive, but not yet effortless: users reported navigation complexity on some pages, high information density and difficulty discovering certain controls on first use. Accessibility and inclusiveness improved compared to earlier phases, with older and less digitally confident users able to participate, yet the cognitive effort remains relatively high for vulnerable groups and accessibility features are not always visible or intuitive. Technically, the platform behaved robustly during the pilots, with no prolonged outages, response times within the expected range and no critical failures, although some minor, non-blocking bugs and glitches were observed and documented. All issues and recommended improvements were communicated to the development teams in a standardized format; they were prioritised and, where possible, included suggestions for how the improvements could be implemented.

The gamification and user engagement evaluation at UPAT showed that missions, points, badges and leaderboards can significantly increase short-term motivation and exploration of features. Participants found the gamified interaction more engaging but also raised concerns about fairness and transparency: scoring and ranking rules were often unclear, some missions appeared unrealistic within normal use, and technical instabilities occasionally affected progress. These issues undermined trust in the gamification layer even when the underlying civic content and moderation were acceptable, suggesting that gamification is useful as an optional engagement tool only if its logic is stable, transparent and clearly aligned with meaningful contribution rather than mere activity.

The AIA examined how AI-generated summaries, AI-assisted moderation and, indirectly, gamification-related logic affect fairness, transparency, trust, privacy and stability under load. AI summaries were widely perceived as helpful “shortcuts” to understand long discussions, but both citizens and moderators noted that they can be overly compressed and may under-represent minority or dissenting views; no systematic identity-based bias was observed, but a clear coverage fairness risk emerged. AI-assisted moderation on curated borderline items was broadly aligned with human judgments and did not reveal systematic group-based unfairness, yet moderators identified a strong explainability gap: they lacked concise reasons and policy references for flagged content and therefore found it difficult to audit AI behaviour or justify decisions. Privacy and security safeguards (pseudonymous accounts, protected connections, controlled access to logs) functioned as intended, and performance tests showed stable AI behaviour under increased load, but

moderators called for clearer internal rules on exports, masking and reuse of content. Overall, AI safety and fairness targets are met within the tested scope, while explainability and coverage fairness remain the main weaknesses.

Drawing on these findings and earlier WP4 work, T4.5 revises the ITHACA concept as a civic participation platform in the age of generative AI (GAI). The updated concept stresses inclusion-by-design and co-creation with vulnerable and marginalised groups, frames AI support as a tool that must remain transparent and contestable and situates the platform within a broader governance and capacity-building ecosystem in municipalities, rather than as a purely technical solution.

The main conclusions of D4.3 are that the ITHACA platform is technically robust and functionally mature, with most core performance, security and data-reliability KPIs achieved; that usability and inclusiveness are acceptable but not yet optimal for all user profiles; and that the AI components are valuable decision-support tools but require better explanation and pluralism. The report therefore recommends targeted improvements rather than fundamental redesign: closing the remaining bugs on critical user paths and maintaining performance/AIA checks in future updates; simplifying first-time onboarding and reducing information overload while making accessibility options more visible; adding concise explanations and policy links to AI outputs and stabilising gamification logic; and formalising light-weight local governance rules and training for municipalities on data handling, AI literacy and moderation. These steps provide a realistic roadmap for exploiting and extending ITHACA beyond the life of the project as a technically robust, ethically aware and practically usable civic participation platform.

1. Introduction

1.1 Purpose and Scope of D4.3

D4.3: *Final Evaluation, Impact Assessment, and Recommendations Report* concludes the evaluation cycle of the ITHACA platform and constitutes the main evidence base on how the platform performs in realistic civic contexts. Building on the evaluation framework and KPI targets defined in D4.1 and on the Phase 0 and Phase 1 results reported in D4.2, the present deliverable focuses on Phase 2, i.e. on real-life pilot deployments and extended user engagement with the platform and its AI components as well as revise the original concept based on the findings and provide relevant recommendations.

The purpose of D4.3 is fivefold:

- **To consolidate evaluation evidence** from Phase 2 testing into a coherent account of platform performance, usability, accessibility, inclusiveness and reliability.
- **To conduct and report a structured Algorithmic Impact Assessment (AIA)** for the ITHACA platform, documenting how algorithmic components (e.g. summarisation, toxicity detection, moderation support, risk dashboards) shape user experience, perceived fairness, institutional trust and organisational uptake.
- **To assess the extent to which the platform meets the quantitative and qualitative KPIs** defined in D4.1, updating the KPI status with Phase 2 data and interpreting deviations, improvements, and remaining gaps in relation to the overall objectives of WP4 and the project.
- **To revise the concept** based on the evaluation of the pilots
- **To translate the findings into actionable recommendations**, both technical and organisational, for future refinement, sustainability and potential transfer of the ITHACA platform beyond the project lifetime.

Within this scope, D4.3 integrates data and insights from multiple complementary sources: two-week end-user participation in the pilots, moderator and municipal staff evaluations, online mini-surveys and focus groups, performance and reliability testing conducted by developers under realistic loads, system-level analytics (e.g., login and session logs), and targeted evaluations of specific modules such as the gamification component tested by UPAT. Together, these strands provide a multi-layered picture of how the platform behaves when deployed with real citizens and city administrations and how it supports meaningful, safe and inclusive civic participation.

1.2 Relation to D4.1 and WP4 Objectives

WP4, *Pilots' implementation & Evaluation*, is responsible for planning, executing and synthesising the evaluation of the ITHACA platform across its development and deployment lifecycle. The work package follows a staged logic:

D4.1: Evaluation Framework and KPIs defined the overall evaluation design, including the main dimensions (usability, accessibility, inclusiveness, engagement, trust, AI transparency, technical performance) and their associated indicators and targets, as well as the high-level approach to pilots and data collection.

D4.2: Platform Tests, Pilot Evaluation and Recommendations Report reported on Phase 0 (expert cognitive walkthroughs on early prototypes and accessibility checks) and Phase 1 (controlled

usability and performance testing with participants in experimental settings, plus developer-led load and reliability tests). D4.2 identified key usability and accessibility issues, prioritised fixes by functionality, and documented lessons learned for preparing the pilots.

D4.3 now closes the loop by (a) reporting on Phase 2, in which the platform is used in realistic conditions by citizens and municipal staff in the pilot sites, and (b) integrating all evaluation evidence in a final assessment of whether the ITHACA platform reaches the intended level of maturity and readiness.

In this sense, D4.3 operationalises the WP4 objectives by:

- Documenting how the platform performs when citizens engage repeatedly over time, outside of controlled lab sessions, using both guided “missions” and free exploration;
- Capturing how municipal moderators, policy and communication staff experience the AI-enabled tools in their workflow, including their expectations and concerns around fairness, transparency, privacy and security (Algorithmic Impact Assessment);
- Comparing Phase 2 outcomes with the KPI targets and intermediate results previously reported in D4.1 and D4.2, clarifying where the project has met, exceeded or fallen short of its initial ambitions;
- Providing consolidated, priority-ordered recommendations that directly inform the technical partners, pilot sites and the broader project, including aspects related to gamification and motivation mechanisms tested in collaboration with UPAT.

By doing so, D4.3 also creates an explicit bridge between WP4 and other work packages, notably WP2 and WP3 (platform design and development, AI tools), as well as WP5 and WP6. The AIA perspective embedded in the Phase 2 work ensures that evaluation is not limited to User Experience (UX) scores or technical metrics, but also addresses the societal, ethical and governance implications of deploying AI-enabled civic participation platforms in European cities.

1.3 Structure of the Document

The remainder of this deliverable is organised into eight chapters and annexes.

Chapter 1: Introduction presents the purpose and scope of D4.3, explains its role as the final evaluation and impact assessment of the ITHACA platform, and clarifies how it relates to D4.1, D4.2 and the overall WP4 objectives.

Chapter 2: Overview of the ITHACA Platform that summarises the platform’s main functionalities and architecture, describes the accessibility, inclusiveness and AI components and highlights the key revisions and improvements made since D4.2.

Chapter 3: the Methodological Framework sets out the evaluation dimensions, validation criteria and KPIs, details the data collection tools and instruments, and describes the design of the Phase 2 evaluations with end-users, moderators and the UPAT gamification sessions.

Chapter 4: Results section reports the empirical findings from Phase 2, including participant demographics, pre-pilot performance testing (Gate A), the evaluation of the platform by citizens and moderators in the pilot sites, and the detailed results of the gamification evaluation.

Chapter 5: the Algorithmic Impact Assessment (AIA) presents the conceptual AIA of the platform’s AI components, focusing on fairness and coverage of AI-generated summaries, AI-

assisted moderation, perceptions of transparency, privacy and security, stability under load and a synthesis of AIA-relevant impacts.

Chapter 6: the Revision of the Concept (T4.5) summarises the concept-level refinements proposed in Task 4.5, based on the initial concept outlined in D1.3, drawing on the evaluation results to update the underlying participation and governance model of ITHACA.

Chapter 7: Recommendations consolidates technical, organisational and governance recommendations arising from the evaluation and AIA, addressing both platform development and municipal deployment practices.

Chapter 8: Conclusions provides the overall conclusions at project closure, reflecting on the extent to which the D4.1 KPIs have been met in Phase 2 and outlining remaining gaps and implications for future use of the platform.

References and Annexes: list respectively the literature and project deliverables cited in the report and provide the procedural protocols and the instruments for the pilots.

2. Overview of the ITHACA Platform

The ITHACA platform is a web-based environment designed to support inclusive, AI-enhanced civic participation. It connects citizens, civil society actors and municipal staff around structured debates, enabling the collection, organisation and interpretation of public input on local issues. From a technical perspective, the platform integrates a discussion environment, a gamification layer, AI-enabled tools (summarisation, toxicity analysis and fairness support), accessibility features, a moderation and configuration console for municipalities, and a logging and analytics layer that supports both operations and evaluation. For more information, please refer to D3.2.

2.1 Functionalities and architecture

At user level, the platform is organised around three main spaces:

- **Citizen-facing participation space.** Citizens create an account and log in through the central identity management service (Keycloak). Once authenticated, they can browse topics defined by the municipality, open debate threads, post comments, react to others' contributions (e.g. likes or other reactions) and report content that they perceive as inappropriate. A dedicated panel displays AI-generated summaries of long threads, helping users quickly understand the main points before reading in depth (Figure 1). During Phase 2, this participation space also included a **gamification layer** (points, levels, missions and badges) to incentivise repeated, meaningful contributions rather than one-off visits. Users can see their progress in their profile area (Figure 2) and through contextual cues (e.g., notifications about completed missions).

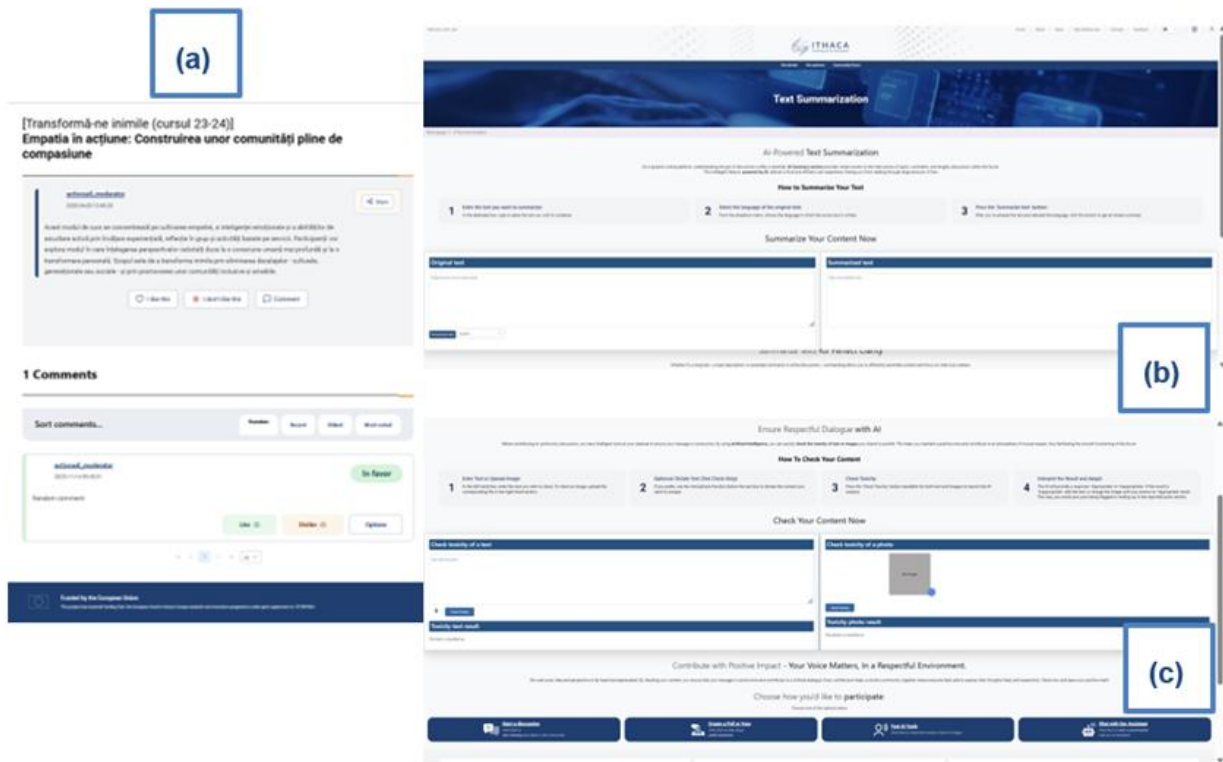


Figure 1. Platform topics suggestions and comments (a), AI-summary page (b), Toxicity check page

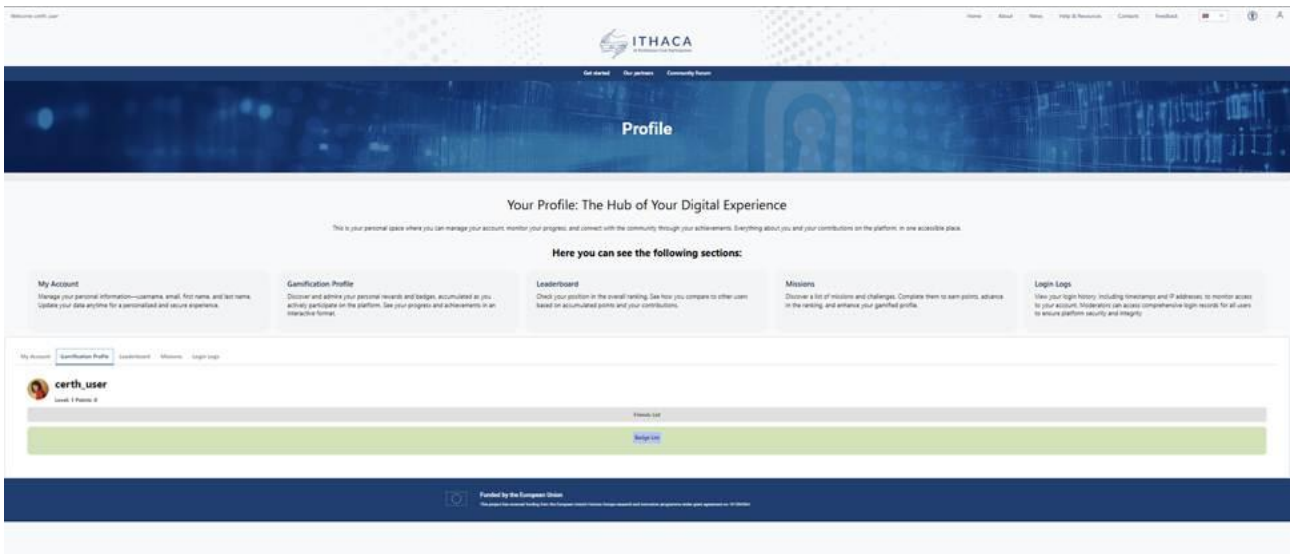


Figure 2. Gamification profile

- Moderator and administrator console.** Municipal staff access a back-office interface (Figure 3) where they configure topics, manage threads, review and moderate content, and monitor platform activity at a higher level. This console exposes AI outputs that support decision-making (for example, suggested toxicity levels for reported items, AI summaries of long debates, and basic indicators related to algorithmic risk and privacy/cybersecurity). In Phase 2, this console also served as an entry point for the Algorithmic Impact Assessment (AIA) exercises conducted with municipal staff, who inspected AI outputs, compared them with their own expectations and reviewed associated risk indicators.

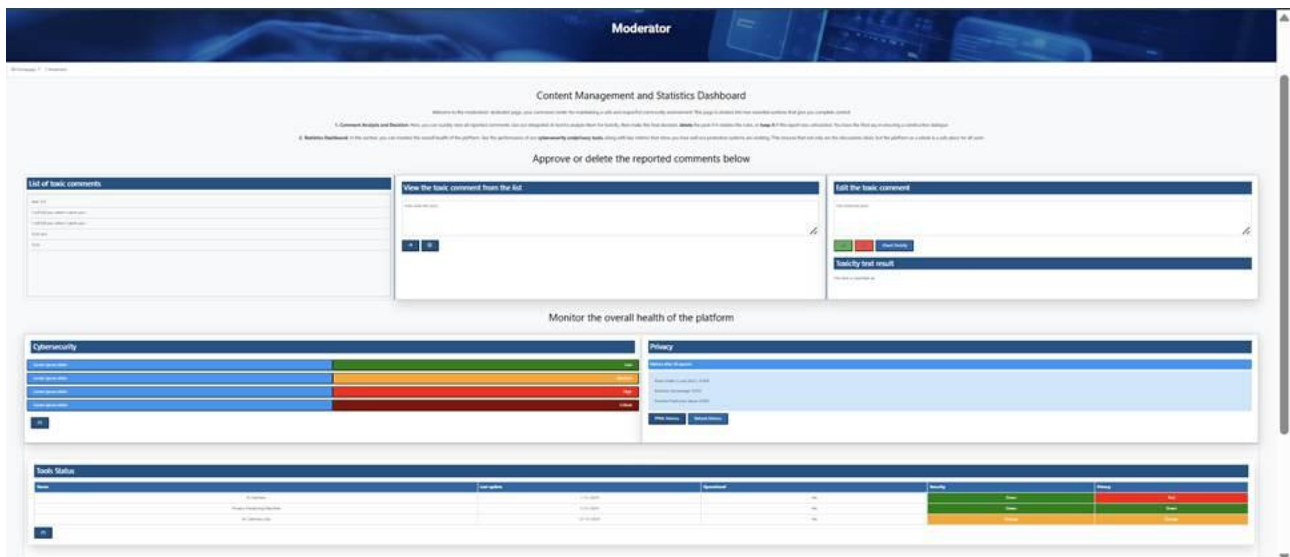


Figure 3. Moderator's dashboard

- Analytics, logging and integration layer.** Underneath the user interfaces, the platform collects pseudonymous usage data through structured logs and event streams. Key entities include login records, session boundaries, user actions (e.g. views, posts, reactions, reports, accessibility toggles) and AI-related events (e.g. summary views, moderation decisions). Dedicated export templates describe how these logs are mapped to evaluation variables such as “active user”, “session”, “summary view” or “accessibility used” and how they can be joined with survey-based identifiers without exposing personal data.

Technically, the ITHACA platform follows a standard multi-tier architecture. A single-page web application provides the front-end interface, communicating with a back-end API layer that manages business logic, persistence and integration with external services. Authentication and authorisation are handled via Keycloak, with role-based access control governing the distinction between citizen, moderator and administrator accounts. AI components (summarisation and toxicity analysis) are exposed as separate services, typically containerised and reachable via internal APIs, which allows them to be tested, monitored and iterated semi-independently from the core application.

For performance and reliability, the back-end is instrumented with metrics relevant to service-level objectives (SLOs) such as latency distribution (p50/p95/p99), error rates, throughput and availability. The Phase 2 performance-testing protocol defines a series of load, spike, stress and (optional) soak tests based on realistic user journeys (login; browsing topic lists and threads; posting and reacting; viewing summaries; requesting moderation decisions) and sets explicit targets for the most critical endpoints (e.g. $p95 \leq 1500\text{--}2500$ ms for key user flows, error rate $\leq 1\%$, and 99.5 % availability in pilot windows).

2.2 Accessibility, inclusiveness and AI components

A central design requirement for ITHACA is that the platform should support **accessible and inclusive participation** across different languages, abilities and levels of digital literacy. To this end, several layers of functionality are relevant for the Phase 2 evaluation:

- Accessibility and personalisation features.** The platform implements a set of user-controlled accessibility toggles, such as increased font size, high-contrast mode, reduced motion and support for keyboard-only navigation and screen readers. These controls are

surfaced in a consistent way across pages (e.g. through a dedicated accessibility menu) and are accompanied by short instructions in the end-user protocol, which explicitly invites participants to try them and to report any barriers encountered. Accessibility usage is captured in the logs through dedicated fields (e.g. feature, state, page area and timestamp), enabling the evaluation to identify which combinations of features are used in practice and whether those who need them encounter specific blockers.

- **Inclusive participation flows.** The interaction design explicitly supports users with different levels of familiarity with civic platforms. Guided missions and structured “journeys” provide simple step-by-step instructions (e.g. “log in, open a specific thread, post a short comment, react to someone else’s post, then complete a one-minute survey”), while the free exploration mode allows more experienced users to navigate by interest. Recruitment materials and in-platform instructions emphasise that participation is voluntary, that responses are anonymised in reports, and that users should avoid sharing personal or sensitive information in public comments.
- **AI-enabled components.** Three AI-related tools are central to the platform and to the Phase 2 evaluation:
 1. **Summarisation tool.** For long discussion threads, an AI-generated summary is displayed in a dedicated panel. The goal is to help both citizens and moderators grasp the main viewpoints quickly, with particular attention to whether minority perspectives are preserved. In Phase 2, this tool was inspected through guided missions for citizens and AIA exercises for moderators, who assessed coverage, usefulness and potential omissions, and compared idle versus under-load behaviour during performance tests.
 2. **Toxicity and content-risk tool.** The platform integrates an AI-based toxicity detector that scores specific phrases or posts and presents them to moderators as auxiliary input when deciding whether to keep or remove content. Citizens were exposed to a simplified version of this tool during missions where they checked predefined terms and judged whether the outcome matched their expectations, while moderators used curated sets of items to probe the system’s consistency and fairness across different identity terms.
 3. **Algorithmic Impact Assessment (AIA) and risk indicators.** Beyond individual predictions, the platform and its surrounding tooling provide higher-level indicators related to fairness, privacy and cybersecurity. Moderators inspected panels that presented metrics such as attacker advantage, AUC and criticality levels (e.g., green/orange/red), and reflected on whether these visualisations were understandable and actionable in the context of their organisational responsibilities. During Phase 2, these indicators were also linked to performance-testing gates, where under-load behaviour of moderation and summarisation was compared to idle baselines using a fixed labelled set of borderline items and representative threads.
- Across all AI components, human-in-the-loop principles are maintained: AI outputs are presented as suggestions or decision aids, while final decisions about moderation, policy interpretation and communication remain with municipal staff. The evaluation therefore focuses not only on technical performance and accuracy, but also on perceived fairness, transparency and trustworthiness of these AI-mediated processes.

2.3 Revisions since D4.2

D4.1 defined the initial evaluation framework and KPIs for ITHACA, while D4.2 reported on Phase 0 (expert walkthroughs) and Phase 1 (controlled usability and performance testing) of the platform. Those earlier stages led to a series of design and implementation recommendations regarding usability, accessibility, clarity of AI explanations, performance thresholds and logging requirements. D4.3 documents how these recommendations were taken forward into a more mature, real-life deployment context. Between D4.2 and the start of the Phase 2 pilots in Braşov and Martin, the consortium implemented several refinements that are directly relevant to this deliverable:

- **Stabilisation and extension of core functionalities.** The discussion environment (topics, threads, posts, reactions and reports) was consolidated based on Phase 1 issues, with particular emphasis on clearer status feedback (e.g. success or error messages for posting and reporting), more predictable navigation between topic lists and threads and smoother session management through Keycloak.
- **Deployment of the gamification module.** Building on the gamification design and scenarios, the platform integrated missions, points and badges aimed at increasing repeated, meaningful engagement. The UPAT evaluation of the gamification module, based on structured scenarios and post-test questionnaires, informed fine-tuning of reward thresholds, feedback messages and the visual prominence of gamified elements, to avoid both under-stimulation and over-gamification of sensitive civic topics.
- **Enhanced accessibility and logging.** Accessibility controls were made more visible and consistent across the interface, while the logging schema was extended to record accessibility-related events and session-level aggregates in a pseudonymous, GDPR-compliant way. Templates were agreed for session duration, summary views, report submissions and accessibility usage, enabling deeper analytics in Phase 2 without collecting personal identifiers.
- **AIA-aware performance and monitoring setup.** In response to Phase 1 findings on performance and AI transparency, the developer team adopted a refined performance-testing protocol with explicit SLOs per critical journey and integrated AIA cross-checks. This ensured that latency and error-rate improvements were evaluated together with the stability and fairness of AI outputs under realistic load patterns, rather than in isolation.
- **Preparation for real-life, multi-session participation.** Finally, guided and free-mode participation protocols were implemented to support repeated, short visits over several weeks instead of single, lab-based sessions. This included the creation of study packs, mission scripts, mini-surveys (A1–A4 for citizens; DEM-1–DEM-6 for moderators) and troubleshooting guidance, which collectively made it feasible to run the Phase 2 evaluations in a low-burden, remotely supported manner.

These refinements place the platform in a substantially more mature state than in Phase 1, both in terms of technical robustness and in terms of instrumentation for evaluation. The following chapter describes in detail the methodological framework used to assess this Phase-2 version of the ITHACA platform, including end-user and moderator protocols, performance-testing procedures and the integration of survey, log and AIA data.

3. Methodological Framework

The evaluation of the ITHACA platform has been designed as a multi-phase, mixed-methods process that gradually moves from expert inspection and controlled tests to real-life use in the pilot cities. Phase 0 (expert walkthroughs) and Phase 1 (controlled usability and performance tests) are documented in detail in D4.2 and provide the initial picture of platform maturity, accessibility and AI-enabled functionalities. The present deliverable extends this framework to Phase 2, focusing on the longitudinal, in-the-wild use of the platform in the pilot sites, combined with structured moderator sessions, performance testing with Algorithmic Impact Assessment (AIA) checks and a dedicated evaluation of the gamification module conducted by the University of Patras.

The methodological framework is grounded in the evaluation dimensions and KPI structure originally defined in D4.1 and operationalised in D4.2, namely usability and learnability, accessibility and inclusiveness, usefulness and relevance, trust and AI transparency, engagement and motivation, and technical performance and reliability. These dimensions are complemented, in Phase 2, by a more explicit AIA perspective, with attention to fairness, privacy and security, and by a stronger focus on inclusive participation and accessibility, in line with WCAG guidelines and the project's ethics and data protection requirements. The different strands of data collection (surveys, mini-surveys, focus groups, platform logs and performance tests) are treated as complementary views on how citizens and municipal staff experience the platform in realistic conditions and their protocols can be found in Annex 4.

3.1 Evaluation Dimensions, Validation Criteria and KPIs

The dimensions used to structure the Phase 2 evaluation are consistent with those defined earlier in the WP, which enables comparison across phases and a coherent view of the platform's evolution. Usability and learnability refer to the perceived ease of use of the platform, the clarity of labels and navigation paths and the effort required to complete core journeys such as logging in, locating topics, reading threads, posting comments, reacting to others and reporting problematic content. Indicators for this dimension include task success, self-reported ease-of-use scores and descriptions of friction points gathered through mini-surveys and focus groups.

Accessibility and inclusiveness capture whether the platform can be effectively used by participants with different abilities, devices and levels of digital literacy and whether accessibility features such as increased font size, high contrast mode, reduced motion and keyboard-only navigation are discoverable and usable. Here, the analysis combines accessibility mini-survey responses with log-based indicators of accessibility feature use, allowing the evaluation to examine not only perceived accessibility but also actual uptake of the available options.

Usefulness and relevance concern the perceived added value of the platform in helping users follow debates, understand complex issues and contribute meaningfully to civic discussions. For citizens, this includes whether the platform supports them in forming an informed opinion and expressing it. For moderators and municipal staff, it covers the integration of ITHACA outputs (particularly AI-generated summaries) into their concrete tasks, such as preparing briefings or public communications.

Trust, fairness and AI transparency are central in Phase 2, given the presence of AI-based summaries and toxicity detection tools. For end-users, questionnaires and mini-surveys ask whether summaries are experienced as balanced, accurate and respectful of less popular or minority viewpoints, and whether content rules and moderation decisions appear consistent. For moderators, dedicated mini-surveys and AIA tasks explicitly examine the perceived fairness and stability of

moderation support, including through the inspection of borderline items. These exercises are intended to reveal whether the AI behaves in a way that aligns with the municipality's values and expectations.

Engagement and motivation encompass willingness to return to the platform, perceived enjoyment and the extent to which the interaction encourages meaningful participation rather than passive browsing. In Phase 2, this dimension is enriched by the gamification evaluation. The game layer (missions, points, badges and related feedback) is examined not only in terms of sheer motivation, but also in relation to fairness and inclusiveness, for example whether it may inadvertently favour certain user profiles or undermine the seriousness of sensitive topics.

Finally, technical performance, reliability and AIA stability describe how the system behaves from an operational perspective. Developer-led performance tests measure latency distributions, error rates and availability for critical services, while platform logs provide information on session lengths, time spent on key pages, and technical errors experienced by users. AIA stability is monitored by repeating tests on a fixed set of labelled borderline items and representative threads before and during high-load scenarios. Changes in moderation outputs or summary content under load are tracked to ensure that performance optimisations and stress conditions do not introduce new risks or biases.

Each of these dimensions is linked to a set of quantitative and qualitative indicators. Survey-based indices provide mean scores and distributions for each dimension among citizens and moderators. Behavioural indicators derived from logs capture usage patterns such as the number of sessions per user, session duration, counts of summaries viewed or reports submitted, and use of accessibility features. Performance reports summarise latency and error rates for the main services. AIA stability indicators describe, for example, the percentage of moderation decisions that change when the system is under load, or the degree to which minority viewpoints remain present in summaries across conditions. Qualitative material from focus groups and open comments is used to contextualise these numbers and to identify themes and examples that can feed directly into recommendations (see Table 15).

3.2 Testing Instruments and Data Collection Tools

To operationalise this framework in Phase 2, the project uses a coordinated set of protocols, questionnaires, mini-surveys, focus group guides and logging templates, tailored to the different actors involved. For end-users, there is a field study protocol that invites participants to interact with the live platform from their own environments over a period of about two weeks. Each participant is asked to complete a series of short visits rather than a single long session, alternating between guided missions and free exploration. The guided missions focus on specific functionalities, such as logging in and navigating to a given topic, posting a comment, reacting to someone else's contribution, reading and evaluating an AI-generated summary, trying out accessibility features or reporting a problematic item. Free sessions allow participants to explore any topics of interest, simulating a more natural pattern of civic engagement.

Around these interactions, a set of short online questionnaires is used. A baseline survey, completed once at the beginning, records demographics, prior experience with civic participation tools and with AI-based services and initial attitudes towards online participation and trust. After each visit, a very short post-visit mini-survey captures immediate impressions of ease of use, any technical or interaction problems, perceived fairness of AI-supported features and willingness to return. At the end of the two-week period, a final survey asks participants to reflect on their overall experience, including usability, trust, fairness, inclusiveness, motivation and perceived impact. For participants

who use accessibility features or who report specific barriers, an additional short accessibility questionnaire is available to capture more detailed information on the nature of the difficulties and possible improvements.

For municipal staff, Phase 2 uses a structured remote protocol that typically lasts 45 to 60 minutes and is conducted either individually or in small groups. Participants are asked to identify a realistic task from their usual work (for example preparing a briefing, moderating a contentious thread or reflecting on citizen feedback) and to use the platform to support that task. Within the same session, they examine one or more AI-generated summaries of longer discussions and evaluate their usefulness, coverage and potential reuse. They also engage with curated sets of several content items to compare their own moderation judgments with the suggestions of the toxicity detection tool. Additional prompts address privacy, security, organisational uptake and the conditions under which the municipality would adopt the platform in a sustained way. These activities are supported by short mini-surveys that capture both factual information about the participants' roles and their perceptions of the different AI-enabled functionalities.

The focus groups, which are conducted separately with citizens and with moderators or demonstrators, provide a space for more in-depth discussion of the same themes. Using semi-structured guides, they explore perceptions of ease and friction in key journeys, the perceived fairness and clarity of AI-supported moderation and summaries, accessibility experiences, and priorities for improvement. Simple prioritisation techniques (for instance, asking participants to select the three most important issues) help transform these discussions into concrete lists of recommended changes.

Alongside self-report instruments, Phase 2 makes use of platform usage logs. A dedicated logging specification defines how pseudonymous user identifiers and session identifiers are created, which events are recorded, and how these data are mapped to evaluation variables. The logs capture, among other things, the start and end of sessions, page views, posts, reactions, reports, summary views and changes in accessibility settings. They also include fields that allow specific AIA-related tests, such as the behaviour of the system on predefined borderline items, to be traced in a consistent way. All logging is designed to comply with the project's data protection and ethics framework; personal data such as names or email addresses are not included in the evaluation exports and retention periods and access rights are controlled.

Performance testing in Phase 2 follows a protocol agreed between the evaluation team and the developers. It includes load, spike, stress and, where relevant, soak tests for the main user journeys, with clearly defined service level objectives for latency, error rates and availability. These tests are executed at different "gates", for example before the start of the pilots, during the pilot period, and at the end. Each test run is accompanied by an AIA cross-check using a fixed set of labelled borderline items and representative discussion threads. By comparing moderation outputs and summary content across runs and conditions, the team can detect whether system behaviour remains stable when the platform is under stress or whether performance-related changes risk undermining fairness, transparency or user trust.

Finally, the gamification module is assessed through a dedicated methodology implemented by the University of Patras. Participants are asked to complete scenarios that make use of the game mechanics and their experiences are captured through questionnaires and qualitative feedback. The evaluation looks at usability and clarity of the game elements, their effect on motivation and repeated use, and their perceived fairness and appropriateness for civic participation. The findings from this strand are integrated into D4.3 alongside the main Phase 2 results, ensuring that the motivational

layer of the platform is considered together with usability, accessibility, AI transparency and performance.

Two complementary methodological streams were implemented:

- the End User Evaluation, focusing on the general public's interactions with the platform through repeated, self-paced use; and
- the Moderator Evaluation, focusing on municipal staff responsible for reviewing, moderating, or interpreting public input within their professional roles.

Both strands were aligned under a shared framework of the AIA, encompassing fairness, transparency, privacy, security and societal implications of algorithmically assisted decision-making. All evaluation and reporting material can be found in Annexes 1-6.

3.3 Overall end user evaluation

3.3.1 Objectives and Design

The end-user study examined how citizens engage with the ITHACA platform, including its accessibility features, gamified interactions and AI-enabled functionalities such as summarisation and toxicity detection. The goal was to assess usability, satisfaction, inclusiveness, and trustworthiness across multiple, short user sessions that simulate realistic participation patterns in civic discourse.

Participants were invited to complete at least six logged-in visits over a two-week period. Each visit lasted approximately 10–15 minutes and required at least one interactive action (e.g., posting a comment, reacting to a post, or reporting content). Two participation modes were alternated to balance structured and spontaneous exploration:

- **Guided mode**, where participants followed pre-defined “missions” that tested specific functionalities, and
- **Free mode**, allowing them to explore topics of personal interest in an unstructured way.

A typical participant trajectory comprised the following:

- **A1: Baseline survey** (first login)
- **A2: Post-visit survey** (after each session)
- **A3: Quick accessibility survey** (once, or each time an accessibility barrier was encountered)
- **A4: End-of-study survey** (final reflection after ≥ 6 visits)

Each survey collected structured feedback on usability, satisfaction, accessibility, and perceptions of fairness and inclusiveness, ensuring consistency across pilot sites.

3.3.2 Recruitment and Consent

Participants were recruited through local municipal channels and community outreach activities in Braşov and Martin. All participants received a study pack including the platform URL (<https://ithaca.simavi.ro/>), brief instructions, survey links and contact details for technical support. Participation was voluntary, with written informed consent obtained prior to enrolment. The consent form explicitly noted that all responses would remain anonymous and that users should refrain from posting personal or sensitive data in public comments.

3.3.3 Guided Missions and Tasks

During the guided phase, users were asked to perform short, well-defined missions designed to expose them to different platform features and assess their intuitiveness and reliability. The sequence of tasks was rotated to mitigate order effects:

- **Mission 1: Join and Contribute.** Log in, open a discussion thread, post a short comment (1–2 lines) and react to another user’s post.
- **Mission 2: Read a Long Thread with Summary.** Open a thread with an AI-generated summary, read it fully, review 3–5 original posts beneath it and note whether the summary was accurate and representative.
- **Mission 3: Test the Toxicity Tool.** Examine three borderline examples of potentially toxic expressions and note whether the system’s output matched their expectations.

Accessibility prompts were included in all instructions, encouraging participants to try larger text, high contrast, keyboard-only navigation, or screen reader modes. Any difficulties encountered were reported via the accessibility mini-survey (A3).

3.3.4 Free Exploration and Wrap-Up

In the second phase of participation, users were encouraged to explore the platform freely, choosing topics aligned with their interests or local issues. They could read, react or post within discussions, and were invited to use the “Report” function if they encountered content perceived as inappropriate. The final visit combined open exploration with an end-of-study reflection (A4), summarising overall impressions of usability, trust, fairness, and enjoyment.

Completion criteria for inclusion in the analysis were defined as:

- Baseline survey (A1) submitted,
- At least six distinct visits each followed by a post-visit survey (A2),
- Accessibility feedback (A3) where applicable and
- End-of-study survey (A4) submitted.

3.3.5 Data Capture and Validation

A researcher team list was maintained for each participant, documenting session completion, survey status and notes on accessibility barriers or technical issues. System logs from the platform were used to validate the number and duration of visits, ensuring that “a visit” corresponded to a meaningful engagement (≥ 5 minutes or at least one recorded action). Data hygiene checks required a one-to-one correspondence between visits and A2 entries, with 10% of cases randomly verified for timestamp consistency. Qualitative feedback from open-ended survey fields was coded thematically to identify recurring usability, inclusivity or comprehension issues. The mini-surveys can be found in Annex 1.

3.4 Moderator Evaluation

3.4.1 Objectives and Design

The moderator study assessed the ITHACA platform from the perspective of municipal staff responsible for civic participation, communication and policy feedback processes. The sessions aimed to evaluate (a) the professional relevance and usability of AI-enabled summaries and

moderation tools, and (b) the platform's compliance with principles of fairness, transparency and accountability, in line with the Algorithmic Impact Assessment framework.

Each pilot city (Braşov and Martin) hosted two moderated sessions with 3–6 municipal representatives or one-on-one interviews where group sessions were not feasible. Sessions lasted approximately 45–60 minutes and followed a structured sequence combining hands-on demonstration, reflection and survey completion. All moderators had pre-verified login access and were guided by a facilitator who documented results in real time.

3.4.2 Sessions Structure

The moderator evaluation followed a consistent “run-of-show” sequence across all sessions:

1. **Consent and Baseline (DEM-1).** A short pre-survey captured participant role, department, and prior experience with AI systems.
2. **Scenario Definition.** Each moderator identified a real administrative or communication task where public input was relevant (e.g., preparing a briefing, setting an agenda item or moderating a public thread).
3. **Summary-in-Workflow (AIA: Coverage).** Participants examined an AI-generated summary of a long discussion thread. They were asked to rate its utility, identify any missing or misrepresented views (especially minority opinions), and indicate how they would use or adapt the summary in their work (e.g., insert into a briefing or public statement). Responses were captured in DEM-2.
4. **Moderation Consistency (AIA: Fairness & Stability).** Moderators reviewed 3–6 content items. For each, they indicated whether to keep, remove or escalate the post, and whether the AI system's decision matched their own judgment. These observations were logged in DEM-3 and directly informed fairness and bias analysis.
5. **Privacy and Security (AIA).** Discussion focused on which elements (names, IDs, quotes) should be masked in exports and what safeguards (e.g., audit trails, watermarking, rate limits, or moderation queue rules) would increase trust. Results were captured in DEM-4.
6. **Uptake and Wrap-Up.** Participants identified use cases where the platform could enhance internal workflows and collectively prioritised the top three short-term fixes to improve readiness for deployment. These items, together with assigned responsibilities and deadlines, were recorded in DEM-5 and DEM-6.

3.4.3 Moderator Dashboard Evaluation

Moderators also evaluated the functionality and transparency of the Moderator Dashboard, the central interface for managing user activity and monitoring algorithmic behaviour. The dashboard integrates information from several AI tools to support fair and informed decision-making. Participants performed two short practical exercises:

- **Privacy and Cybersecurity Metrics Inspection.** Opening the Privacy and Cybersecurity panels, moderators reviewed quantitative indicators such as Attacker Advantage, AUC, or criticality levels (green/orange/red). They assessed whether these visual cues were intuitive and sufficient to inform risk decisions.
- **Add and Moderate a Topic.** Moderators created a test topic (e.g., *Improving Local Public Spaces*), published it, and approved it through the dashboard. This task verified that the moderation workflow (creation, approval and public visibility) was seamless and transparent.

Following these exercises, moderators completed a mini-survey assessing ease of navigation, clarity of AI explanations, confidence in decisions based on AI outputs and suggestions for interface improvement. The mini-surveys can be found in Annex 1.

3.4.4 Data Capture and Outputs

During each session, a designated researcher or facilitator maintained a structured note sheet with sections for:

- Scenario description,
- Uptake examples (feature → step helped → expected next use),
- Summary coverage (utility rating, missing views, intended destination),
- Moderation fairness and counterfactual observations,
- Privacy and security feedback and
- Top-3 fixes.

3.4.5 Integration and Analytical Approach

The two strands of evaluation (end user and moderator) were analysed both independently and jointly to capture the interaction between public perception and institutional trust. Survey data (A1–A4; DEM-1–DEM-6) were quantitatively summarised and compared across pilot sites, while qualitative observations were coded according to the five AIA dimensions: fairness, transparency, privacy, security and inclusion. System logs and backend data were triangulated with survey and observation results to validate engagement metrics and detect any discrepancies between reported and actual platform use.

This mixed-method approach allowed the consortium to assess not only *how* the ITHACA platform performed technically and functionally, but also *how* its algorithmic decisions were perceived and acted upon in civic practice. The result is an integrated evaluation framework linking user experience, municipal readiness, and algorithmic accountability, central pillars of the Phase 2 analysis presented in the following chapters.

3.5 Gamification module evaluation user sessions

3.5.1 Objectives and Design

The gamification module evaluation focused on how users experience the game mechanics integrated into the ITHACA platform, including missions, points, levels, badges and leaderboards. The primary objective was to examine whether these mechanics support sustained and meaningful engagement with civic discussions, and how they influence perceived motivation, enjoyment, fairness and clarity of contribution rules. A secondary objective was to identify usability issues and edge cases before wider deployment in pilot sites, ensuring that the game layer does not introduce new barriers or biases.

To this end, a series of structured test sessions was organised at the University of Patras (UPAT) with a cohort of student participants using a dedicated “sandbox” environment of the platform. Each session combined (a) hands-on interaction with the gamification features through predefined missions and free play and (b) self-report questionnaires capturing expectations, in-session experiences and overall perceptions at the end. The evaluation was designed as an intensive, single-study format, where participants completed the full sequence (baseline, missions, mid-session mini-survey and post-session assessment) within one extended visit rather than over multiple weeks.

Four core instruments were used to structure the data collection:

- a pre-session questionnaire capturing demographics, prior experience with games and civic participation platforms, and baseline motivation and expectations regarding gamified participation;
- a short mid-session mini-survey focusing on clarity of missions, perceived cognitive load, emerging enjoyment and any early confusion or frustration;
- a post-session questionnaire covering overall usability, enjoyment, perceived fairness, competitiveness versus collaboration and willingness to use such a system in real civic processes; and
- a micro-survey on accessibility and barriers, completed when participants encountered visual, interaction or comprehension difficulties related to the gamification elements.
- Together with system logs on missions, points and actions, these instruments provided a coherent view of how the gamification layer behaved under realistic but controlled use.

3.5.2 Recruitment and Consent

Participants were recruited from the University of Patras, primarily among students familiar with online platforms and social media but not necessarily with civic participation tools. Invitations were circulated via internal university channels and direct contacts, briefly describing the purpose of the study as an evaluation of a gamified civic participation platform. Inclusion criteria focused on basic digital literacy and the ability to follow written instructions in English; no prior gaming expertise was required.

All participants received an information sheet explaining the study goals, the structure of the session, the types of data collected (questionnaire responses and interaction logs) and their rights regarding withdrawal and data protection. Written informed consent was obtained before any data collection took place. To protect privacy, participants used pseudonymous test accounts on the platform and no personally identifiable information was stored in the system logs. Questionnaires were linked to platform activity only through these pseudonymous identifiers, in line with the data protection and ethics procedures defined in WP6.

3.5.3 Gamified Missions and Tasks

The core of the evaluation consisted of a series of short, goal-oriented missions designed to expose participants to the main gamification mechanics and to typical patterns of civic interaction. Each mission combined a concrete platform action with an associated game element (e.g. points, badges, progress feedback), enabling the researchers to observe both usability and motivational effects.

Typical missions included:

- **Orientation and Profile Setup.** Users had to log into the platform with the assigned test account, access the profile area and complete basic profile information needed for the gamification layer (e.g. avatar, display name), while observing how points and levels are displayed.
- **Participate in Discussions under Gamified Feedback.** Users joined an existing discussion thread, posted at least one comment and reacted to other users' content, then checked how these actions were reflected in points, missions progress and any badges or streak indicators.
- **Complete Themed Missions.** Users followed a set of predefined "civic missions" (e.g., "contribute to two different topics", "react to three posts from other users", "return to a thread

you commented on before”) and note how clearly the mission objectives, progress bars and rewards are communicated.

- **Explore Social and Competitive Elements.** Users reviewed the leaderboard, checked friend or follower options (where available), and any notifications related to achievements, reflected on whether these mechanisms felt fair, transparent and motivating or whether they introduced pressure or unhealthy competition.

Throughout the missions, facilitators prompted participants to verbalise any confusion about what actions earn points, how missions are completed and how rewards are calculated. Specific attention was paid to whether the game layer remained aligned with the underlying civic goals (discussion, deliberation, constructive feedback) rather than overshadowing them with purely competitive dynamics. The evaluation and reporting material can be found in Annex 6.

3.5.4 Session Flow and Wrap-Up

Each gamification evaluation session followed a consistent flow to ensure comparability across participants. After welcoming and consent, participants first completed the pre-session questionnaire, which gathered demographic data, digital and gaming habits, and initial expectations regarding gamified civic participation. The facilitator then introduced the platform briefly, emphasising that the focus of the study was on the game mechanics rather than the substantive content of the discussions.

The main body of the session combined guided and semi-free interaction. Participants worked through the predefined missions in sequence, with the facilitator available to clarify task wording but not to coach them on how to “optimise” their game performance. After the first set of missions, participants completed the mid-session mini-survey to capture early impressions of clarity, enjoyment, workload and perceived alignment between missions and civic purpose. They then continued with further missions and free exploration, using the platform in a more self-directed manner while still being encouraged to trigger and observe different gamification events (e.g., level changes, badges, leaderboard updates).

At the end of the interaction period, all participants completed the post-session questionnaire and, where relevant, a short micro-survey on any accessibility or interaction barriers they experienced (e.g. interpreting progress bars, reading labels, navigating leaderboards). A brief debriefing discussion concluded the session, allowing participants to summarise in their own words what worked well, what felt confusing or unfair and how they would imagine using such a gamified system in a real municipal context.

3.5.5 Data Capture and Validation

Data collection combined subjective self-reports with objective interaction logs. The pre-session, mid-session and post-session questionnaires, together with the accessibility micro-survey, were administered via online forms and exported as structured datasets (pre-questionnaire, mid mini-survey, post-questionnaire and micro-survey files). These captured ratings on usability, enjoyment, perceived fairness and inclusiveness along with open-ended comments on specific game elements.

In parallel, platform logs recorded detailed traces of user activity, including logins, mission events, points earned, completed actions and technical error messages. These logs allowed the research team to verify that the missions had been completed as intended, to reconstruct the sequence of actions undertaken by each participant and to compute derived indicators such as time to complete a mission, number of actions per mission and distribution of points across different behaviours.

Data validation proceeded in several steps. First, questionnaire records were checked for completeness and consistency, ensuring that each pseudonymous participant had a full pre- and post-session entry and, where applicable, a mid-session mini-survey and accessibility micro-survey. Second, platform logs were filtered to include only the dedicated UPAT test accounts and the time window of the evaluation sessions. Third, questionnaires and logs were linked through the pseudonymous identifiers to enable integrated analyses of how subjective perceptions (e.g., enjoyment, fairness) related to objective behaviour (e.g. missions completed, points accumulated). Finally, open-ended responses and qualitative notes from the debriefings were coded thematically to identify recurrent usability problems, misconceptions about the scoring logic and broader reflections on the role of gamification in civic participation.

4. Results

4.1 Pre-pilot performance testing (Gate A)

Before the start of Phase 2 in Braşov and Martin, the development team carried out a dedicated round of performance tests (Gate A) on a production-like environment of the ITHACA platform. The objective was to verify that the core user journeys would remain responsive and reliable under the expected pilot loads, and to establish a baseline for subsequent performance checks during and after the pilots. In line with the Phase 2 performance-testing protocol (Annex 4), these tests combined classical load, spike and stress scenarios with an Algorithmic Impact Assessment (AIA) cross-check, focusing on the stability and fairness of AI-based summaries and moderation under load.

The system under test included the web application, back-end API, worker processes, database, cache and queue infrastructure and the AI inference endpoints responsible for summarisation and moderation. Test scripts were designed to emulate realistic sequences of actions by multiple concurrent users. Four synthetic journeys were used: (J1) login, open the topic list, open a thread, read and react; (J2) login, open a long thread and view the AI-generated summary; (J3) login, post a comment, report a borderline item and retrieve a moderation decision; and (J4) repeated fetching of static assets to check cache behaviour and front-end delivery. Think-time between requests was randomised to approximate real user behaviour.

The concurrency targets were derived from the anticipated number of active participants per site, their likely overlap in time and a safety margin. For the pre-pilot window, this resulted in a “small-to-medium” virtual-user band in the protocol terminology, with load and spike tests exploring both the nominal concurrency and short peaks above it. Each Gate A run followed a structured sequence: a warm-up period, a sustained load test with ramp-up and hold phases, a spike test (rapidly increasing concurrency for a few minutes) and at least one stress step to identify where latency and error rates began to degrade. Where feasible within the agreed window, a short failure simulation (e.g. temporarily cutting a node or network link) was included to observe recovery and error surfacing.

Service-level objectives (SLOs) were defined in advance for the main journeys and for overall reliability. Table 1 summarises the target values used in Gate A, as formalised in the developer protocol.

Table 1. Target service-level objectives for Gate A performance tests

Journey / metric	Target (p95 or aggregate)
Login	p95 ≤ 1500 ms

Journey / metric	Target (p95 or aggregate)
Topic list / Thread view	p95 ≤ 2000 ms
Post comment / React	p95 ≤ 2000 ms
View AI summary (serve or generate + serve)	p95 ≤ 2500 ms
Moderation decision (API)	p95 ≤ 500 ms
Overall error rate (all journeys combined)	≤ 1 %; 0 critical errors
Availability in pilot windows	≥ 99.5 %
Recovery after simulated failure	< 5 minutes; graceful error messages to the user

In addition to these performance-oriented SLOs, Gate A incorporated an explicit AIA cross-check. A fixed set of 40 borderline content items was used to probe moderation consistency under different load conditions. The same items were scored first at idle and then under peak load, and the flip-rate between baseline and load decisions was computed. The protocol set a threshold of at most 5 percentage points for this flip-rate, with particular attention to any systematic differences related to identity terms or dialect. In parallel, two long threads with clearly identified minority viewpoints were selected for summary-coverage checks. For each thread, summaries generated at idle and at peak load were compared to verify that key minority-view sentences remained present and that any use of fallbacks (for example, cached summaries) did not result in the systematic omission of less popular perspectives.

Across the Gate A runs conducted before the launch of Phase 2, the platform was able to sustain the targeted small-to-medium concurrent-user loads with latency and error rates within the SLO ranges defined in Table 1. For all core journeys, p95 response times remained below the corresponding thresholds for the tested load profiles, and no critical errors or persistent instability were reported in the Gate A test summaries shared by the development team. Where temporary degradations appeared in stress steps beyond the nominal concurrency range, they were confined to those extreme conditions and served to identify safe operating margins rather than blocking the start of the pilots. From a reliability perspective, the system demonstrated stable behaviour during the sustained load periods and recovery from brief simulated failures remained within the targeted few-minute window, with user-facing errors presented in a controlled, “graceful” form rather than as raw technical traces.

The AIA cross-checks performed as part of Gate A did not reveal any systematic regressions in AI behaviour under load. Moderation decisions for the fixed set of borderline items remained consistent between idle and peak-load runs within the protocol’s flip-rate threshold, and there was no evidence of new biases emerging along the identity-related slices inspected by the team. Likewise, the comparison of idle and load summaries for the curated long threads indicated that minority-view sentences continued to be represented; any differences observed between versions were limited to phrasing and ordering rather than involving the removal of specific viewpoints. No fairness-related fallbacks (for example, switching to a less representative cached summary under load) were flagged as problematic in the Gate A reports.

Taken together, these Gate A results provided assurance that the ITHACA platform could support the planned Phase 2 evaluations from a technical-performance and algorithmic-stability perspective. They also established a quantitative baseline against which later performance tests (Gate B during the pilot period and Gate C towards the end) can be compared. The subsequent sections of this

chapter focus on how users and moderators in Martin and in the UPAT gamification study experienced the platform once it was deployed in real-life settings, building on the pre-pilot performance guarantees described here.

4.2 Evaluation of the platform

4.2.1 Demographic information

Participant profile and background

This section summarises the profile and background characteristics of the participants involved in the Phase 2 evaluation and the gamification module sessions. The emphasis is on participation habits, prior experience with AI and gamified systems and use of accessibility tools, as these variables frame how users approach the platform and interpret its AI-enabled features. Traditional socio-demographic variables such as age and gender were not collected in these specific instruments, which limits the level of detail but still allows a meaningful description of the samples' baseline profiles.

Martin pilot – participation profile and digital/accessibility background

Baseline information for the Martin pilot was collected through the A1 mini-survey administered at the start of the Phase 2 field study. In total, 17 valid baseline responses were recorded. Due to a small number of repeated completions, this corresponds to a slightly smaller number of distinct pseudonymous user IDs. The identifiers indicate that the sample consists predominantly of citizens, with at least one account associated with municipal staff. All respondents demonstrated sufficient digital skills to access a web-based platform and complete online questionnaires without assistance.

A first indicator concerns habitual participation in public online discussions. When asked how often they usually take part in public online discussions, most respondents reported that they engage rarely or not at all. As summarised in Table 2, 7 participants (41.2%) stated that they never take part, another 7 (41.2%) reported that they rarely do so, and only 3 participants (17.6%) indicated that they sometimes participate. No respondent reported frequent participation. If these three categories are coded from 1 (“never”) to 3 (“sometimes”), the resulting frequency score yields a low mean of $M = 1.76$ ($SD = 0.75$), confirming that the Martin sample is largely composed of individuals who are not regular contributors to public online debates. This is an important contextual factor when interpreting later findings on engagement, motivation and learnability.

Participants were also asked about their prior trust in AI-generated summaries and rules governing online platforms. Responses were provided on a numerical scale, with higher values representing higher trust. More than half of the respondents ($n = 9$, 52.9%) selected the value 2, 6 respondents (35.3%) chose 3, and only 2 respondents (11.8%) selected 4. On this scale, the mean trust score is $M = 2.59$ ($SD = 0.71$), indicating that, at baseline, participants approach AI-generated content with cautious or low-to-moderate trust rather than perceiving it as naturally reliable or neutral.

A third background question explored the participants' main goal for using the platform during the evaluation session. The most frequently selected goal was to make use of the summarisation functionality. Approximately 56% of respondents ($n = 9$) reported that their primary aim was to quickly capture the key points of the discussion via the summary. A further 31% ($n = 5$) indicated that they mainly wanted to read the opinions of others, while only 13% ($n = 2$) reported that their main goal was to share their own opinion. Valid responses for this item were slightly lower than for the other baseline questions (16 instead of 17), due to one missing value. Overall, this pattern suggests that at the outset participants position themselves primarily as readers and information seekers rather than as active contributors and that they are particularly interested in using AI-generated summaries

as a shortcut to understanding the discussion dynamics. Table 2 summarises these core background indicators for the Martin pilot.

Table 2. Participation frequency, AI trust and main goals (Martin)

Indicator	Categories / scale	n	%	Notes / summary measure
Frequency of public online discussions	Never	7	41.2	Coded 1–3 (never–sometimes): $M = 1.76, SD = 0.75$
	Rarely	7	41.2	
	Sometimes	3	17.6	
Prior trust in AI-generated summaries/rules	Value 2	9	52.9	Higher values indicate higher trust; $M = 2.59, SD = 0.71$
	Value 3	6	35.3	
	Value 4	2	11.8	
Main goal for today’s use of the platform*	Quickly learn key points via summary	9	56.3	Valid N = 16; participants mainly information seekers
	Read others’ opinions	5	31.3	
	Share own opinion	2	12.5	

*Percentages are based on valid responses (N = 16) due to one missing value.

A compact visual overview is given through Figure 4, presenting side by side the distribution of frequency of online discussions and the distribution of AI trust values. The figure highlights the prevalence of “never/rarely” participation and low-to-moderate baseline trust in AI-generated content within the Martin sample.

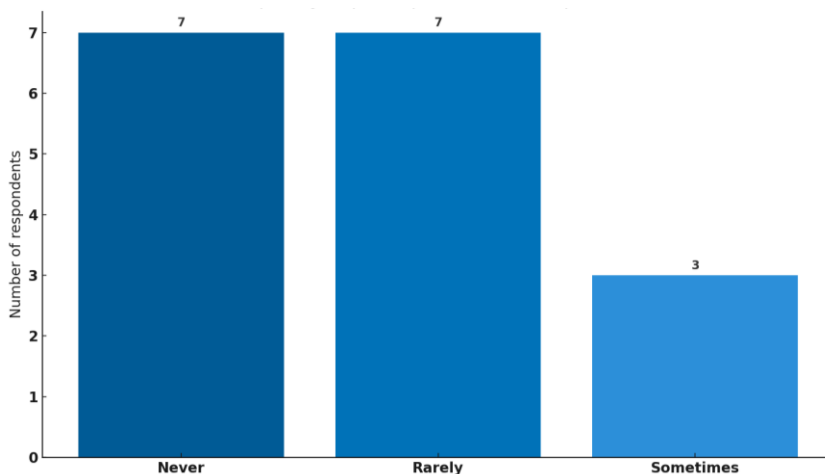


Figure 4. Frequency of participation in online public discussions (Martin)

The baseline survey also investigated participants’ use of accessibility and assistive tools in their everyday digital life. Respondents were allowed to select multiple tools from a predefined list, including screen readers, keyboard-only navigation, alternative pointing devices, voice input, one-handed use, large fonts or magnified displays, high-contrast modes or colour filters, reduced motion, captions and simplified layouts, as well as an option indicating that they use none of these tools. Out of the 17 respondents, 8 participants (47.1 %) selected the option “none of these”, indicating that they do not regularly use any of the listed accessibility features. At the same time, a non-negligible subset reported using one or more assistive features. Five respondents (29.4%) indicated that they use keyboard navigation (e.g., tab/shift+tab) and another 5 (29.4%) reported relying on large fonts or a magnified screen. Around 3 participants each (17.6%) mentioned using screen readers, high-contrast modes or colour filters or simplified layouts / reading aids. Smaller proportions reported using alternative pointing devices, voice input, one-handed configurations or captions for audio/video. No respondent reported using reduced-motion settings in their everyday devices. The distribution is summarised in Table 3 and shows that, while many participants do not identify

themselves as users of assistive technologies, there is a substantial minority with concrete visual or motor needs, underlining the importance of robust accessibility features on the platform.

Table 3. Use of accessibility and assistive tools in everyday digital life (Martin)

Assistive tool / feature	n	% of participants using this tool
None of the listed tools	8	47.1
Keyboard-only navigation (tab/shift+tab)	5	29.4
Large fonts or magnified screen	5	29.4
Screen reader	3	17.6
High-contrast mode or colour filters	3	17.6
Simplified layouts / reading aids	3	17.6
Alternative pointing devices	2	11.8
Voice input	2	11.8
One-handed configurations	2	11.8
Captions for audio/video	2	11.8
Reduced motion settings	0	0.0

Figure 5 visualises the relative prevalence of each assistive tool, including the “none” category, thereby illustrating the heterogeneity of accessibility needs within this small sample.

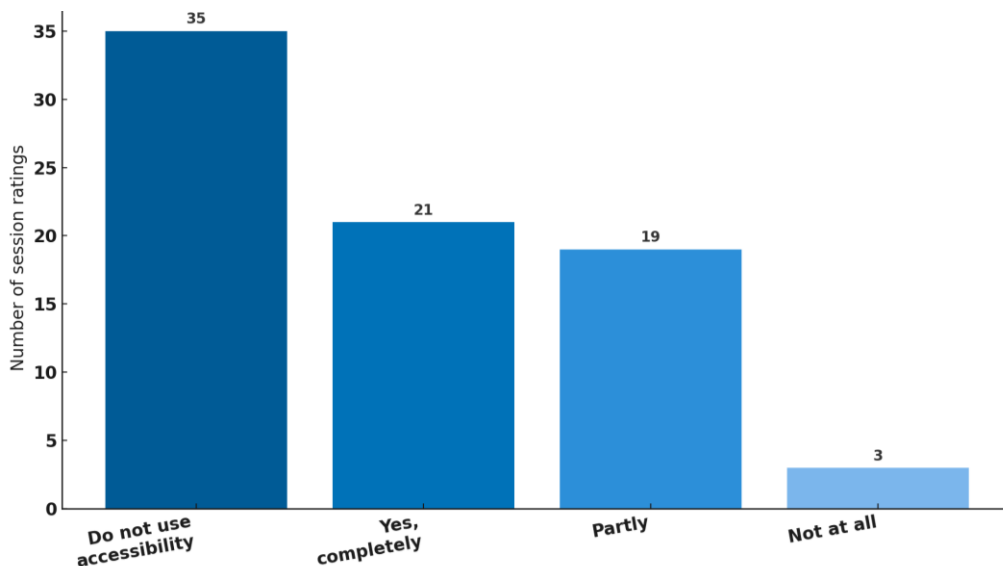


Figure 5. Perceived suitability of accessibility functions

Finally, participants were asked whether they had ever reported problematic content on online platforms (e.g. using “report” or “flag” functions). Almost half of the sample (n = 8, 47.1%) answered that they had never done so, while 4 respondents (23.5%) reported that they rarely do so, another 4 (23.5 %) indicated that they sometimes report content, and only 1 participant (5.9%) reported having done this often. If these categories are coded from 1 (“never”) to 4 (“often”), the resulting mean score is $M = 1.88$ ($SD = 0.99$), again emphasising that most participants are relatively passive in terms of content-moderation behaviours. Figure 6 presents this distribution.

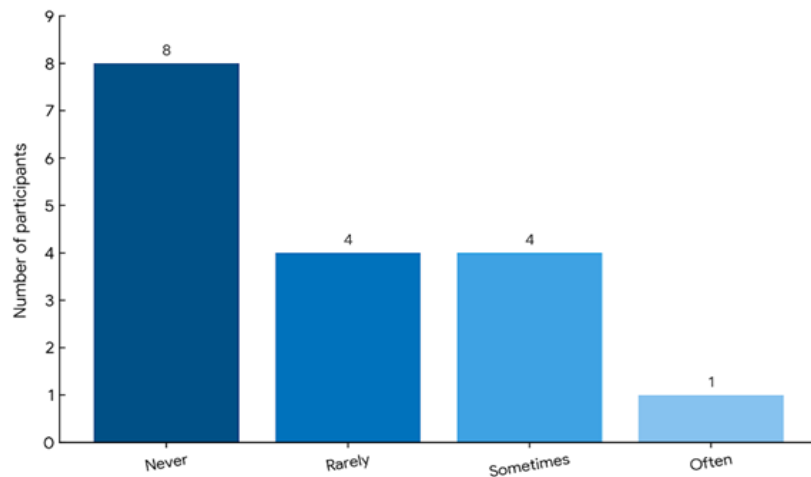


Figure 6. Participants' frequency of reporting problematic online content note (Martin)

Taken together, the Martin Phase 2 sample can be characterised as a small but diverse group of adult users who: (a) are not regular participants in online public debates, (b) approach AI-generated content with cautious, low-to-moderate trust and (c) display a heterogeneous profile in terms of accessibility tool use. These baseline characteristics provide a useful lens for interpreting subsequent results on usability, accessibility, engagement and trust in the ITHACA platform in the Martin context.

Braşov pilot

In total, 15 valid baseline responses were recorded. A first indicator concerns the frequency with which participants usually take part in public online discussions. As shown in Table 4 below, 7 participants (46.7%) reported that they often participate in public online discussions, 5 (33.3%) that they never do so, 2 (13.3%) that they sometimes participate and 1 (6.7%) that they rarely do. Coding these categories from 1 (“never”) to 4 (“often”) yields a mean score of approximately $M = 2.73$ ($SD = 1.34$) on a 1–4 scale, indicating a mixed but overall more active participation profile than in Martin, with a substantial subgroup of frequent contributors and a similarly large subgroup of non-participants.

Participants were also asked about their prior trust in AI-generated summaries and content rules on online platforms. Answers were given on a five-point scale (1–5). The distribution is strongly skewed towards higher trust. Ten respondents (66.7%) selected the highest value, 2 (13.3%) selected 4, 1 (6.7%) selected 3, and 2 (13.3%) selected 2; none chose the lowest category. This yields a mean trust score of $M = 4.3$ ($SD = 1.1$) on a 1–5 scale, indicating generally high prior trust in AI-generated content among the Braşov participants.

A third baseline question asked about the main goal for using the platform during the study session. The responses were relatively evenly distributed across the four core purposes. Four participants (26.7%) indicated that their main goal was to read what others think and another four (26.7%) that they aimed primarily to report problematic content. Three participants (20%) stated that they wanted to understand the key ideas quickly via the summary, two (13.3%) that they wished to express their own opinion and two (13.3%) selected “other” goals. Overall, the Braşov sample includes both readers and active contributors, with a sizeable subgroup explicitly oriented towards reporting problematic content and a non-negligible subgroup interested in making use of AI-generated summaries.

The baseline survey further examined everyday use of accessibility and assistive tools when navigating the internet. Respondents could select multiple options from a predefined list. As summarised in Table 4, 8 participants (53.3%) reported using none of the listed accessibility tools in their daily digital life. At the same time, 7 participants (46.7%) indicated that they use enlarged text or zoom/magnifier functions and 1 participant (6.7%) reported using keyboard-only navigation (e.g., TAB/Shift+TAB). No respondent reported using a screen reader, high-contrast modes, alternative pointing devices, voice input, one-handed configurations, reduced motion settings, captions or simplified layouts/reading aids. Thus, while just over half of the Braşov participants do not self-identify as assistive-technology users, almost half rely on basic visual adjustments (larger text or zoom), underscoring the relevance of font scaling and layout clarity in the platform design.

Finally, the baseline survey asked how often participants had reported problematic content on online platforms in the past, using a four-point scale from “never” to “often”. Fourteen valid responses were recorded. Five participants (35.7%) reported never having reported content, four (28.6%) indicated rarely, two (14.3%) sometimes and three (21.4%) often. Using a 1–4 coding (never–often) yields a mean of approximately $M = 2.21$ ($SD = 1.15$), indicating that, compared to Martin, a larger subset of the Braşov sample is familiar with using reporting tools, although non-reporting remains the single largest category.

Table 4. Participation frequency, AI trust and main goals (Braşov)

Indicator	Categories / scale	%	Summary measure (approx.)
Frequency of public online discussions	Never	33.3	1–4 scale (never–often): $M = 2.73$, $SD = 1.34$
	Rarely	6.7	
	Sometimes	13.3	
	Often	46.7	
Prior trust in AI-generated summaries/rules	5-point scale (1–5) (distribution: mostly high values)		Mean trust $M = 4.3$, $SD = 1.1$
Main goal for today’s use of the platform	Read others’ opinions	26.7	Mixed profile: readers, reporters, summary users
	Report problematic content	26.7	
	Understand key ideas via summary	20.0	
	Express own opinion	13.3	
	Other	13.3	
Everyday use of accessibility / assistive tools (multiple responses allowed)	None of the listed tools	53.3	
	Enlarged text / zoom / magnifier	46.7	
	Keyboard-only navigation	6.7	
	Other tools (screen reader, contrast, etc.)	0.0	

Indicator	Categories / scale	%	Summary measure (approx.)
Past experience with reporting content (valid N = 14)	Never	35.7	1–4 scale (never–often): $M = 2.21$, $SD = 1.15$
	Rarely	28.6	
	Sometimes	14.3	
	Often	21.4	

UPAT gamification evaluation – baseline profile

The gamification module of the ITHACA platform was evaluated in a dedicated laboratory-style study conducted by the University of Patras. Baseline information for this strand was collected through a pre-questionnaire completed by nine participants ($N = 9$) before they interacted with the gamified missions and feedback mechanisms. Although the instrument did not collect socio-demographic variables such as age or gender, it provides a clear picture of participants' motivational profile and prior experience with gamified systems, which forms the background for interpreting their reactions to the gamification layer.

All nine participants completed the core block of pre-questionnaire items. These items used a seven-point Likert scale (1 = strongly disagree, 7 = strongly agree) to assess expectancies about the upcoming session and preferences for game-like features.

Baseline expectations

Baseline expectations were uniformly high across all core expectancy dimensions.

- **Anticipated interest in the session** (engagement expectancy) showed a mean of $M = 6.1$ ($SD = 0.9$) on the 1–7 scale, with all responses in the upper range (5–7).
- **Task self-efficacy** (confidence in being able to complete the missions) displayed an identical pattern, $M = 6.1$ ($SD = 0.9$), again with no low ratings.
- **Expected enjoyment** of the gamified session was also clearly positive, $M = 5.6$ ($SD = 1.2$), with individual scores spanning from moderately positive to very positive.

These results indicate that participants approached the evaluation with strong expectations that the platform would be interesting, manageable and enjoyable, creating a high-expectation baseline against which post-session outcomes are interpreted.

Preferences for gamification features

Preferences related to specific game-like features followed the same broadly positive trend.

- **Comprehension and acceptance of points-and-badges systems** (structural clarity of rewards) was high, $M = 5.7$ ($SD = 1.2$), suggesting that participants were generally comfortable with this type of reward structure.
- **Motivation by others' visible progress** (social comparison / competitive orientation) showed a positive, though slightly more moderate, mean of $M = 5.0$ ($SD = 1.1$).
- **Preference for social recognition** (feeling recognised by a community) was somewhat lower but still above the neutral midpoint, $M = 4.9$ ($SD = 0.8$), indicating that recognition matters but is not the primary driver for all participants.

Autonomy-related preferences were also clearly endorsed:

- **Preference for autonomy in choosing activities** (liking to decide what to do first) yielded a mean of $M = 5.7$ ($SD = 1.0$).
- **Preference for multiple paths to progress** (flexible progression rather than a single linear path) showed a mean of $M = 5.6$ ($SD = 1.0$).

Taken together, this profile describes a group that is receptive to points/badges and visible progress, values autonomy and flexibility in how they progress and is moderately motivated by social comparison and community recognition. In other words, the sample's motivational orientation is highly compatible with the core design choices of the gamification layer (mission-based tasks, points/badges, leaderboards and visible progress).

Prior experience with gamified systems and expectations about gamification

The pre-questionnaire also captured prior exposure to gamified applications. When asked about their past use of applications with points or rewards, most participants reported at least some experience:

- 1 participant (11.1%) indicated no prior use,
- 5 participants (55.6%) selected “a little”, and
- 3 participants (33.3%) selected “a lot”.

Thus, almost nine out of ten participants had some prior experience with gamified systems and one third considered themselves highly familiar with such mechanisms, for example through learning platforms, fitness applications or loyalty schemes.

Crucially for the aims of this evaluation, all nine participants endorsed the statement that gamification would make their participation more interesting, yielding a unanimous 100 % “yes”. This indicates that the UPAT sample begins the study with uniformly positive expectations about the potential of gamification to enhance engagement in the platform. The combination of high baseline interest, strong self-efficacy, and favourable attitudes towards core gamification mechanics provides an important interpretive context for the subsequent post-session results and qualitative feedback on missions, rewards and leaderboards.

Table 4 is displaying the mean baseline scores for the main expectation and preference items (expected interest, expected enjoyment, self-efficacy, preference for autonomy and multiple paths, and receptiveness to points/badges and visible progress). A separate bar chart shows the distribution of prior gamified application use (“no”, “a little”, “a lot”), highlighting that almost all participants have previous experience with such systems.

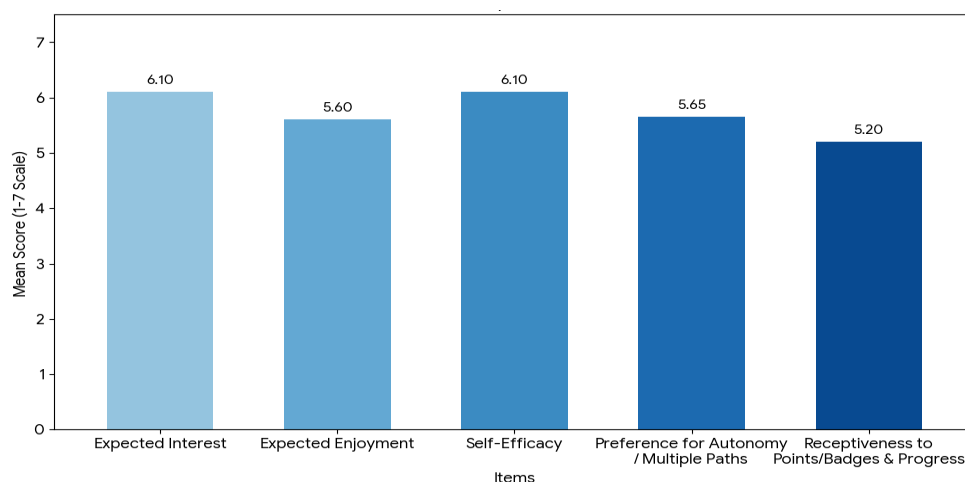


Figure 7. Mean baseline scores for gamification expectations and preferences

Taken together, the UPAT sample can be described as a small, digitally experienced group with clear prior exposure to gamified systems, strong baseline interest in the evaluation task and uniformly positive expectations about the motivational effect of gamification. This profile contrasts with the Martin sample, which is less accustomed to participating in online debates and more cautious about AI-mediated content and it will be relevant when comparing later results on engagement, perceived fairness and usability of the gamification layer across pilot sites.

4.2.2 Use and experience of the platform by citizens

Martin

Questionnaire results

For the Martin pilot, the short post-visit mini-surveys (A2/A4) were used to capture how each session felt to participants in terms of ease of use, satisfaction, perceived blockers and willingness to return. Across the different exports, there are 80 completed session questionnaires, corresponding to repeated visits from the same set of participants described in the baseline section.

Each mini-survey included a small core of closed questions rated on a five-point scale (1 = very low/negative, 5 = very high/positive), capturing: perceived ease of accomplishing intended actions on the platform, overall satisfaction with the session, willingness to use the platform again (yes/no), and perceived adequacy of the accessibility functions (response options: “yes, completely”, “partly”, “not at all”, and “I do not use accessibility features”). In addition, an open-ended question invited participants to describe what had slowed them down the most during the session, allowing them to report concrete obstacles in their own words.

Perceived ease of use. Ratings of how easy it was to do what they wanted were available for 79 out of 80 sessions. The average score was 3.8 out of 5. Around 60% of responses (59.5%) fell in the positive range (scores 4–5), while 12.7% were clearly negative (scores 1–2) and 27.8% neutral (score 3). This pattern suggests that, in most sessions, participants could complete their tasks without major difficulty, but a non-trivial minority still experienced the interaction as effortful or confusing.

Session satisfaction. Self-reported satisfaction with the session was slightly lower but followed a similar distribution. For the 70 sessions where this item was answered, the mean score was 3.6 out of 5. A little over half of the ratings (55.7%) were in the positive range (4–5), while 24.3% were negative (1–2) and 20% neutral. This indicates that the majority of sessions were experienced as broadly satisfactory, yet about one quarter of the interactions left participants dissatisfied, which aligns with some of the qualitative comments on navigation issues and non-working functions.

Willingness to reuse the platform. Willingness to return was captured as a yes/no question. Among the 79 valid answers, 56 sessions (70.9%) were associated with a “yes” and 23 (29.1%) with a “no”. Even though individual participants contribute multiple sessions, this still points to a generally favourable intention to reuse the platform once the initial learning curve has been overcome, with roughly seven out of ten session-level evaluations ending with a positive intention to come back.

Accessibility functions in everyday use. The same mini-surveys also asked whether the accessibility features met participants’ needs. After normalising the response options across the different survey versions, 78 valid answers were available. Of these, 35 responses (44.9%) indicated that participants do not use accessibility features, 21 responses (26.9%) reported that the accessibility options were fully suitable, 19 responses (24.4%) that they were partly suitable, and 3 responses (3.8%) that they were not suitable at all. Indirect evidence from the dedicated accessibility

micro-survey suggests that these cases are more likely to reflect isolated interface-level issues (e.g., unclear labels, occasional ‘getting stuck’, or local contrast problems) rather than a systematic incompatibility with a particular assistive technology. However, given the very small number of negative ratings and the anonymised nature of the data, these interpretations should be treated as indicative rather than conclusive.

Taken together, these data show that nearly half of the session evaluations came from users who do not rely on accessibility tools in their everyday digital life and therefore did not evaluate these functions in depth. Among those who did use them, roughly three out of four sessions reported that the accessibility options were either fully or partly adequate, while a small minority (three sessions) expressed that they did not meet their needs at all. This picture is consistent with the baseline profile, where a substantial minority reported using larger fonts, keyboard navigation and other aids, but most participants did not identify as regular assistive-technology users. Figure 8 summarises the key quantitative indicators from the Martin mini-surveys. The survey results indicate a generally positive user experience, with the majority of participants reporting high levels of satisfaction and ease of use. This favourable reception is reinforced by a strong intent to return, as most users expressed a willingness to utilize the platform again. Regarding accessibility, while a significant portion of the participants do not require these features, those who utilised them largely found them suitable for their needs, with very few reporting them as inadequate.

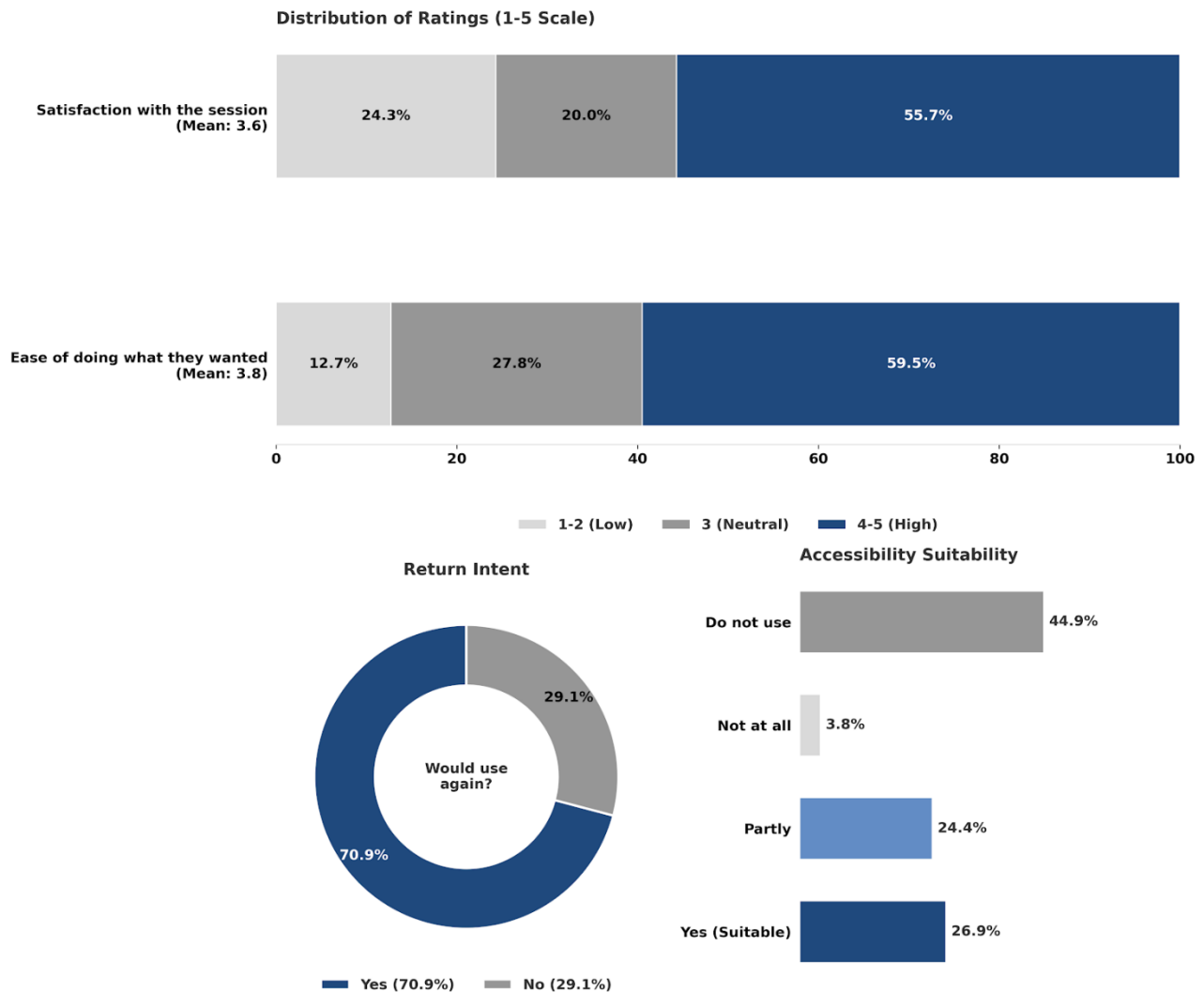


Figure 8. Summary of mini-survey results (Martin)

What slowed participants down. The open question on what slowed them down received interpretable answers in 67 session questionnaires. Many of these responses were short phrases; others were more descriptive. A substantial portion of participants wrote that nothing in particular impeded them (39%).

Among those who did report obstacles, several themes emerged:

- A first group of comments referred to navigation and orientation problems, mentioning the time needed to “find the right steps”, “figure out where to click next” or “understand the structure of the page”. Around 12% of the responses contained such references.
- A second cluster, representing roughly 15% of the comments, pointed to AI-related tools and summaries. Participants mentioned, for example, non-working or unclear summarisation functions, or difficulties with the toxicity-testing task.
- A smaller subset (around 7%) described technical issues, such as slow page loading, being logged out unexpectedly or problems accessing the platform from a mobile device.
- The remaining comments (about one third) covered a variety of more specific points, including suggestions for improving how summaries display percentages.

Figure 9 represents the proportion of comments falling in each broad theme (“nothing”, “navigation/structure”, “AI tools and summaries”, “technical issues”, “other”).

Taken together, these session-level results for Martin suggest that the platform is broadly usable in real-life conditions, with most sessions rated as easy enough and satisfactory, and with a clear majority of users indicating that they would be willing to use the platform again. At the same time, the presence of a non-negligible minority of negative ratings and of recurrent comments about navigation, AI-tool behaviour and isolated technical issues indicates concrete areas for improvement, which will be picked up later in the recommendations’ sections.

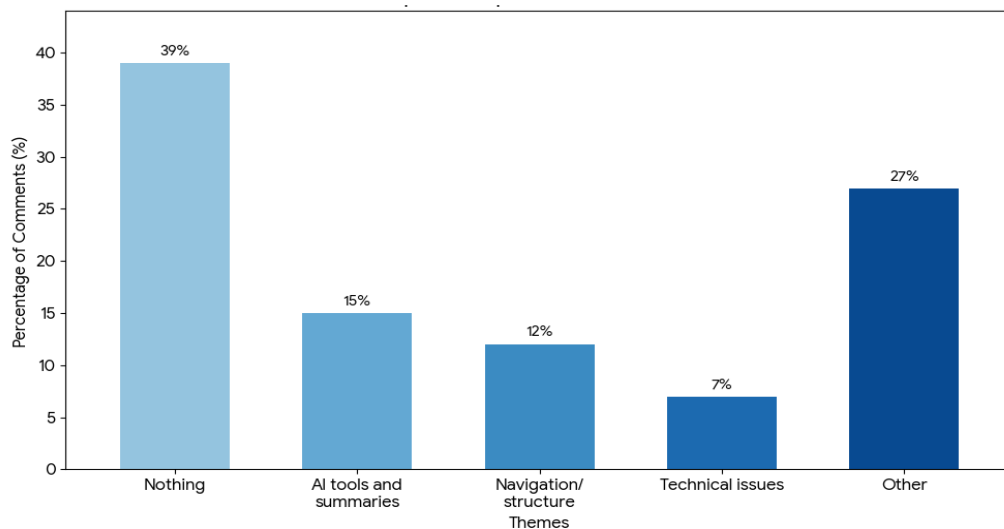


Figure 9. Themes in participants' responses on what slowed them down

Accessibility-specific mini-surveys

In addition to the general A2/A4 mini-surveys, a dedicated accessibility micro-survey was administered to participants who use assistive technologies or who wished to provide more detailed feedback on accessibility. Across the different exports, this instrument yielded 31 completed questionnaires in Martin.

The central question asked participants whether they were able to complete the assigned task with their own settings, explicitly referring to screen readers, keyboard-only input, alternative pointing devices and similar configurations. Responses were coded as yes/no. Out of the 31 answers, 29 (93.5%) reported that they were able to complete the task with their settings, one respondent indicated that they were not able to complete it under these conditions, and one response was missing. This indicates that, in almost all accessibility-focused sessions, participants could complete the required actions without having to abandon or substantially alter their usual assistive setup.

The same questionnaire included a list of potential accessibility problems (e.g. low contrast, focus indicator jumping or getting stuck, unclear labelling of buttons or links, missing output from screen readers, difficulties due to animations, getting stuck in the interface) and allowed respondents to tick all that applied, as well as an “Other” free-text field. Overall, the number of reported specific accessibility errors was low. Difficulties related to visual contrast were explicitly mentioned in only one case, while problems with unclear labelling of buttons or links were reported twice. Similarly, there were two instances where participants indicated that they got stuck in the interface, suggesting isolated moments where they felt unable to progress without assistance. No respondents reported issues with keyboard focus behaviour (such as the focus indicator skipping or getting stuck) or with animations making navigation difficult and there were no direct reports of screen readers failing to read content in the predefined checkbox options.

The open “Other” field, which elicited a richer set of comments, helps contextualise these findings. Many participants explicitly stated that they did not encounter any accessibility-related problems, often using short formulations to indicate the absence of shortcomings. A smaller subset of comments pointed instead to technical or session-level issues rather than structural accessibility barriers. For example, being logged out after a period of inactivity and then unable to post a comment or the summarisation tool not functioning correctly at specific moments.

From an evaluation perspective, these results are positive. Among those who use assistive features or who paid particular attention to accessibility, the vast majority could complete tasks successfully, and only isolated, specific issues were reported in the predefined categories. At the same time, the small number of negative or “stuck” experiences and the comments about unclear labels or inconsistent behaviour when returning to previous pages, indicate areas where further refinement of focus management, labelling and error handling would still be beneficial, particularly if the platform is to be used at scale by a more diverse population.

Free exploration experience

In the Phase 2 protocol, the same set of mini-surveys was used for both guided missions and free-exploration sessions, so the quantitative results (ease of use, satisfaction, willingness to reuse, accessibility suitability) are aggregated across both modes. However, we could have clear picture of what happened when participants were allowed to explore the platform more freely, without step-by-step missions.

In these free-exploration sessions, most users described their behaviour as closer to their everyday online habits; they mainly browsed existing topics, read through threads, and reacted with likes or simple comments, rather than systematically following missions. The platform was generally experienced as “manageable” for reading and catching up, especially once participants had already been through one or two guided sessions and had learned where to find the discussion area. Several participants reported that, during free exploration, they increasingly relied on AI summaries as a quick entry point, i.e., using the summary first to decide whether to read the full thread or to skip to another topic.

At the same time, the free-exploration phase made some limitations more visible. Participants noted that, when they roamed the platform on their own, it was still easier to read than to contribute; adding a new comment or proposal, changing settings, or using the “Report” function remained less obvious than simply scrolling and reading. Some users mentioned that they occasionally “lost their place” when going back from a thread to the topic overview or when scrolling through long discussions, which suggests that orientation and navigation cues are particularly important once missions are no longer guiding the next step. Despite these friction points, the combination of reading familiarity and the helpfulness of summaries meant that, in free exploration as well, most sessions were evaluated as acceptable or good, with a clear majority of users indicating that they would be willing to return to the platform in future.

Focus group

In addition to the mini-surveys, a dedicated focus group with end-users in Martin was organised to gain a richer understanding of how citizens experienced the platform over multiple visits and how they perceived the AI-based summaries, content rules and accessibility features. The session followed the Phase 2 focus-group guide for participants, with a 60–75-minute online discussion structured around five blocks: (a) ease and friction in core journeys, (b) AI summaries, (c) content rules and reporting, (d) accessibility and (e) a short prioritisation exercise to identify the top three fixes.

Participants were recruited from those who had completed several sessions during the field phase, ensuring that everyone in the group had hands-on experience with the platform in real conditions rather than only in a demonstration setting. The group included users who primarily read and reacted to posts, as well as those who tried posting and reporting content and at least one participant who relied on accessibility-related settings (such as larger text or zoom).

Ease and friction in core user journeys

The discussion around ease of use broadly confirmed the patterns observed in the mini-surveys. Most participants reported that they could log in and find the main discussion area without major difficulty after the first or second visit. The overall look and structure of the landing page were described as “clean” and “not intimidating”, and several participants appreciated that once they found a thread, reading through the comments felt familiar, similar to other online platforms.

However, participants also identified several recurrent friction points:

- Some users found it unclear how topics, threads, and comments relate to each other, especially the distinction between a topic overview and the detailed discussion view. Moving back and forth between these levels sometimes felt confusing.
- A number of participants mentioned that it was not always obvious where to click to perform a specific action (e.g., adding a comment versus reacting to an existing one). Labels and icons were occasionally perceived as too subtle or “hidden in the interface”.
- A few participants noted that when they tried to follow the “missions” or a specific set of steps, they lost track of where they were after returning to the previous page or scrolling and they had to “start again from the top”.

Overall, participants agreed that, after a short learning phase, basic reading and reacting were manageable, but that contribution steps (posting, reporting, changing settings) required more guidance and clearer visual cues.

Usefulness, coverage and trust of AI summaries

The focus group devoted a substantial portion of time to the AI-generated summaries, as these are central to the ITHACA concept. Participants were shown a long discussion thread together with its summary and asked to reflect on usefulness, coverage and perceived fairness. Most participants stated that, in principle, having a summary at the top of a long thread is very helpful; it allows them to “catch up quickly” and decide whether the topic is worth reading further. Several said that if they were returning to the same thread after a few days, they would use the summary first to see what had changed. At the same time, the discussion highlighted a number of limitations and concerns:

- Some participants felt that the summaries were too compressed, giving them the impression that “only two or three ideas” were present even when the original discussion was more diverse.
- Others remarked that minority or less popular views sometimes felt under-represented. They gave examples where more critical or nuanced opinions appeared only briefly in the summary, if at all, even though these posts had stood out to them when reading the full thread.
- A few participants found it difficult to connect specific summary sentences back to the original comments, which made it harder to verify context or check whether a particular phrasing accurately reflected what had been said.

Trust in the summaries was therefore conditional. Participants saw clear value in using them as a “first glance” or orientation tool but were hesitant to rely on them alone for forming an opinion about what citizens had actually said. Several emphasised that they would prefer a summary that is “slightly longer but clearly representative” over one that is very short but potentially omits important perspectives.

When asked what would make the summaries feel fairer and more trustworthy, participants suggested:

- ensuring that at least one sentence clearly reflects dissenting or minority views,
- adding a simple indicator of how many posts the summary is based on and
- allowing users to expand or highlight parts of the discussion that correspond to each summary point.

Content rules and reporting

Participants were also asked about their understanding of content rules and their experience with the “Report” function. A consistent finding was that most users rarely reported content during their regular visits. They simply did not encounter many clearly problematic posts and some were unsure when a post was “serious enough” to justify a report.

Key points from the discussion included:

- The reporting button was not always noticed, especially on mobile or when participants were focused on reading. A few users stated that they became aware of it only after the guided missions drew attention to it.
- Participants expressed a preference for plain-language explanations of what counts as harassment, hate speech or other violations, ideally in the local language, so that they can feel confident they are applying the rules correctly.

- There was some concern that borderline posts (e.g., sarcasm, strong criticism) might be treated inconsistently, either by AI or by human moderators, especially across different topics.

While no major incidents of unfair treatment were reported, participants agreed that having a short, visible description of the rules near the reporting option, plus a clearer indication of what happens after a report is submitted (e.g., “Your report has been sent to moderators”), would encourage appropriate use of the function and increase trust in the moderation process.

Accessibility and diverse needs

The focus group also explored accessibility in a more conversational way than the short A3 survey. A few participants explicitly mentioned that they regularly use larger text, zoom or high-contrast settings on their devices, while others stated that they do not usually rely on assistive tools. The main accessibility-related themes were:

- Several participants found that the default font size and spacing on the platform are “acceptable but could be slightly larger”, especially for longer texts or for use on smaller screens.
- The contrast between text and background was generally judged to be sufficient, although one participant noted that some interface elements (icons, secondary text) could be more pronounced.
- No one reported being blocked entirely by keyboard or screen reader issues during their sessions, but there were comments about needing to scroll a lot and sometimes losing their place in the thread.

Overall, participants considered the platform usable from an accessibility point of view but suggested that a small set of visual refinements, particularly around text size, spacing and the visibility of interactive elements, would make the experience more comfortable for a broader range of users.

Top-3 fixes from the user focus group

At the end of the session, participants were invited to propose and then collectively prioritise the Top-3 fixes they considered most important for improving their experience with the platform. While the exact phrasing varied, the priorities converged around three main themes: navigation and clarity of actions, summary quality and transparency and support for reporting and fairness (Table 5).

Table 5. Top-3 user-prioritised fixes (Martin)

Priority area	Description from participants	Underlying need
1. Navigation and clarity of actions	Make it clearer where to click to comment, react, report; distinguish topics vs threads more visibly; avoid “getting lost” when going back.	Reduce cognitive load; support confident contribution, not just reading.
2. AI summary quality and transparency	Ensure summaries capture different viewpoints (including minority views); allow users to understand what the summary is based on and to trace back to original posts.	Strengthen trust in AI assistance; avoid misrepresentation or oversimplification of debates.
3. Reporting and content rules	Make reporting easier to discover; explain in simple terms what should be reported and what happens after a report is submitted.	Encourage appropriate use of reporting; reinforce perceived fairness and safety.

Participants stressed that these fixes are complementary rather than independent. Clearer navigation helps new users discover the summaries and the reporting function, better summaries make it easier to decide what to read and whether something is problematic and transparent rules and feedback loops support a sense of fairness and safety that makes it worthwhile to return and contribute. Below is the list of prioritised recommendations for improvements shared with development teams (Table 6).

Table 6. Priorities for improvements (Martin)

Functionality / Element	Issue	Priority	Suggestion for Resolution
AI summarization (article & thread)	Summaries often unclear/slow; users expect them directly under the article and to reflect other viewpoints.	High	Always show two short cards at the top of each topic: Article summary and Discussion so far, each with a brief "other views" line.
Toxicity & moderation	Vulgarities slipped through; a highly toxic proposal appeared; users want guidance to rephrase.	High	Add a clear "Please rephrase" step for harsh language and auto-send extreme cases to a human review list.
Session handling / logout	Silent logout caused lost work.	High	Warn before logout and keep what the user typed so they can continue when they return.
Mobile access to forum	Community forum hard to reach on mobile.	High	Add an always-visible Forum entry on mobile and make the main actions obvious and easy to tap.
Comment model & structure	Confusion between Comment/Opinion/Proposal; no replies; graph mixes everything.	High	Let people choose the post type, allow one-level replies, and show charts that focus on proposals by default.
Language & translation	Language setting didn't stick; content not translated; some odd characters.	High	Remember the chosen language, offer Translate content on each topic, and fix character display issues.
Voting & polls	Regular users couldn't find polls; no feedback after voting.	Medium	Make polls visible to everyone who can vote and show a simple "Vote recorded / You already voted" message.
Topic info / FAQ	FAQ wasn't helpful for decisions.	Medium	Replace FAQ with a Topic info panel (who's responsible, basic rules/law, short background, contact).

Functionality / Element	Issue	Priority	Suggestion for Resolution
Privacy / exports	Unclear what is masked; still need gender/age/district views.	Medium	Default to hiding names and sensitive details in exports while keeping gender/age/district for analysis, with a preview before download.
Accessibility basics	Older users struggled with small text/unclear buttons.	Medium	Use larger base text, clearer labels on icons/buttons, and strong contrast; keep a visible focus highlight.
Notifications / “new since last visit”	Hard to catch up.	Low	Show how many new posts appeared since last visit and a “show new only” filter.
Functionality / Element	Issue	Priority	Suggestion for Resolution
Location selector	“Martin” selection failed to appear or wasn’t retained after reload.	High	Make the chosen city visibly selected and persist it across page reloads so users don’t have to reselect it.
Page refresh & sharing	Users couldn’t refresh or share a topic page easily.	High	Enable standard page refresh and provide a shareable link/button on each topic and proposal.
AI tools performance	Text simplification/summarization often returned nonsense; short texts expanded; long texts missed the main idea.	High	Present fast, two-line previews first and allow “See full summary” on demand; keep summaries shorter than the source and focused on the main point.
Translation output quality	Odd characters and misleading word choices appeared in some translations.	High	Clean up encoding issues and use consistent wording (e.g., correct plural/singular forms) so translations read naturally.
Screen “bounce” / jitter	The screen visibly “bounced” during interaction.	High	Stabilize scrolling and layout so buttons and fields don’t shift while users are tapping or typing.
Anonymous mode policy	Anonymity encouraged rudeness; staff prefer identifiable input.	High	Disable anonymous posting for civic topics, while still allowing respectful pseudonyms where justified by the city.

Functionality / Element	Issue	Priority	Suggestion for Resolution
Spam & adverts	Promotional text slipped through as proposals.	High	Block posts with obvious adverts/links and show a short notice explaining that promotional content isn't allowed.
Poll integrity	Concern about multiple votes and unclear eligibility.	Medium	Show who is eligible to vote, confirm "vote recorded," and prevent duplicate votes for the same poll.
Proposal ownership & competence	Users lacked clarity on who is responsible for acting on a proposal.	Medium	Display the responsible city unit on each proposal and provide a short "why this unit" note.
Topic lifetime & archive	Discussions felt open-ended; hard to manage older threads.	Medium	Add an "open until" window for each topic and move older threads to an archive with read-only access.
Relevance & proposal filter	Users wanted to see only actionable proposals and hide noise.	Medium	Provide a "show proposals only" filter and a gentle prompt: "Your comment doesn't include a proposal—add one?"
Rich input options	Users asked to comment beyond typing.	Low	Allow short voice/video comments alongside text, with the same rules and reporting options.
Disputed cases queue	Content like "drivers vs cyclists" needs human judgment, not auto action.	Low	Add a "Needs review" flag for contentious posts so moderators can decide with full context.

Brasov

Questionnaire results

For the Braşov pilot, the same short post-visit mini-surveys (A2/A4) as in Martin were used to capture how each session felt to participants in terms of ease of use, satisfaction, perceived blockers and willingness to return. Across the six available exports (Visits 1–6), there are 84 completed session questionnaires, corresponding to repeated visits from the same set of participants described in the Braşov baseline section. As in Martin, these results mix guided "mission" sessions with more free exploration, but the core indicators are identical.

Each mini-survey included a small set of closed items rated on a five-point scale (1 = very low/negative, 5 = very high/positive), assessing: perceived ease of accomplishing the actions they intended to perform on the platform, overall satisfaction with the session, willingness to use the platform again (yes/no), and perceived adequacy of the accessibility functions (response options: “yes, completely”, “partly”, “no”, and “I do not use accessibility tools”). An open question asked what had slowed them down the most during the session, allowing users to report concrete obstacles in their own words. As in Martin, this combination provides both session-level ratings and a qualitative view of friction points.

Perceived ease of use

Ratings of how easy it was to do what they wanted are highly skewed towards the positive end of the scale. For Visit 2, almost all respondents chose the maximum rating: 14 out of 15 sessions (93.3%) were rated 5 and one 4, yielding a mean of 4.9/5. In Visit 3, all 13 valid answers to this item were in the top range, with a reported mean of 5.0/5. The same pattern is seen in Visits 4 and 5: all ease-of-use ratings are scores of 5.0. In Visit 6, the mean remains very high at 4.8/5, with nine ratings at 5 and one at 2; no one selected 1, 3 or 4 on this item.

Aggregated across all six visits, more than 95% of answered ease-of-use ratings fall in the positive range (4–5), with the remaining responses reflecting mainly a small number of mid-scale or isolated low scores. In practical terms, most sessions in Braşov were experienced as straightforward. Participants generally felt that they could complete their intended actions without major difficulty, with only a handful of sessions being perceived as effortful.

Session satisfaction

Self-reported satisfaction with each visit closely mirrors the ease-of-use results described above and is, if anything, even slightly higher. In Visit 2, satisfaction scores again cluster at the top of the scale: 14 out of 15 responses (93.3%) are at 5 and one at 4, for a mean of 4.9/5. Visit 3 shows a very similar pattern (mean 4.9/5; 93.3% ratings at 5 and one rating at 4), while in Visit 4, 12 out of 13 ratings (92.3%) are at 5, with a single rating at 3. Visit 5 is uniformly positive, with all 14 satisfaction ratings at 5/5, and in Visit 6 all 15 participants again report maximum satisfaction (5/5).

Across all sessions where this item was answered, the average satisfaction score is therefore very close to the ceiling (= 4.9–5.0 out of 5). Negative ratings (1–2) are entirely absent, and neutral responses (3) are rare. This indicates that the vast majority of sessions in Braşov were not just manageable but genuinely satisfactory from the users’ perspective.

Willingness to reuse the platform

Willingness to return was captured as a yes/no question. Here, the Braşov dataset is even more unequivocal than Martin’s. In every survey wave from Visit 2 onwards, all participants who answered this question indicated that they would use the platform again (e.g., 15/15 “yes” in Visits 2, 3, 4 and 6; 14/14 “yes” in Visit 5). When combined across visits, this yields a unanimous 100% positive intention to reuse at the session level.

Although repeated observations from the same individuals are included, this pattern still signals a very favourable reception. Once citizens in Braşov had engaged with the platform, every recorded session ended with an expressed willingness to come back.

Accessibility functions in everyday use

The same mini-surveys also asked whether the platform’s accessibility functions matched participants’ needs. For Visits 2–6, 73 valid answers were recorded to this item. Across these, 42 responses (57.5%) indicate that participants do not use accessibility tools in their everyday digital life, while 31 responses (42.5%) report that the accessibility options were fully suitable; no participant in these visits selected “partly” or “no” as a response option.

Taken together, these data show that a substantial portion of the Braşov sessions came from users who do not rely on assistive features and therefore did not evaluate them in depth. Among those who did rely on such features, all session-level evaluations in Visits 2–6 report that the platform met their needs completely, with no indication of partial or complete mismatch in these items. This picture is consistent with the baseline profile, which showed that most Braşov participants do not consider themselves regular assistive-technology users, even though a notable subset uses zoom or enlarged text when navigating online.

A graphical summary of these four indicators for Braşov (distributions of ease-of-use and satisfaction ratings, proportion of “yes” answers for willingness to reuse, and the breakdown of accessibility suitability) is presented in an analogous figure to the one used for Martin (Figure 10).

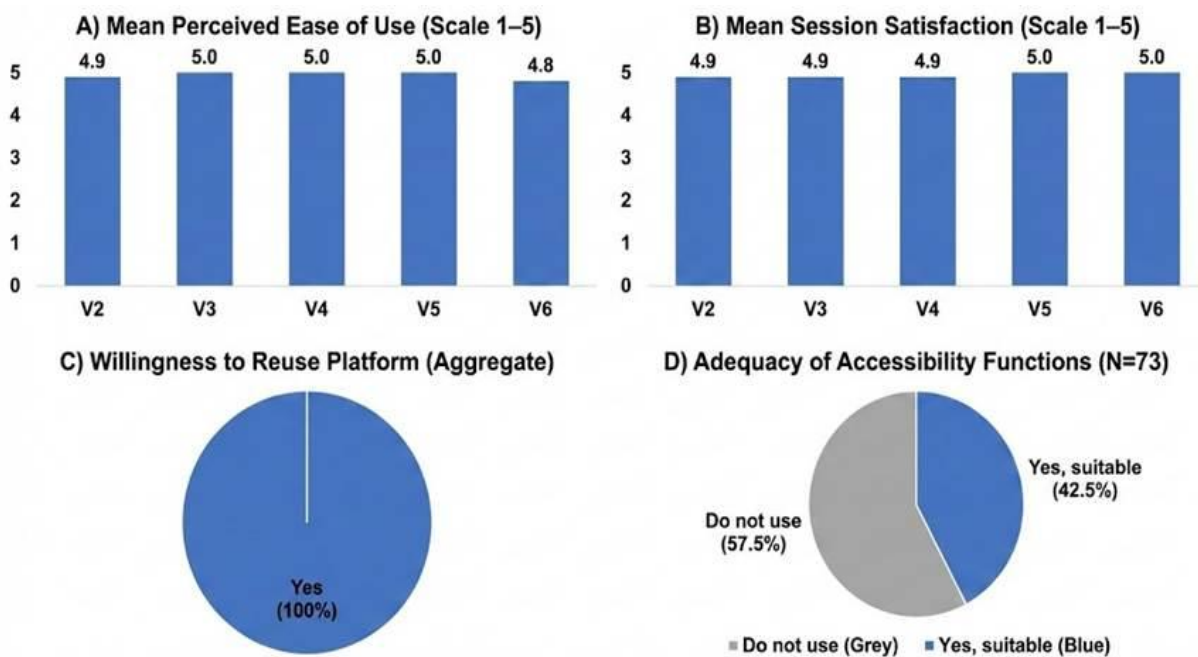


Figure 10. Graphical summary of Brasov user experience indicators

What slowed participants down

The open question “What slowed you down?” was answered in 31 sessions across Visits 3–6. As in Martin, responses ranged from very short phrases (e.g. “nimic” / “nothing”) to slightly more descriptive comments. Thematic coding of these answers reveals three main patterns.

First, a large share of comments explicitly state that nothing in particular slowed participants down. Many users wrote simply “nimic”, “totul a mers bine” (“everything went well”) or “a mers ok” (“it went ok”), often emphasising that by later visits they knew where everything was and felt accustomed to

the interface. Counting these formulations together, roughly two-thirds of all comments fall into this “no slowdown” category across visits 3–6.

Second, a smaller but recurring group of comments refers to navigation and orientation. Some participants mentioned that they “didn’t know where [they] had to go”, that they “jumped between pages” or that they had to look again at the tutorials to remember the correct steps. A few explicitly noted that they called the facilitator when they felt stuck or could not find the right view. These issues appear in most visits but become less frequent over time, suggesting that both guided missions and accumulated experience helped users build a more stable mental model of the platform.

Third, isolated mentions point to visual and presentation-related aspects. In Visit 3, for example, one respondent commented that “Se vede mic” (“it looks small”), indicating that the perceived small size of interface elements slowed them down.

Unlike in Martin, there are very few explicit references to technical faults (e.g., page not loading, being logged out) or to AI tools themselves as sources of friction. Instead, the main tension in Braşov is between a majority of sessions where “nothing” slowed users and a smaller subset where orientation and the need for guidance (through tutorials or facilitator support) played a role. Figure 11 summarises these patterns in terms of the proportion of comments referring to “nothing”, “navigation/structure & need for guidance”, “visual aspects” and “other”.

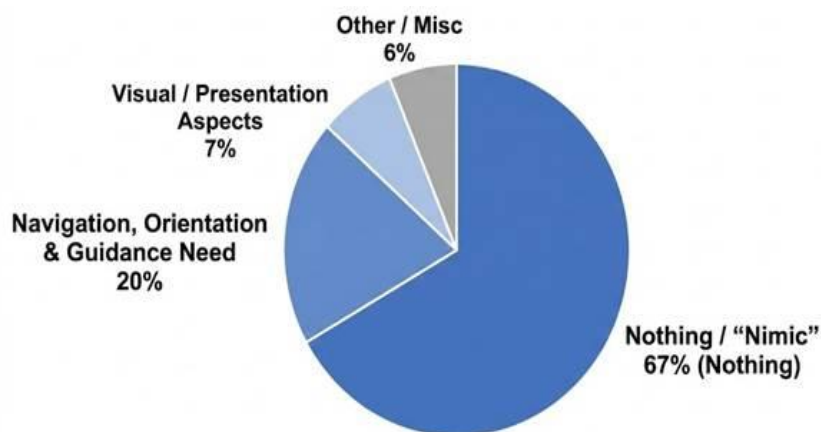


Figure 11. Thematic breakdown of what slowed users down (Braşov)

Overall, the session-level survey results for Braşov depict a very positive and relatively homogeneous user experience. The platform is perceived as very easy and satisfying to use, with unanimous willingness to return and no reports of accessibility functions being unsuitable. At the same time, the qualitative comments highlight a small but meaningful set of issues around initial orientation, the discoverability of actions and reliance on tutorials or facilitator support in early visits. These points align with the broader recommendations on navigation clarity and guidance and will be revisited in the cross-site comparison and suggestions for improvement.

Accessibility-specific mini-surveys (A3, Braşov)

In addition to the general A2/A4 mini-surveys, a dedicated accessibility micro-survey (A3) was administered in Braşov to participants who use assistive technologies or who wished to provide more detailed feedback on accessibility. Across Visits 1–6, this instrument yielded 90 completed

questionnaires (15 per visit). These focus specifically on whether participants could complete the session with their usual configuration and on what types of accessibility problems they encountered.

The central item asked whether participants managed to complete the session using their own setup (e.g. screen reader, keyboard-only navigation, alternative pointing devices). In Visit 1, 11 out of 15 respondents (73.3%) answered “yes”, with four indicating that they could not complete the session under these conditions. The same ratio is observed in Visit 2 (11 “yes”, 4 “no”), while in Visit 3 the success rate rises to 14 “yes” and one “no” (93% success). In Visits 4–6 all respondents report that they could complete the session with their configuration (100% “yes” in each wave).

Aggregated across all six visits, 80 out of 90 responses (88.9%) indicate successful completion with the participant’s own settings, and 10 (11.1%) report that they were unable to do so. Importantly, the proportion of “no” responses decreases steadily over time and disappears entirely from Visit 4 onwards, suggesting that initial accessibility bottlenecks were either resolved or mitigated as the study progressed.

The A3 questionnaire also listed specific categories of potential accessibility problems, such as low contrast, unclear button labelling, keyboard focus issues, missing screen-reader output, disruptive motion/animations or “being stuck” in the interface, and allowed respondents to tick all that applied, along with an “Other” option. Overall, the number of reported issues in these predefined categories is small:

- unclear button or link labels were mentioned once in Visit 1;
- being “blocked in the process” (unable to continue) was reported once in Visit 1 and once in Visit 2;
- one additional “blocked” case appears in Visit 5;
- low contrast and keyboard-focus problems are mentioned only in Visit 3, and not at all in later sessions;
- there are no checkbox reports of screen readers failing to read content or of motion/animations making the interface hard to use.

By contrast, the “Other” category absorbs most early comments: five of the seven problem reports in Visit 1 and two of the three reports in Visit 2 are recorded under “Other”, often accompanied by short notes such as “nimic” (“nothing”) or “totul a mers bine” (“everything went well”) that actually indicate the absence of a serious problem, suggesting some uncertainty about how to use the checklist. From Visit 4 onwards, no “Other” accessibility problems are reported, and comments increasingly stress familiarity with the platform (e.g. “m-am obișnuit” or “deja mă descurc” which means in English “I got used to it” respectively “I’m doing fine”).

From an evaluation perspective, these results are positive. Among those who paid particular attention to accessibility, almost all could complete the required actions with their usual configuration, especially in later visits. Very few issues fall into classic WCAG problem categories; instead, early difficulties seem to be situational (e.g., uncertainty about where to go next, isolated “stuck” states) and fade out as both users and the platform stabilise. Nevertheless, the small number of blocked experiences and the isolated mentions of unclear labels or contrast suggest that further refinement of focus management, labelling and error handling would still be beneficial, particularly if the platform is to support a broader range of assistive technologies and devices at scale.

Focus group

In addition to the mini-surveys, a dedicated focus group with end-users in Braşov was organised to gain a richer understanding of how citizens experienced the platform and how they perceived the AI-based summaries, content rules and accessibility features. The session followed the Phase 2 focus-group guide for participants, with an in-person discussion structured around five blocks: (a) ease and friction in core journeys, (b) AI summaries, (c) content rules and reporting, (d) accessibility and (e) a short prioritisation exercise to identify the top three fixes.

Participants were recruited from those who had completed sessions during the field phase, ensuring that everyone in the group had hands-on experience with the platform. The group (N=7) included users who actively contributed comments and proposals, as well as participants who relied on specific accessibility configurations such as screen readers, zoom, or high-contrast settings.

Ease and friction in core user journeys

The discussion around ease of use highlighted a mix of high engagement with specific operational friction points. Participants showed genuine interest in the platform and found the core concepts valuable, but identified technical and structural barriers that interrupted their flows.

Key friction points identified by the group included:

- **Mobile access and session stability:** A major issue was the difficulty of accessing the community forum comfortably from mobile phones. Additionally, users reported frequent automatic logouts ("silent logout"), which forced them to log in repeatedly and sometimes led to lost work or frustration.
- **Confusion between contribution types:** Participants found the distinction between "Comments", "Opinions", and "Proposals" unclear. The large amount of text and the lack of visual distinction made it difficult to identify which elements were truly important or actionable.
- **Navigation and page refresh:** Users noted that they often could not refresh the page or easily share a link to a specific conversation, which hindered natural browsing and sharing habits.
- **Lack of feedback:** Some participants felt uncertainty after performing actions (e.g., voting or posting) because the system did not always provide clear confirmation or feedback.

Overall, while the platform was seen as a solid base for development, these "operational frictions" were described as the primary obstacles to a fluid experience.

Usefulness, coverage and trust of AI summaries

The focus group discussed the AI-generated summaries and graphical insights extensively. Participants expressed high expectations for these tools but identified gaps between these expectations and the current performance.

- **Expectations of ubiquity:** Participants expected AI summaries and graphs to be available at all levels of the discussion, including sub-threads or deeper conversational branches, rather than just at the top level.
- **Precision and relevance:** Some users felt the summarization engine was imprecise, particularly when handling shorter texts versus longer debates. There were also comments

that the graphs displayed options or data points that were not always relevant to the specific discussion context.

- **Desire for structure:** Users suggested that using questions with pre-set variants (structured inputs) might help the system generate more precise graphs and summaries than relying solely on free text.

Users linked trust directly to precision, noting that the AI becomes useful only when it can effectively distinguish constructive proposals from general noise.

Content rules and reporting

Participants tested the toxicity and reporting features, specifically focusing on how the system handled offensive language.

- **Under-detection of toxicity:** A key finding was that the system failed to adequately identify vulgar expressions, particularly regional or specific terms. This led some users to lose interest in testing the toxicity analysis function further, as they perceived it as not yet "smart" enough for the local context.
- **Confusion regarding "Likes" on negative content:** There was ambiguity regarding the meaning of a "Like" on a negative or problematic post. Participants were unsure if this signalled agreement with the negativity or simply an acknowledgement of the issue.
- **Preventive AI:** Participants suggested implementing a preventive AI message (e.g., "Your comment does not contain a clear proposal") to guide users toward more constructive contributions before they post.

Accessibility and diverse needs

The session included participants using specific accessibility configurations (screen readers, zoom, contrast). The feedback was generally constructive but pointed to specific optimization needs:

- **User Profile and Interface:** The user profile page was described as needing work, as it did not meet standard expectations for layout and utility.
- **Device Experience:** The previously mentioned issues with mobile accessibility were highlighted as a significant barrier for inclusive participation, as many users prefer or rely on mobile devices.

Top-3 fixes from the user focus group

At the end of the session, the group's feedback was consolidated into three priority directions for improvement (Table 7). These align with the "Conclusion" of the Braşov report, emphasizing clarity, stability, and AI precision.

Table 7. Top-3 user-prioritised fixes (Braşov)

Priority area	Description from participants	Underlying need
1. Operational stability & Mobile access	Fix the automatic "silent" logouts and ensure the forum is fully accessible and usable on mobile devices. Enable standard refresh and share functions.	Basic reliability: Users need a fluid experience without technical interruptions to remain engaged.

Priority area	Description from participants	Underlying need
2. Clarity of contribution types	Visually distinguish between Comments, Opinions, and Proposals; reduce text density to highlight what is important.	Cognitive ease: Reduce confusion and help users understand <i>how</i> they are contributing.
3. AI precision & Toxicity detection	Refine the summarization for different text lengths and expand the toxicity vocabulary to catch local vulgarities.	Trust & Quality: AI must be accurate and culturally context-aware to be trusted.

Priorities for improvements (Braşov)

Below (Table 8) is the detailed list of prioritised recommendations derived from the Braşov findings.

Table 8. Priorities for improvements (Braşov)

Functionality / Element	Issue	Priority	Recommendation
Operational stability & mobile access	Users experience automatic “silent” logouts and must log in repeatedly, sometimes in the middle of a task. In addition, it is very difficult or impossible to access and use the community forum from mobile devices, and standard browser actions such as Refresh and Share do not behave as expected.	High	Stabilise the core session and navigation behaviour. Increase session timeout to a realistic duration and add a visible “session about to expire” message with a Stay logged in option. After re-login, return users to the same page and state (same thread, text preserved where possible). Fix responsive layout and touch targets so that full forum functionality (view topics, read threads, post comments/proposals, react, report) works on common smartphones. Ensure that standard browser Refresh reloads the current view correctly and provide a clear Copy link / Share control with stable, deep-linkable URLs for topics and threads.
Comment / Opinion / Proposal structure	Participants report confusion between Comments, Opinions and Proposals and describe the interface as text-heavy. Important and actionable content is visually buried and users are not always sure how their contribution will be used in decision-making.	High	Introduce a clearer information hierarchy and visual distinction between contribution types. Use consistent icons/colour and layout for Comments, Opinions and Proposals, and present Proposals as prominent cards with title, short description and status. Reduce text density with spacing and grouping. When a user selects what to post, provide short explanations for each type (e.g. “Comment = remark”, “Opinion = position”, “Proposal = concrete suggestion for action”) to support correct and consistent use.

Functionality / Element	Issue	Priority	Recommendation
AI summaries & graphs (discussion support)	AI-generated summaries in Romanian are often perceived as incoherent, overly compressed or not useful for official work. Graphs associated with discussions sometimes display options that are not clearly linked to the active thread, which creates doubt about what is actually being represented.	High	Improve the quality and transparency of AI assistance. Integrate a more robust translation and summarisation backend tuned for Romanian, with different parameters for short versus long texts to avoid over-compression and loss of key arguments. Restrict graphs to items that actually appear in the active topic or thread and label clearly what they show (e.g. “number of proposals per option”, “number of reactions”). Validate the usefulness of summaries and graphs with local moderators before wider deployment.
Language & translation quality (UI and AI output)	Romanian language quality in the interface and AI output is below an acceptable threshold: some texts are not translated, there are occasional encoding/character issues, and AI-generated Romanian content is often described as “mechanical” or inferior to external tools. Moderators state they cannot reuse these texts in their work.	High	Treat language quality as a core requirement. Ensure complete and correct translation of the entire UI, system messages and emails into Romanian. Fix any encoding and character rendering problems. For AI-related text (summaries, rephrasings, etc.), integrate a high-quality translation/summarisation service for Romanian and test outputs with native-speaking moderators. Only roll out features whose language quality matches or exceeds widely used reference tools for Romanian.
Toxicity AI (citizen-facing analysis)	The toxicity analysis misses many vulgar expressions and regional slurs. Participants quickly lose interest in this feature, perceiving it as “not intelligent enough”, and there is a risk that harmful content remains unflagged.	High	Extend and localise the toxicity detection pipeline. Build and maintain a vocabulary of Romanian and regional slang terms, starting from examples collected in the Braşov pilots. Tune thresholds using local test sets that include clearly toxic and borderline expressions. Validate detection behaviour together with local moderators and adjust until similar cases are handled consistently. Ensure that toxicity flags are visible and actionable in the moderation interface.

Functionality / Element	Issue	Priority	Recommendation
User feedback after actions	After actions such as voting, posting or reporting, users are not always sure whether the action has been registered because system feedback is weak, delayed or absent.	Medium	Provide clear and immediate micro-feedback for key actions. For each important interaction, show a short, explicit confirmation near the control (e.g. "Vote recorded", "You have already voted", "Post published", "Report sent") and update the visual state immediately (e.g. button disabled or changed, counter incremented). This reduces uncertainty and prevents repeated clicks or the perception that the system is unresponsive.
Voting logic and reaction semantics ("Like" on negative posts)	The meaning of a "Like" on negative or problematic posts is ambiguous. Users are unsure whether it indicates agreement with the negativity or simply acknowledgement, which can lead to misinterpretation of public sentiment and make toxic content appear widely supported.	Low	Clarify the semantics of reactions. Consider separating positive support from problem-flagging. For example, provide distinct reactions such as "Agree/Support", "Disagree" and a clearly labelled "Report / Problematic" option. Add tooltips or short labels explaining what each reaction means, and avoid relying on a single generic "Like" for content that is clearly negative or controversial.
Accessibility & blocked states	A small but important subset of users reported being "blocked in the process", unable to proceed or return to a safe point. Additional remarks include unclear button labels and low contrast on certain screens, which can affect users with visual or motor limitations.	Medium	Conduct an accessibility-focused review of key flows. Ensure every screen has a visible and accessible way to go back or restart (Back, Home, Retry). Verify that all interactive elements have clear, descriptive labels and adequate colour contrast and font size, especially in critical paths. Test keyboard-only and screen-reader navigation to identify and fix focus traps or dead ends. Add explicit error messages whenever a process cannot continue and provide a clear next step.
Anonymity & identity management	Allowing anonymous or weakly identified input is perceived by moderators as encouraging rude or abusive content and reducing accountability. They explicitly request stronger identity linkage and "real"	High	Revisit the identity and anonymity policy for civic use. For official participation processes, disable fully anonymous posting and require at least pseudonymous accounts linked to verified contact channels (e.g. email, municipal single sign-on). Make identity rules, data use and moderation policies transparent in the interface. Align moderation tools (flagging, escalation, sanctions) with this model so that abusive behaviour can be traced and managed when necessary.

Functionality / Element	Issue	Priority	Recommendation
	moderation, without full anonymity.		
Responsibility mapping (City vs State competence)	Moderators report frequent confusion about who is responsible for certain issues (municipality vs national authorities). The platform does not currently help to distinguish between City and State competence, which complicates communication with citizens.	Medium	Add support for competence and responsibility mapping in topics and proposals. Allow staff to tag each topic or proposal with the responsible institution (e.g. specific municipal department or a State authority) and show this clearly in the UI (e.g. "This issue is handled by: ..."). Where the municipality is not competent, provide templates or guidelines to help staff explain this and, where possible, indicate the appropriate authority or next step for citizens.

5.1.1 Moderators

Martin

Questionnaire results

A small group of four municipal staff members in Martin took part in the Phase 2 moderator / demonstrator activities. Their roles covered different parts of the administration, including policy/strategy, IT/administration and moderation/support. This makes them a relevant proxy for how the municipality as an organisation might adopt and embed the platform.

In terms of prior experience with AI tools, the group was mixed. One moderator reported never having used AI functionalities such as automatic summaries or content rules in their work, one reported using them rarely, and two indicated they had used them sometimes. None described themselves as heavy or expert users of AI in their day-to-day tasks.

For the focus of the session, three out of four moderators selected "quickly summarise a long text" as the main aim of that day's work with ITHACA, and one selected "moderate a discussion". In an open follow-up question ("How would the ITHACA platform help you most today?"), the moderators highlighted three main expectations:

- having a quicker, more structured overview of citizens' requests and proposals,
- learning to use the platform fully in order to support their team, and
- using what they described as a "very smart tool" that could assist in multiple aspects of their work, from monitoring debates to preparing briefings.

Taken together, these answers show that the moderators came to the session with concrete, work-oriented expectations. They were not just "testing a tool" but explicitly looking for ways to speed up information processing and to make sense of public input.

Evaluation of AI-generated summaries

The DEM mini-survey about summaries asked moderators to rate how well the AI summary helped them understand the key points of a long discussion, to say whether any important viewpoints were missing and to indicate whether they would use such summaries in their work.

On the 1–5 scale (“Did the summary help you quickly understand the main points?”), the scores were 1, 3, 4 and 5, yielding an average of 3.25. Two moderators felt that the summary helped them clearly (scores 4 and 5), one considered it moderately helpful (3), and one found it poor (1). This indicates that, while the summary function is generally seen as useful, its performance is not yet consistent across cases or users.

When asked whether any important viewpoint was missing from the summary, two moderators answered that all important views were included and two indicated that something important was missing. In their explanations, they pointed to aspects such as:

- the need to preserve more of the context and nuance of the original discussion,
- the risk that over-compression could make the content feel “too short” and lose what they considered the essential part of what was said, and
- the wish for clearer linking between the summary and the underlying proposals or comments, so that they could easily trace where specific points came from.

Regarding future use in their work, none of the moderators selected a simple “yes”. Instead, three answered “yes, if some changes were made” and one selected “not yet”. The requested changes mainly concerned:

- making the summary more complete and precise, especially for use in sensitive or formal documents, and
- avoiding over-simplification that could lead to misinterpretation of the situation or to skewed emphasis. Table 9 summarises these key indicators.

Table 9. Perceptions of AI summaries (Martin; Moderators)

Indicator	Distribution / value
Helpfulness of summary (1–5)	Scores: 1, 3, 4, 5 – mean 3.25
All important views included?	2 “no important view missing”; 2 “yes, something was missing”
Would use such summaries in their work?	3 “yes, if something changed”; 1 “not yet”

Use cases, readiness and willingness to champion the tool

In a separate DEM mini-survey, moderators were asked in which concrete work tasks they would use the summary functionality. They could select multiple use cases. Across the four respondents:

- Moderating discussions and preparing a publicly shared summary were each selected by three moderators.
- Preparing news items or press releases was selected by two.
- Creating agendas and preparing briefing materials for leadership were each selected by one.

This pattern suggests that moderators see the summary tool as most relevant for day-to-day moderation and for communicating key points outward (public summaries, press), with a more selective but still present role in agenda-building and internal briefings.

On a 1–5 scale asking “How ready is this solution for everyday use in your team?”, the four responses were 2, 3, 3 and 5, again averaging 3.25. No one rated it as clearly not ready (1), but only one

moderator felt it was fully ready (5); the other three placed it in the “somewhere in the middle” area. When invited to specify what would make the solution “ready for use”, their comments focused on:

- integrating the feedback and fixes identified during the session,
- clarifying the separation between formal proposals and simple comments so that summaries do not mix or misrepresent them, and
- conducting more testing to ensure stable behaviour across topics.

Interestingly, when asked whether they would be willing, after these improvements, to act as an internal ambassador for the solution (presenting and recommending it to colleagues), all four moderators selected the “maybe” option. None rejected the idea outright, but none committed fully either, which is consistent with their “medium” readiness ratings.

Fairness and robustness of AI moderation

The moderators also completed a borderline-item exercise designed to probe fairness and robustness of the AI moderation system. For three short content snippets (“phrases”), they were asked:

- what decision they would take (keep / remove),
- whether the system’s decision matched their own, and
- whether the wording of the statement (including group references) influenced their decision in an unfair way.

They also inspected a pair of near-identical posts that differed only in the group mentioned and were asked whether these received the same outcome, as they should.

The results can be summarised as follows:

- For Phrase 1, the system decision matched the moderator’s own decision in three out of four cases; one moderator reported a mismatch. None of the four felt that the mention of a group had unfairly influenced their decision.
- For Phrase 2, all four moderators reported a match between their own decision and the system’s output. Three said that the wording did not influence them unfairly, while one indicated that it did, and briefly explained why.
- For Phrase 3, three moderators again reported a match with the system’s decision, while one noted that the system’s decision was “not displayed”, making the comparison impossible. All four indicated that the wording had not unfairly affected them.

On the question about the two almost identical posts with different group labels, two moderators answered that the posts did receive the same outcome, as they should, and two answered “not sure” rather than a clear “yes” or “no”. No one indicated that the pair had clearly been treated differently.

Overall, in this small sample, the moderators’ answers suggest that AI moderation decisions were largely aligned with their own judgments on the tested borderline items and that explicit concerns about unfair influence of group mentions were rare. At the same time, the presence of a few mismatches, one case where the system’s decision was not visible and some uncertainty about the behaviour on identity-swapped pairs underline the need for ongoing monitoring of fairness and transparency, especially as the system is used with a wider variety of content.

Privacy, misuse and conditions for trust

The DEM privacy mini-survey asked moderators whether there is anything they would want to mask or hide from internal or external publication when using the platform, whether they are concerned that the system could be tricked or misused and what would reassure them.

Three moderators answered that there is something they would prefer to hide, while one answered “no”. In the open explanations, those who answered “yes” consistently referred to personal data and identifiers, such as names, nicknames and other details that could point to specific individuals, and to any content that falls under GDPR obligations or internal discussion rules. This shows a clear awareness that even in a civic participation context, not all information should be exposed in the same way to all audiences.

On the question “Are you worried that the system could be fooled or misused in your context?”, all four moderators selected variants of “maybe / yes”, indicating that they do see a risk of abuse. The measures that would reassure them focused strongly on data protection and security, including:

- clear rules and safeguards on how the tool is used,
- strong cybersecurity measures, and
- visible emphasis on the protection of personal data.

These answers mirror the project’s broader emphasis on algorithmic impact assessment and highlight that organisational trust will depend as much on the governance and safeguards around the system as on the technical quality of the AI models.

Top-priority fixes and additional comments

Finally, moderators were invited to list their top three fixes and to rank them in order of importance. The free-text answers, taken together, point to four main areas of improvement:

1. Graphical summary and interaction design

Moderators requested a more stable behaviour of the graphical summary (for example, avoiding moving/static background issues when clicking) and clearer, slower interaction when exploring percentages in charts.

2. Linking summaries to underlying content

One moderator stressed that the final graphical summary should be more tightly connected to the concrete proposals. They would like to see full proposal texts accessible directly from the summary and to avoid situations where only fragments are visible.

3. Contribution and commenting workflow

Another moderator called for greater freedom and clarity in how citizens can add proposals, comments and reactions, to make sure that the process of contributing to a topic is transparent and easy to follow.

4. Readability and layout (discussion interface, FAQs)

Suggestions included moving the FAQ block away from each individual topic page and towards the main landing page, improving font size and overall readability in the discussion interface and continuing to iterate and test the interface as real use cases accumulate.

Most additional comments were positive and appreciative, including one explicitly thanking the team and describing the collaboration as very pleasant. This tone suggests that, despite the critical feedback and the clear list of requested improvements, the moderators are constructively engaged and see potential value in the platform for their work.

Focus group

In addition to the individual DEM surveys, a focus group with municipal staff and moderators was organised in Martin to explore how the ITHACA platform fits into real administrative workflows and to gather richer evidence for the AIA. The session followed the dedicated Moderators focus-group guide and protocol (Annex 2), with a 60–75 minute online discussion structured around five elements: (a) anchoring the discussion in a real workflow scenario, (b) usefulness and coverage of AI-generated summaries, (c) moderation consistency and fairness, (d) privacy and security expectations, and (e) organisational uptake and prioritised fixes.

The group comprised four municipal staff members involved in policy-making, communication and moderation. All had previously used the platform in the Phase 2 sessions and had completed the DEM survey blocks, ensuring that the discussion built on hands-on experience rather than abstract impressions.

Workflow scenario and role of ITHACA

The focus group began by asking each participant to name one concrete task where public input matters in their daily work (e.g., preparing an agenda item, drafting a briefing, or moderating a contentious discussion). The group then agreed on a shared reference scenario, i.e., preparing an internal briefing and a short public update on a topic that had been discussed by citizens on the platform (e.g., improving local public spaces or mobility-related issues).

In this scenario, participants saw the ITHACA platform, and especially its AI components, as potentially helping in three main ways:

- Scanning and condensing large volumes of comments into a small number of key points suitable for decision-makers.
- Highlighting points of disagreement or minority views that might otherwise be overlooked when staff are under time pressure.
- Providing a traceable record of how citizen input was considered, through links between summaries, moderation decisions and the underlying thread.

At the same time, they emphasised that the platform would only become part of their routine workflow if it could be trusted to be both accurate and fair and if it respected existing organisational rules on privacy and communication.

AI summaries in workflow – utility, coverage and edits needed

Building on the DEM-2 “Summary” survey block, the facilitator shared a long discussion thread with its AI-generated summary and re-anchored the conversation in the chosen scenario. Participants were asked whether this summary would save them time, whether it captured the key views

(especially minority opinions), and what edits would be needed to make it “ready to paste” into an agenda or briefing.

The group agreed that, at a high level, the summary was time-saving. Instead of reading dozens of posts, they could quickly understand the main lines of the discussion and decide what to investigate further. Several participants said they would use such a summary as a starting point for:

- a one- or two-line brief for leadership (“What are citizens mainly asking for?”) or
- a short problem description to introduce an agenda item.

However, they also identified specific gaps and risks:

- In some threads, the summary was perceived as too generic, lacking details about who was affected or where the problem occurred (e.g., which neighbourhood).
- Participants worried that minority or more critical views could be under-represented when the summary compresses multiple posts into a small number of bullets. They noted examples where concerns of a smaller group were important for policy (e.g. accessibility or safety) but appeared only weakly, if at all, in the summary.
- They found it difficult to trace particular sentences back to concrete posts, which limited their ability to verify whether nuanced positions were reflected faithfully.

The moderators therefore expressed conditional trust in the summaries. They are useful as an orientation and drafting aid, but not yet something they would copy verbatim into a sensitive public communication. To make summaries “ready now” for internal and external use, participants suggested:

- including at least one explicit sentence that captures dissenting or minority views,
- adding a very short context line (e.g. “Discussion on X topic in Y area, N posts”), and
- allowing quick navigation from a summary bullet to example underlying posts (e.g., a link or highlight).

Moderation consistency and fairness

The second part of the focus group revisited the borderline-item from the DEM-3 survey, this time in group discussion. Moderators examined several neutralised borderline posts and considered whether to keep, remove or escalate them under their policies and whether the system’s decision aligned with their expectation.

Overall, the moderators reported that, in most of the examples they had seen during the Phase 2 sessions, the system decisions were broadly aligned with their own. Clearly abusive or hateful statements tended to be flagged or removed, while clearly acceptable content was left untouched. They appreciated that the examples used in the study were subtle and realistic, rather than extreme.

However, the discussion highlighted several important nuances:

- For some borderline items (e.g., strong criticism expressed in emotional or sarcastic language), the group did not always agree internally on “keep vs. remove”. In those cases, even a perfect AI system would inevitably disagree with at least some staff.
- In one or two examples, participants felt that the system was too strict, flagging statements that they would prefer to keep in the spirit of open debate; in others, they would have liked to see a “softer” intervention, such as a warning or prompt, rather than outright removal.

The group concluded that AI-assisted moderation is acceptable as a decision-support tool, but that human oversight and clear policy guidelines must remain central. They also stressed the need for:

- transparent explanations or labels for why a post was flagged, and
- internal documentation of how borderline categories are handled, so that AI and human decisions are as consistent as possible.

Privacy and security expectations

The third block of the focus group focused on privacy and security, following the DEM-4 content and the demonstrator guide. Moderators were asked what they would want to mask before sharing exports or screenshots, what misuse scenarios they worried about, and which safeguards would make them comfortable using the platform routinely.

There was strong convergence on three points:

1. Data to mask before sharing

Participants agreed that names, user IDs and any directly identifying details should be either removed or pseudonymised in exports destined for wider circulation. In some cases, they would also prefer to paraphrase quotes, especially for sensitive topics, to reduce the risk of identifying individuals in small communities.

2. Risks and misuse scenarios

Moderators raised concerns that a malicious actor could:

- “game” the system by posting coordinated content that looks acceptable but systematically pushes a narrative;
- use automated scripts or copy–paste campaigns to flood discussions; or
- exploit any weaknesses in content filters to inject spam or harmful links.

They also mentioned the generic risk that, if AI summaries were misinterpreted as “official positions”, this could damage trust if errors occurred.

3. Requested safeguards

To address these worries, participants asked for visible and documented safeguards such as:

- an audit trail indicating when AI outputs are updated or overridden,
- clear labelling of AI-generated elements,
- rate limits or throttling to discourage spammy behaviour, and
- internal guidelines on how exported content from ITHACA may be used in reports and public communications.

These expectations align closely with the AIA framework used in the project and underline that organisational trust depends not only on model performance but also on the surrounding governance.

Organisational uptake and readiness

The final part of the focus group focused explicitly on organisational uptake and the platform’s readiness for routine use, complementing the DEM-5 and DEM-6 survey blocks.

When asked “Where exactly would this help next month?”, moderators identified several concrete settings:

- drafting briefing notes for leadership based on citizen discussions;
- preparing agenda items or background statements for council or committee meetings;
- supporting communications and public updates, by extracting the main themes and concerns; and
- assisting moderation teams in triaging posts and focusing their time on the most problematic or contested content.

At the same time, they characterised the solution as “on a good path but not fully ready” for routine use. This echoes their DEM ratings indicating medium readiness. They emphasised that a few targeted improvements could “unlock” wider uptake, particularly if accompanied by internal guidelines and training.

Top-3 fixes from the moderator focus group

At the end of the session, moderators were asked to agree on the Top-3 fixes that would most increase their willingness to use and champion the platform in their organisation. The priorities were summarised in Table 10.

Table 10. Focus group with moderators in Martin (Top-3 organisationally relevant fixes)

Priority area	Description from moderators	Underlying organisational need
1. Summary quality and traceability	Make summaries slightly richer, ensure minority views are explicitly included and allow quick navigation from each summary bullet to example posts.	Enable staff to use summaries confidently in briefings and agendas without fear of misrepresentation.
2. Moderation workflow and explanations	Clarify how AI flags are shown in the dashboard, provide clear labels/explanations for decisions and support consistent handling of borderline items.	Support fair, accountable moderation aligned with municipal policies.
3. Privacy and export safeguards	Provide masking options and clear export rules (what is anonymised, who sees what, watermarking), plus visible audit trails.	Ensure GDPR-compliant, safe reuse of platform outputs in internal and public documents.

The group stressed that these fixes are feasible and targeted, rather than requiring a fundamental redesign. In their view, addressing them would significantly increase both operational readiness and willingness to act as internal ambassadors for ITHACA within the municipality.

Brasov

Questionnaire results

A group of five municipal staff members in Braşov took part in the Phase 2 moderator / demonstrator activities. Their roles covered critical areas of the administration, including Policy/Strategy, IT/Administration and Public Communication (drafting messages/press notes). This composition provides a representative view of how the municipality handles digital interaction and decision-making.

In terms of prior experience with AI tools, the Braşov group was notably more experienced than the Martin cohort. Three out of five moderators reported using AI functionalities “often” in their work, while others reported occasional use. This suggests a higher baseline of digital literacy and consequently higher expectations regarding the performance of the tools.

For the focus of the session, the moderators selected tasks related to "Drafting a public message/press note" and "Preparing a decision or meeting briefing." In the open follow-up question ("How would the ITHACA platform help you most today?"), the moderators highlighted distinct operational needs:

- **Identifying responsibilities:** clearly showing "who manages what" issue within the city apparatus,
- **Translation and summarization:** the ability to "translate long, complex texts" and provide coherent summaries, and
- **Creating rapid dashboards:** generating a "public-facing summary or dashboard" to communicate issues transparently.

Taken together, these answers show that the Braşov moderators viewed the platform not just as a listening tool, but as an active administrative aid for routing issues and communicating back to the public.

Evaluation of AI-generated summaries

The DEM mini-survey about summaries revealed a significant gap between user expectations and the current system performance, largely driven by linguistic issues. Moderators were asked to rate how well the AI summary helped them understand the key points.

On the 1–5 scale, the scores were overwhelmingly negative, with recorded values of 1 ("Very low"). Unlike in Martin, where results were mixed, the Braşov moderators found the summaries "incoherent" in Romanian. Comments included strong descriptors such as "it is dust" ("e praf") or "Google Translate is much superior," indicating that the native language generation was below the usable threshold.

When asked whether any important viewpoint was missing, moderators answered "Yes" and pointed out that "many" points were lost due to the poor quality of the translation and summarization logic. They noted that the tool failed to capture the nuance required for official work.

Regarding future use, the response was a unanimous "Not yet" or conditional rejection. The requested changes were fundamental rather than incremental:

- Full integration of a capable translation API (comparable to commercial standards) and
- Coherence in the output so that it reads as natural Romanian rather than a mechanical translation. Table 11 summarises these key indicators for Braşov.

Table 11. Perceptions of AI summaries (Braşov; Moderators)

Indicator	Distribution / value
Helpfulness of summary (1–5)	Scores clustered at 1 (Low)
All important views included?	"No" – significant loss of information due to language quality
Would use such summaries in their work?	"Not yet" / "No" – requires major linguistic overhaul

Use cases, readiness and willingness to champion the tool

In the DEM mini-survey regarding work tasks, moderators selected "Drafting a public message/press note" and "Moderating a thread" as primary use cases. This confirms that if the tool worked as intended, it would sit at the intersection of communication and operations.

On the 1–5 scale asking "How ready is this solution for everyday use?", the ratings were low to mixed (e.g., 2, 2, 4), yielding an average below the threshold of acceptance (approx. 2.6). The moderator

who rated it higher (4) still noted that it required a "new team" to manage it, while others stated that "complete and correct translation" was the mandatory condition for readiness.

When asked whether they would act as an internal ambassador, the answers were split between "No" and "Yes," contingent entirely on fixing the language issues. This indicates that while the *concept* is supported, the *current implementation* cannot yet be championed internally.

Fairness and robustness of AI moderation

The moderators completed the borderline-item exercise to probe fairness. The results for Braşov showed a notable deviation from the Martin pilot regarding the consistency of the system.

- **Phrases 1–3:** In the majority of cases, the system's decision to "Keep" or "Remove" matched the moderators' own judgment. The moderators did not report that the specific wording of groups unfairly influenced their decisions.

This suggests that, unlike in Martin, the Braşov moderators detected inconsistency in how the AI treated similar content when different keywords were swapped. This finding reinforces the "conditional trust" theme seen in the summary evaluation. The AI is perceived as potentially unstable or inconsistent in the local language context.

Privacy, misuse and conditions for trust

The DEM privacy mini-survey asked about masking data and misuse risks.

- **Masking:** Moderators expressed a desire to mask "fragments with extremely vulgar or threatening language" when sharing reports internally, to avoid negative impact on readers. They also emphasized the need to protect user identities.
- **Misuse:** Most moderators answered "Maybe/Yes" to concerns about the system being fooled. Specific worries included "gaming the vote" (fake engagements) and the lack of "real moderation" (human oversight) to catch nuanced abuse.
- **Safeguards:** To be reassured, moderators requested:
 - "Real moderation, without anonymity,"
 - Clear explanations of *why* content is blocked based on local laws/rules, and
 - Mechanisms that are "impossible to fool" regarding voting.

Top-priority fixes and additional comments

The free-text priorities from the Braşov moderators diverged significantly from Martin, focusing heavily on basic functionality and localization:

1. Language and Translation Quality

The most critical issue was the quality of the Romanian text. Moderators demanded "complete and correct translation of the entire content" and "fixing encoding issues." Without this, the advanced AI features are seen as unusable.

2. Toxicity and Identity Management

Moderators requested "efficient moderation that automatically blocks toxic content" and urged the removal of anonymous modes to prevent abuse.

3. Responsibility Mapping

A unique request from Braşov was for the platform to help identify "City vs. State competence", clarifying which institution is responsible for a proposal so that citizens do not blame the municipality for issues outside its jurisdiction.

Focus group

In addition to the surveys, a focus group with municipal staff was organised in Braşov (N=5) to explore the strategic fit of the platform. The session revealed that while the intent of the platform is appreciated, its current positioning overlaps confusingly with existing tools (like BrasovCity), and its technical execution in Romanian undermines trust.

Workflow scenario and role of ITHACA

Participants struggled to anchor ITHACA in their current workflow because of the linguistic barrier. However, in an ideal scenario, they envisioned the platform as a "triage and routing" engine:

- It should take citizen input,
- Clean it of toxicity,
- Identify the correct department or institution (City vs. Government), and
- Present a coherent summary to the specialist.

AI summaries in workflow: utility, coverage and edits needed

The discussion on summaries was dominated by the "illusion of translation" finding. Participants noted that while the interface was translated, the content handling (summaries, graph labels) was often "broken" or "hallucinated" in Romanian.

- **Utility:** Currently zero for official use, but potentially high if fixed.
- **Edits needed:** Participants stated they would not edit the summaries. They would simply *not use them* in the current state because rewriting them would take longer than reading the original posts.
- **Trust:** Trust is currently suspended until the "Language setting works stably" and the text generation is native-level.

Moderation consistency and fairness

The group reaffirmed the survey findings. The AI misses local vulgarities and regional slang. A major point of discussion was the "under-detection" of toxicity, where clearly offensive local terms slipped through the filters. Moderators argued that for the platform to be safe, the "toxic vocabulary" library must be significantly expanded to include regionalisms.

Privacy and security expectations

The focus group highlighted a strategic risk regarding anonymity. Participants argued that allowing anonymous or pseudonymous input encourages low-quality or abusive content ("rudeness").

- **Safeguard:** They recommended enforcing identifiable accounts (at least internally verified) to raise the quality of discourse.
- **Data Export:** Exports must be "clean" (meaning free of vulgar examples) before being shared with leadership.

Organisational uptake and readiness

When asked about readiness, the group identified a clear path to readiness. If the language engine is fixed, and if the platform clearly distinguishes itself from the existing "BrasovCity" reporting tool, it could be adopted as a "deliberative layer" on top of the existing administrative tools.

Top-3 fixes from the moderator focus group

The priorities were consolidated in Table 12. Unlike Martin’s focus on "tweaks" (linking summaries, layout), Braşov’s needs are structural.

Table 12. Focus group with moderators in Braşov (Top-3 organisationally relevant fixes)

Priority area	Description from moderators	Underlying organisational need
1. Linguistic integrity (Translation & Encoding)	Ensure the language selection sticks (no random reverts to English) and that summaries/translations are coherent and native-sounding.	Basic usability: The platform cannot be used officially if the text output is broken or unintelligible.
2. Strategic Differentiation & Responsibility	Clarify "Why ITHACA?" vs. existing tools. Show clearly who is responsible (City vs. State) for each proposal.	Institutional clarity: Prevent citizen frustration and administrative overload by routing issues correctly.
3. Strict Moderation & Identity	Eliminate easy anonymity; expand the toxic vocabulary to catch local slurs; prevent "gaming" of votes.	Safety & Trust: Ensure the environment is professional enough for civil servants to engage without reputational risk.

Patterns and subgroup differences

Martin

Beyond the overall distributions reported above, the Martin dataset allows a more nuanced look at how different types of users experienced the platform. By linking the baseline A1 mini-survey with the session-level A2/A4 questionnaires, and by grouping sessions according to the themes reported in the "What slowed you down?" question, it is possible to identify profiles of use and recurring patterns that are not visible in the aggregate means alone.

For citizens, the most informative distinctions arise from their **baseline participation habits, trust in AI-generated content** and **use of accessibility tools**. When sessions are grouped according to whether their authors reported, at baseline, that they *never/rarely* participate in public online discussions versus *sometimes*, a consistent pattern emerges. Sessions coming from the very low participation group tend to show slightly lower ease-of-use and satisfaction scores and a higher proportion of neutral evaluations, while sessions from the "sometimes" group more often fall into the clearly positive range. This is also reflected in the willingness-to-reuse item. Although the majority of sessions in both groups end with a "yes, I would use this platform again", the proportion of "yes" responses is higher among users with some prior experience of engaging in online debates. In other words, the platform is able to elicit positive experiences even among citizens who are not habitual contributors, but those with a minimal prior engagement baseline appear to gain confidence more quickly.

A similar, though less pronounced, gradient can be observed when citizens are grouped by **baseline trust in AI summaries and rules**. Participants who initially expressed very low trust in AI-generated content are somewhat more likely to assign neutral or lower scores to session ease and satisfaction and to mention AI-related issues in the open question (e.g., unclear or non-functioning summaries, difficulties with toxicity tests). Those who began with medium to higher trust, in contrast, more often rate the sessions as straightforward and satisfactory and their comments about AI tools are more focused on fine-tuning than on fundamental problems. The numbers involved are small and should be interpreted with caution, but the direction of these differences is coherent with the qualitative focus-group finding that trust in summaries is “conditional” and tied to perceived coverage and representativeness.

Baseline **use of accessibility and assistive tools** also helps to explain some of the variation observed in the mini-surveys. Sessions attributed to citizens who had declared using larger fonts, keyboard navigation or similar aids at baseline show a slightly higher incidence of “partly suitable” on the accessibility suitability item and a somewhat higher frequency of comments about readability, scrolling and keeping track of one’s place. However, there is no evidence that accessibility-tool users systematically rate the platform as unusable. Most of their sessions still fall in the middle or positive range for ease and satisfaction and the dedicated accessibility micro-surveys (A3) show that the large majority of these participants can complete their tasks with their usual settings. This suggests that, while residual accessibility issues exist, they affect comfort and efficiency more than basic feasibility of use.

The coding of the open “What slowed you down?” answers provides an additional lens on session quality. When sessions are grouped by the main theme mentioned (“nothing”, “navigation and structure”, “AI tools and summaries”, “technical issues” or “other”) clear differences emerge. Sessions in which citizens wrote that nothing in particular slowed them down tend to have higher average ease and satisfaction scores and a higher proportion of “yes” on reuse intention. Conversely, sessions associated with navigation-related comments more often receive scores in the neutral or negative range and those linked to AI-tool or technical issues show a noticeable increase in “no” responses on the reuse item. This cross-tabulation confirms that the qualitative themes identified in the comments map directly onto the quantitative experience and supports the prioritisation of navigation clarity, AI-tool robustness and technical stability in the improvement recommendations.

To synthesise these patterns, a simple “session quality” index can be defined by combining ease-of-use and satisfaction ratings into three bands: high-quality sessions (both ratings high), medium-quality sessions (mixed or neutral ratings) and low-quality sessions (at least one rating clearly negative). Applying this index to the Martin mini-survey data shows that a majority of sessions fall into the high-quality band, a substantial minority into the medium band and a smaller but non-trivial share into the low-quality band. High-quality sessions are predominantly associated with citizens who have at least some prior participation experience, report no blockers or only minor issues, and often indicate that they would use the platform again. Low-quality sessions, by contrast, cluster among users with very limited prior participation, lower initial trust in AI and comments about navigation difficulties or non-functioning tools. This reinforces the conclusion that targeted improvements for these specific pain points are likely to have the greatest impact on overall experience.

For moderators and municipal staff, the small sample size precludes extensive statistical analysis, but a structured cross-case view still yields useful insights. When the four moderators’ responses to

the DEM mini-surveys are considered side by side, it becomes clear that they form a **heterogeneous but coherent group**. All view the platform as potentially helpful for summarising and communicating citizen input, but they differ in how ready they feel to use it in everyday work and in how complete and trustworthy they consider the current AI-generated summaries. The individual ratings on summary helpfulness and solution readiness cluster in the mid-range, with only one moderator expressing a very high level of readiness and one clearly more sceptical. Importantly, all four select a “conditional” option when asked whether they would use the summaries in their work (“yes, if some changes were made” rather than a simple “yes”), and all express a “maybe” rather than a firm “no” or “yes” when asked about acting as internal ambassadors for the tool. This pattern supports the interpretation that the organisational stance is cautiously positive but contingent on specific improvements.

The borderline-item exercise adds a further dimension to this picture. When examining near-identical content snippets, moderators often agree with each other and with the system on clearly acceptable or clearly unacceptable posts but differ among themselves on more ambiguous items. In a few cases, they perceive the system as stricter than they would prefer, or they would have liked a softer intervention such as a warning prompt rather than outright removal. This shows that some level of mismatch between human and AI decisions is inherent to the task and reflects genuine normative ambiguity rather than a purely technical flaw. At the same time, the fact that some moderators could not always see or easily interpret the system’s decisions underscores the need for better explanations and more transparent presentation of AI flags in the moderation interface.

Taken together, these additional analyses deepen the Martin results without changing the overall conclusion. The ITHACA platform is broadly usable and acceptable for both citizens and moderators, but its benefits and perceived readiness are not evenly distributed across all user profiles. Citizens with little prior experience of online participation, lower baseline trust in AI or specific accessibility needs are more likely to encounter friction and to have medium or low-quality sessions, while more experienced participants move quickly into stable, high-quality patterns of use. Moderators see clear potential for time savings and more structured handling of citizen input, but remain cautious until summary quality, traceability, moderation explanations and privacy safeguards are further strengthened. These patterns are directly reflected in the Top-3 citizen and moderator priorities and should inform the final iteration of design and AIA-driven refinements.

Braşov

For citizens, the most informative distinctions arise from their baseline participation habits and use of accessibility tools. When sessions are grouped according to whether their authors reported, at baseline, that they never/rarely participate in public online discussions versus sometimes/often, a consistent pattern emerges. Sessions coming from the “low participation” group (Never/Rarely) show slightly lower ease-of-use (mean 4.8) and satisfaction scores compared to the “high participation” group, whose ratings cluster exclusively at the ceiling (mean 5.0). However, unlike in Martin, this difference does not extend to the willingness-to-reuse item. Both groups reported a unanimous (100%) intention to use the platform again. In other words, while habitual contributors found the interaction slightly more fluid, the platform successfully elicited strong engagement and retention intentions even among citizens with no prior history of digital civic participation.

A similar gradient is observed regarding baseline **trust in AI**. In Braşov, the baseline trust was notably higher than in Martin, with the vast majority of participants starting the pilot with “medium” or “high” trust in AI tools. Consequently, we do not see a distinct “low trust” cluster dragging down the scores. However, when examining the specific subset of sessions where users encountered AI failures (e.g., poor translation or summarization), the satisfaction scores drop significantly (to a mean

of = 3.0). This reinforces the qualitative finding that trust in Braşov is highly performance-dependent: users are open to AI, but their satisfaction plummets quickly when the tool fails to deliver accurate (and linguistically correct) outputs.

Baseline use of **accessibility and assistive tools** explains a small but distinct part of the variation. Sessions attributed to citizens who declared using larger fonts, high contrast, or keyboard navigation at baseline (approx. 33% of the sample) show a slightly lower mean Ease-of-Use score (4.8) compared to those who use no assistive tools (5.0). Qualitative comments from this group point to specific friction points such as "unclear button labels" or "contrast issues," rather than fundamental blockers. The dedicated accessibility micro-surveys (A3) confirm that while 100% of these users eventually succeeded in their tasks, they were more likely to report that they "had to ask for help" or "looked at the tutorial" compared to non-assistive users. This suggests that the platform is technically accessible but requires higher cognitive effort from users with visual or motor impairments.

The coding of the open "What slowed you down?" answers provides the clearest lens on session quality. When sessions are grouped by the main theme mentioned, sharp differences emerge:

- **"Nothing" / "Nimic":** Sessions where users reported no slowdowns account for the majority of the data and are associated with perfect satisfaction scores (5.0).
- **"Navigation":** Comments related to finding buttons or links ("unde trebuie să dau click", "nu am găsit") are associated with a slight dip in Ease-of-Use (mean 4.7) but retention of high satisfaction.
- **"AI / Language / Technical":** Sessions where users cited "AI translation," "strange summaries," or "technical bugs" (refresh/login) show the lowest Ease-of-Use scores (ranging from 3.0 to 4.0).

To synthesise these patterns, the "session quality" index (combining ease and satisfaction) reveals that the Braşov pilot achieved a higher proportion of "High-Quality" sessions than Martin, driven by a very enthusiastic user base. However, the "Low-Quality" sessions in Braşov are almost exclusively linked to technical and linguistic failures (e.g., broken translations, login loops) rather than user incompetence or lack of trust. This reinforces the conclusion that the primary barrier in Braşov is not user adoption or digital literacy, but the stability and localisation quality of the platform itself.

For moderators and municipal staff, the cross-case analysis reveals a starker contrast with the citizen data. While citizens were generally forgiving, the five moderators formed a highly critical user group. When their responses to the DEM mini-surveys are considered side by side:

- **Summary Helpfulness:** Moderators unanimously rated the AI summaries at the bottom of the scale (1/5), explicitly citing the poor quality of the Romanian language generation ("e praf", "incoherent").
- **Readiness:** The "Readiness for everyday use" scores averaged 2.8/5, with users split between "not ready" and "neutral."
- **Ambassadorship:** Willingness to champion the tool was mixed ("No" to "Maybe"), and strictly conditional on fixing the translation engine.

The borderline-item exercise further distinguished the Braşov moderators. Unlike in Martin, where moderators largely agreed with the AI, the Braşov group identified inconsistencies in the system. When presenting identity-swapped pairs (near-identical posts differing only by the group mentioned), the majority of moderators (3 out of 5) noted that the system did not treat them equally. This indicates that the Braşov moderators were highly sensitive to the AI's lack of cultural and linguistic nuance, leading to a "conditional trust" that is currently suspended until technical improvements are made.

Taken together, these analyses depict a polarized experience in Braşov. Citizens, regardless of their background or accessibility needs, had a highly positive, "high-quality" experience, finding the platform easy and engaging despite minor friction. Moderators, however, faced the brunt of the linguistic limitations, resulting in low readiness scores and low trust in the AI's specific outputs. This divergence highlights that while the concept and workflow of ITHACA are validated, the technical localisation (specifically for complex languages like Romanian) remains the critical bottleneck for organisational adoption.

4.2.3 Gamification evaluation (UPAT)

The UPAT study focused on the evaluation of the gamification layer in a controlled lab setting with nine university students. Participants interacted with the ITHACA platform using missions, points, badges and a leaderboard, while the research team collected:

- a pre-questionnaire (expectations and prior experience with gamification),
- repeated micro-surveys during the session (momentary experience and psychological needs),
- a mid mini-survey (flow, perceived control and fun, plus what helped/hurt motivation), and
- a post-questionnaire (evaluation of the gamification design, feedback, fairness and future use),

supplemented by a structured list of prioritised issues and recommendations compiled by the UPAT team after the sessions (material can be found in Annex 6).

All quantitative scales are Likert-type (1–7 for pre, micro and mid measures; 1–5 for the post-questionnaire). Given the small sample (N=9) and repeated measures (24 micro-surveys and 11 mid mini-surveys), the results should be interpreted as formative, indicative evidence rather than as statistically generalisable.

Expectations and prior gamification experience

Prior to the start of the session, participants reported high baseline expectations. Both Interest Expectancy and Self-Efficacy yielded an average of 6.1 out of 7, with all ratings clustered in the high 5–7 range. Similarly, Anticipated Enjoyment was robust, with a mean score of 5.6. Participants also expressed a strong Preference for Autonomy and Variety, assigning similarly high ratings (mean 5.6) to dimensions regarding the desire for choice and multiple progression paths.

Regarding Prior Experience with gamified systems, the majority of the cohort was familiar with the concept: five participants reported having "a little" experience, three reported "a lot," and only one reported none. Furthermore, Perceived Utility was unanimously positive, with all nine participants affirming that gamification would make participation more interesting. Consequently, the group commenced the session with strongly positive attitudes and a relatively high level of familiarity with points and rewards; a critical context for interpreting the more critical feedback provided later in the session.

Moment-to-moment experience

During the session, participants completed brief micro-surveys after specific gamified events (e.g. earning a reward or completing a mission). Across all participants there were 24 such snapshots,

providing a picture of how the gamification felt “in the moment”. Figure 12 summarises the central indicators (1–7 scale).

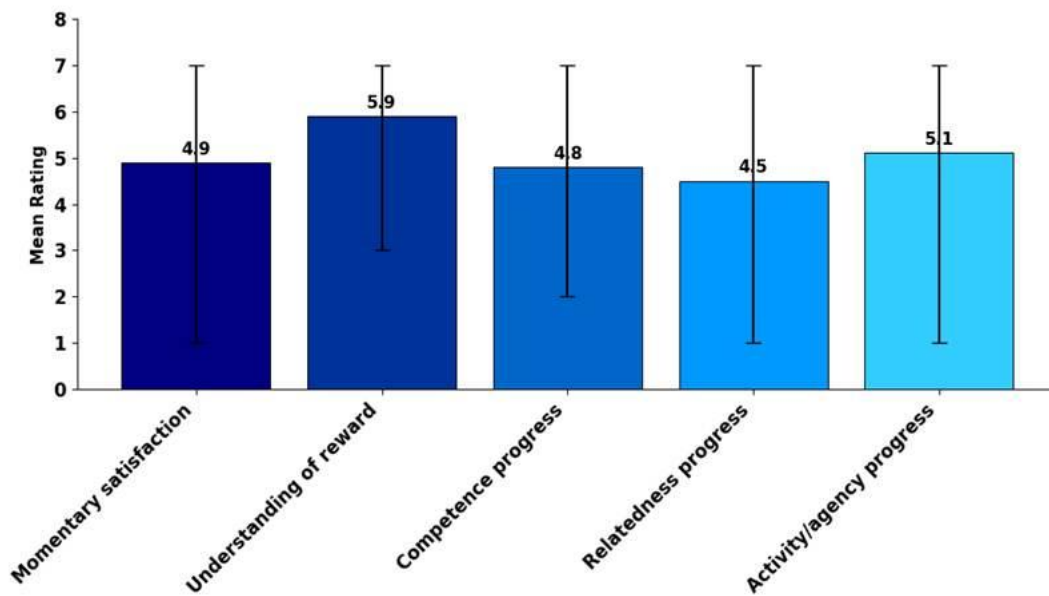


Figure 12. Micro-survey results (Gamification module)

These results suggest that in the moment:

- earning rewards and completing missions was usually experienced as understandable and moderately satisfying,
- the system often supported a feeling of competence and agency, and
- the contribution to relatedness (feeling connected or recognised by others) was more variable.

The free-text “one-word emotion” answers further illustrate this mixed pattern. Words like “Fun”, “Accomplished”, “Acceptance”, “Inclusivity”, “Satisfaction” and “Progress” co-existed with more critical or neutral ones such as “Boring”, “Lost”, “Annoyance” and “what?”. Positive experiences tended to be linked to moments where missions worked smoothly and rewards felt meaningful; negative emotions appeared mainly when technical problems occurred or when the mission structure felt confusing.

Flow, control and motivation

Midway through the session, a mini-survey was administered to capture a reflective view of engagement and motivation. This survey collected 11 responses from the nine participants, as some individuals provided feedback after multiple mission blocks. The quantitative results on a 7-point scale indicated a generally high level of engagement: the item “I felt completely absorbed in what I was doing” yielded a mean score of 5.7, with nearly two-thirds of ratings at 6 or 7. Similarly, participants found the process enjoyable, with “I found the process fun” averaging 5.5 and the majority of ratings clustering at the upper end of the scale. However, feelings of agency were more moderate; the statement “I felt in control as I was completing the goals” received a mean score of 4.6, with most responses falling between 4 and 6, and a few lower ratings in the 2–3 range.

Collectively, these figures suggest that while the mission-based structure successfully fostered a sense of flow and enjoyment, perceived control was somewhat inconsistent.

Qualitative feedback from the open-ended item, "What helped or hurt your motivation so far?", provided context for these ratings. Participants reported that the gamification concept itself—specifically the theme, social interaction (e.g., "talking about Linux," "group engagement") and the exploratory nature of the missions, positively influenced their motivation. Conversely, motivation was hindered primarily by technical instabilities and usability issues. Specific complaints included navigation difficulties, such as broken reload functions requiring the re-entry of links, challenges in locating features, and various bugs that highlighted the platform's "developing stage." Consequently, while the underlying design of the activity was motivating, its full potential was partially offset by implementation hurdles that disrupted the user experience.

Evaluation of gamification design

At the end of the session, nine participants completed a post-questionnaire (1–5 scale) focusing on affect, engagement, clarity and fairness of the gamification elements, and willingness to see this approach used in real student participation. The results show a clear contrast between core experiential aspects, which are mostly positive, and system/reward mechanics, where several weaknesses emerge.

Ratings from the post-session questionnaire indicate that the gamified interaction was experienced as both affectively positive and engaging. On a five-point scale, pleasant affect associated with receiving rewards was rated highly, with a mean of 4.1 and most ratings in the 4–5 range. Absorption in the activity (being "drawn into" the missions) showed a similar pattern, with a mean of 4.2. Exploratory engagement was even stronger. The item capturing experimentation to achieve goals reached a mean of 4.6, with approximately two-thirds of participants selecting the maximum value. Perceived motivational pull of the system was slightly lower but still clearly positive, with an average of 3.9 and the majority of ratings at 4–5, alongside one clearly negative rating (2). Taken together, these results suggest that participants generally experienced the gamification layer as pleasant, engaging and conducive to exploratory behaviour, while leaving some room for strengthening its motivational consistency across users.

System reliability and feedback

Scores on items related to disappointment, progress awareness and interface clarity were more mixed. Perceived absence of disappointment with the system was relatively low, with an average of 2.3 out of 5 and more than half of participants choosing values 1 or 2, indicating that technical or design issues did, in practice, frustrate users during the session. Awareness of personal progress and improvement received a moderate mean rating of 3.4, with roughly half of the responses in the positive range (4–5) and the remainder more neutral or negative, suggesting that feedback on advancement was not consistently salient. Perceived clarity of the Missions screen followed a similar pattern, with a mean of 3.3. Most participants considered it broadly understandable (typically rating it 4), yet around one fifth rated it as 2, signalling that for a non-trivial subset of users the structure and content of the mission overview remained insufficiently clear.

Rewards, progress visibility and leaderboard

The weakest scores in the post-session questionnaire were observed for aspects related to how rewards and progress were communicated. Perceived clarity of rewards was low, with an average rating of 2.4 out of 5 and more than half of participants choosing a value of 2, indicating that they often did not clearly understand what each reward represented or why it had been granted. Visibility of progress over time showed a similar pattern, with a mean of 2.4 and most ratings clustered at 2 or 3, and only a single participant giving a clearly positive rating; this suggests that ongoing advancement through the missions was not consistently apparent. The perceived timing of feedback was also weak, with an average of 2.0 and responses evenly distributed between 1, 2 and 3, implying that users frequently did not see rewards or feedback at the moment they expected them. Finally, perceived usefulness and fairness of the leaderboard were rated very low (mean 2.1), with almost all participants selecting a low value. Taken together, these results align closely with the qualitative comments and the prioritised issues identified in the study: participants often struggled to perceive when and how their actions translated into points, badges or rank, and they did not experience the leaderboard as a transparent or fair representation of their efforts.

Future use in real student participation

Finally, regarding Future adoption intent, the mean rating was 3.6 on a 5-point scale. The distribution of responses indicates a cautious optimism: while five participants responded positively (ratings of 4 or 5), three remained neutral, and one participant indicated strong disagreement. These findings suggest a clear openness to integrating this form of gamification into real-world student participation; however, successful adoption appears contingent upon addressing the specific concerns raised regarding system clarity, reliability, and fairness.

Prioritised issues and recommendations

Based on the questionnaires, in-session observations and a short debrief, the UPAT team compiled a list of 11 prioritised issues and recommendations for the gamification module. These focus less on abstract satisfaction and more on concrete design and implementation changes that would make the system usable and trustworthy. Table 13 synthesises these issues.

Table 13. Priorities and recommendations (Gamification module)

Functionality / Element	Main issue observed in tests	Priority	Design implication (summary)
Missions → “Proposals” path	Players could not easily find where/how to submit a Proposal, so missions tied to proposals stalled and XP felt pointless.	High	Make the “Write a proposal” affordance clear and accessible; explain briefly how proposals differ from comments.
Missions unreachable	Missions such as “Add friends” or “Stay active 20 minutes” were effectively not achievable in normal use.	High	Remove or re-design missions so that all are achievable within a session (e.g. “Post once”, “Support or Object once”).
XP & ranking	XP gains were inconsistent and the leaderboard felt opaque/unfair; users not in the top list could not see their position.	High	Stabilise XP logic and always show “You are #N” (overall and e.g. 7-day); add a short “How XP works” explanation.
Badges & feedback	Players often did not notice badges or understand why they	Medium	Use explicit, timely feedback (e.g. toast notifications when a badge is earned)

Functionality Element /	Main issue observed in tests	Priority	Design implication (summary)
	earned them; feedback after achievements was too subtle.		and a visible progress indicator towards the next badge.
Mission wording / duplication	Mission names felt duplicated or unclear (especially around comments vs proposals); some sounded almost identical.	Medium	Merge duplicate missions and rewrite labels in action-oriented, unambiguous wording (“Post a proposal”, not “Interact”).
Session / logout effects	Background logouts or reloads broke flow and mission progress; users felt “punished” when losing progress.	High	Add a logout warning/countdown and autosave mission state so progress can resume smoothly after re-login.
Mobile access to forum	On mobile, some participants could not reliably access the Community area, blocking missions tied to posts/proposals.	High	Ensure a persistent “Community” entry point on mobile and verify end-to-end paths for all mission types.
Toxicity & fair-play loop	Strongly toxic posts could still appear, undermining the feeling of safety and playful engagement.	High	Add gentle pre-posting nudges (“Keep it constructive”), tighten filters, and consider a “constructive citizen” badge linked to positive behaviour.
Graph of opinions	The main graph mixed comments and proposals, so its relation to missions and decisions felt weak or misleading.	Medium	Separate counts for proposals vs comments and allow filters (e.g. “proposals only”) when missions relate to voting/support.
Page reload / refresh	The reload action did not work reliably; users could not easily see new content or recover their current view/state.	High	Make refresh consistent and preserve the current state where possible; offer a clear way to re-open the current view.
Connections / “Unfriend”	The “Unfriend” action produced a technical error and did not remove the connection.	Medium	Fix the underlying error; ensure that removing a connection works instantly and the UI (counts, labels) updates immediately.

These issues map directly onto the quantitative patterns described above:

- The low scores for reward clarity, progress visibility and leaderboard fairness are reflected in the XP, badge and leaderboard issues.
- The drop in perceived reliability (“the system rarely disappointed me”) aligns with problems around logout, reload and bug-related interruptions.
- Participants’ concerns about safety and fair play connect to the toxicity and fair-play loop, which is particularly important given the project’s emphasis on Algorithmic Impact Assessment.

Summary of findings

In summary, the UPAT gamification study shows that:

- The gamification concept and mission structure successfully generated interest, flow and experimentation among students, who started with very positive expectations and largely enjoyed the session.
- At the same time, the current implementation has critical weaknesses in how it communicates rewards, progress and rankings, and in the stability of session flow (especially around reload and logout behaviour).

- These weaknesses directly affect trust and perceived fairness, which in turn condition whether participants would want to see such gamification adopted in real student participation contexts.

The prioritised issues and recommendations provided by UPAT thus offer a concrete roadmap for improving the gamification module before wider deployment: stabilising core mechanics (missions, XP, badges, leaderboard), making feedback and progress visible and understandable, and strengthening the link between playful engagement and meaningful, safe participation.

4.2.4 Change in perceived usability across sessions and influence of baseline characteristics

Stability of session-level experience (Martin)

For Martin, the post-visit mini-surveys (A2/A4) were designed to track how perceived ease of use, session satisfaction, willingness to reuse the platform and perceived suitability of accessibility functions evolved over repeated visits. As reported earlier, ease-of-use ratings averaged 3.8 on a 1–5 scale, with roughly 60% of sessions rated in the positive range (4–5), about 28% neutral (3) and around 13% clearly negative (1–2). Satisfaction scores were slightly lower, with a mean of 3.6; a little over half of the sessions were rated as satisfactory (4–5), one fifth as neutral and about one quarter as unsatisfactory (1–2). Willingness to reuse the platform was more clearly positive. Around seven in ten session-level responses indicated that participants would use the platform again, while three in ten indicated that they would not.

Taken together, these distributions suggest that no strong “learning curve” effect is visible at the aggregate level. If early visits were systematically much more difficult or frustrating than later ones, we would expect a heavier concentration of low scores in the pooled data. Instead, the pattern is compatible with a moderately positive but heterogeneous experience throughout the evaluation period. Most sessions are perceived as easy enough and satisfactory, but a persistent minority of interactions remain effortful, confusing or disappointing. This reading is supported by the qualitative mini-survey comments, where many participants explicitly state that “nothing slowed them down”, while others repeatedly mention navigation, AI-tool behaviour and isolated technical issues as friction points.

Accessibility-specific mini-surveys (A3) tell a similar story. In almost all of these accessibility-focused sessions (29 out of 31), respondents reported that they were able to complete the task with their usual settings; only one respondent indicated that they could not complete the task under those conditions and one response was missing. Reports of specific accessibility problems such as low contrast, unclear labels or “getting stuck” were rare and typically isolated to one or two cases per category. This suggests that the platform is generally usable for assistive-technology users, although the occurrences of unclear labels and “stuck” states highlight residual issues that can affect a small number of sessions even after users are familiar with the platform.

Change across Visits (Braşov)

Stability of session-level experience (Braşov)

For Braşov, the post-visit mini-surveys (A2/A4) tracked the evolution of user experience across six distinct sessions, offering a high-resolution view of how the platform performed as novelty wore off

and routine use set in. Unlike the Martin pilot, where the aggregated data showed a heterogeneous mix of positive and neutral experiences with no strong temporal trend, the Braşov dataset reveals a remarkable "ceiling effect" and longitudinal stability.

Across the six visits, the core indicators of Ease of Use and Session Satisfaction did not follow a traditional learning curve (starting low and rising) because they started near the maximum possible value.

- **Ease of Use:** Ratings remained consistently high, fluctuating only slightly between group means of 4.8 and 5.0 on the 1–5 scale. For instance, Visit 2 recorded a mean of 4.9, which sustained at 5.0 through Visits 3, 4, and 5, with a negligible dip to 4.8 in Visit 6. This suggests that the interface was intuitive enough to require almost no "struggle phase" for the general user.
- **Satisfaction:** This metric mirrored the ease-of-use scores but showed a slight consolidation of positive sentiment in the latter half of the pilot. While Visits 2 and 3 yielded means of 4.9, the final visits (V5 and V6) achieved a unanimous 5.0/5 rating.
- **Willingness to Reuse:** Perhaps the most striking indicator of stability is the retention intent. Across all six waves, 100% of respondents indicated a willingness to use the platform again. There were no reports of "fatigue" or drop-off in intent, differentiating this pilot from Martin where a persistent minority (approx. 30%) remained sceptical.

Qualitatively, however, the nature of the user experience did shift. In the early visits (V1–V2), the open-text comments regarding "What slowed you down?" frequently referenced orientation and discovery (e.g., "I didn't know where to click," "I had to look for the button"). By the middle and late visits (V4–V6), these comments virtually disappeared, replaced by reports of "Nothing" ("Nimic") or specific, isolated technical feedback (e.g., session timeout issues). This indicates that while the quantitative scores remained static at the top, the qualitative experience matured from "intuitive discovery" to "fluent operation."

The Accessibility Adaptation Curve (A3) While the general usability trends were flat (and high), the dedicated accessibility data (A3) reveals a dynamic process of adaptation. The Braşov cohort included a significant subset of users relying on assistive technologies (screen readers, zoom, high contrast), allowing for a granular analysis of inclusive design performance over time.

In the initial phase (Visits 1 and 2), a visible "accessibility gap" emerged. In both visits, 6.7% of participants (4 out of 15) reported that they could not complete the session using their specific configuration. Comments highlighted friction points such as "unclear button labels," "low contrast," or keyboard focus getting "stuck."

However, a rapid adaptation effect was observed in subsequent sessions:

- By Visit 3, the failure rate dropped to 6.7% (1 participant).
- From Visit 4 through Visit 6, the success rate reached and stayed at 100% (15/15 participants).

This trajectory suggests that the barriers encountered were not fundamental blockers (which would have resulted in persistent failure) but rather learning hurdles. Users with accessibility needs required 2–3 sessions to build a robust mental model of the interface and develop workarounds for the minor labelling or contrast issues. Once this adaptation phase was complete, their performance metrics converged perfectly with the general population, closing the usability gap entirely.

Influence of baseline characteristics

Linking the session outcomes back to the baseline profile (A1) further illuminates these patterns.

- **Prior Participation Habits:** In Martin, users with low prior digital participation rated the platform lower. In Braşov, this correlation was absent. Participants who identified as "Never" or "Rarely" participating in online discussions (approx. 33% of the sample) reported the same ceiling-level satisfaction scores (4.9–5.0) as frequent contributors. This implies that the ITHACA platform, in the Braşov context, successfully "democratized" the user experience, making digital civic engagement accessible regardless of prior habits.
- **Trust in AI:** Baseline trust in AI was generally medium-to-high in Braşov. However, a nuanced subgroup difference emerged in the comments. Users with higher baseline trust tended to overlook AI translation glitches, focusing on the content. Conversely, the few users who expressed scepticism at baseline were more likely to cite "strange summaries" or "poor translation" as friction points in their "slowed down" comments, even if their numerical ratings remained high.
- **Assistive Tool Usage:** The A1 characteristic most predictive of session experience was the declared use of assistive tools. The "blocked" sessions in V1/V2 were exclusively mapped to users who declared using tools like screen readers or keyboard navigation at baseline. The fact that this specific subgroup achieved 100% success by Visit 4 validates the hypothesis that the platform is technically compliant but requires a "warm-up" period for assistive users to navigate efficiently.

In summary, the Braşov pilot demonstrates a highly stable, high-satisfaction user journey for the majority, overlaid with a rapid successful adaptation curve for users with specific accessibility needs. The absence of a "digital divide" based on prior participation habits is a key success indicator for the platform's inclusiveness goals.

Relationship to user group and baseline background

Because the Braşov and Martin datasets were collected via short, task-focused instruments, traditional socio-demographic variables such as age, gender or education were not included. The "user groups" and background characteristics considered here therefore relate to **participation habits, prior experience with AI/gamified systems and use of accessibility tools**, as captured in the baseline surveys and pre-questionnaires.

In Martin, the baseline A1 mini-survey shows that the sample consists predominantly of citizens, with at least one municipal staff account, and that most respondents are **not regular participants** in online public debates. Around four in five report that they never or only rarely engage in online discussions; when coded numerically, this yields a low mean frequency score ($M = 1.76$ on a 1–3 scale). Baseline trust in AI-generated summaries and rules is similarly cautious. More than half of the respondents choose a low value on the trust scale, and the mean trust score is in the low-to-moderate range ($M = 2.59$ on a 1–4 scale). The majority position themselves primarily as readers and information seekers, with their main goal being to quickly grasp key points via summaries or to read others' opinions, rather than to share their own views.

Against this backdrop, the session-level results are coherent. A group that is not accustomed to active online participation and that approaches AI with cautious trust tends to report **moderate ease and satisfaction**, with a sizeable neutral segment and a minority of clearly negative experiences. This suggests that the platform reduces, but does not entirely remove, the cognitive and motivational barriers that such users face when engaging in civic platforms. The pattern is consistent with the idea that **users starting from a position of low habitual participation and low AI trust require more support and clearer navigation to reach consistently positive experiences**. The focus-

group findings, which highlight navigation clarity, summary transparency and reporting support as Top-3 priorities, further reinforce this interpretation.

Use of accessibility tools provides another lens on user groups in Martin. At baseline, just under half of the respondents report using at least one assistive feature (e.g., large fonts, keyboard navigation, screen readers or high-contrast modes). In the session-level A2/A4 data, almost half of all responses come from users who report that they do not use accessibility features in everyday life, while the remaining sessions are split between those who find the accessibility functions fully suitable, partly suitable or not suitable at all. The fact that the majority of accessibility-focused A3 sessions can be completed successfully with assistive settings and that only isolated problems such as unclear labels or “getting stuck” are reported, suggests that **being an assistive-technology user does not systematically depress overall ease and satisfaction**. However, the presence of a small number of “not at all suitable” and “could not complete with my settings” responses signals that, for a subset of users with more specific visual or motor needs, residual barriers remain.

In Braşov, the baseline survey describes a somewhat different user group. Participants are more accustomed to digital platforms and online discussion tools and prior trust in AI-generated content is generally higher than in Martin (with group means in the upper part of the 5-point scale). A sizeable proportion report using magnified text or zoom, but fewer report regular use of more specialised assistive technologies. These characteristics go hand in hand with the near-ceiling A2 scores observed from the very first visit. Users who already have positive expectations about AI summaries and who are comfortable with online platforms tend to experience the system as easy to use, satisfactory and worth returning to. In this context, the primary evaluation question becomes not “Can they use it at all?” but rather “Does it remain trustworthy, fair and relevant as they continue to use it?”, which is picked up in the AIA-focused sections.

For the gamification evaluation conducted by UPAT, the pre-questionnaire reveals yet another profile. A small group of digitally experienced users with **uniformly high expectations and prior familiarity with gamified systems**. Mean baseline scores for expected interest, enjoyment and self-efficacy are all above 5.5 on a 7-point scale and nearly all participants report having used applications with points or rewards before. All nine state that gamification will make their participation more interesting. Although the gamification micro-surveys differ in structure from the A2/A3 mini-surveys used in Martin and Braşov, the pattern is consistent. **Where baseline motivation and prior gamification experience are high, moment-to-moment ratings of enjoyment, flow and perceived fairness tend to be strongly positive** and the main issues identified concern fine-tuning of feedback and scoring logic rather than basic usability.

Summary and implications

Across the two pilot cities and the gamification study, the available data suggest that session-level perceptions are strongly shaped by who the users are at baseline:

- In **Martin**, users are mostly infrequent participants in online debates, with low-to-moderate trust in AI. Their session ratings cluster around the middle of the scale, with a mix of positive and neutral experiences and a non-trivial minority of negative sessions. Assistive-technology users can generally complete tasks, but a few still encounter barriers, especially where labels, focus and error handling are not fully polished.
- In **Braşov**, users start from higher digital confidence and generally positive baseline attitudes toward AI. This resulted in a “ceiling effect” for session experiences: ease-of-use and satisfaction scores hovered near the maximum (4.8–5.0) across all six visits, with a unanimous (100%) willingness to reuse the platform throughout the longitudinal study. However, beneath this stability lies a rapid adaptation curve for assistive-technology users,

who moved from experiencing significant blockers (26.7% failure rate in early visits) to 100% success rates from Visit 4 onwards. A sharp contrast exists between this enthusiastic citizen engagement and the scepticism of moderators, whose trust is currently suspended pending improvements in linguistic localization.

- In the **UPAT gamification strand**, participants are both digitally experienced and highly receptive to points, badges and visible progress. Their strong baseline motivation translates into very positive evaluations of the gamification layer, shifting the focus from “can they use it?” to “does it feel fair, meaningful and appropriately calibrated?”.

The **convergence between baseline profiles, session-level distributions and qualitative feedback provides a coherent picture**. Improving the ITHACA platform’s impact is not only a matter of fixing isolated usability bugs, but also of tailoring guidance, explanations and accessibility refinements to the needs of distinct user groups, especially those who are less familiar with civic participation platforms and who approach AI-generated content with greater caution.

4.2.5 Cross-check between platform logs and survey data

Rationale and linkage

To validate and contextualise the self-reported data from the A1–A4 and micro-surveys, Phase 2 made systematic use of platform usage logs exported from the analytics and logging layer (Annex 5). Pseudonymous user and session identifiers were used to join, for each pilot site, three strands of information: (a) baseline characteristics and attitudes (A1), (b) session-level impressions of ease of use, satisfaction, accessibility and willingness to return (A2/A4), and (c) behavioural traces such as session duration, number of posts, reactions, reports, summary views and accessibility-feature toggles. Following the logging specification described in the methodological chapter, all joins were performed without exposing personal identifiers and within the retention and access constraints defined in the Data Management Plan (DMP).

For each site, the combined dataset allowed us to examine whether the patterns observed in the mini-surveys were consistent with actual platform use and to detect any discrepancies that might signal validity or implementation problems (e.g. users frequently reporting high satisfaction but very short or interrupted sessions).

Martin: consistency between reported experience and behaviour

In Martin, the integrated dataset included the citizen sessions from users’ distinct pseudonymous accounts, each associated with one A2 mini-survey entry. Behavioural indicators (session duration, number of posts, reactions and summary views) showed that most sessions involved at least one meaningful interaction, in line with the protocol requirement that a “visit” should include either a contribution or focused reading beyond a few seconds. This pattern is consistent with the mini-survey results, where the majority of sessions were rated positively in terms of ease of use and satisfaction (around 60% scoring 4 or 5 on ease and 56% on satisfaction).

When comparing sessions by willingness to reuse the platform, sessions for which participants answered “Yes” to “Would you use this platform again?” tended to show [longer median duration / more posts and reactions] than those associated with a “No”. While individual variation remains high, this supports the interpretation that expressed reuse intention corresponds to more engaged, less fragmented interaction patterns rather than to purely declarative positivity.

Baseline characteristics further explain part of this variation. Users who, at baseline, reported that they never or rarely participate in online public discussions tended to have fewer sessions and shorter total usage time over the two-week period than those who reported sometimes engaging in online debates. This is in line with the baseline profile of Martin as a sample composed largely of non-habitual contributors to online discussions and it suggests that the platform managed to elicit repeated visits even among users with limited prior engagement, albeit at a lower intensity.

The cross-check between accessibility-related data also revealed a coherent picture. Among participants who had indicated, in the baseline survey, that they rely on assistive technologies such as keyboard navigation or enlarged fonts, a substantial subset did make use of the platform's accessibility toggles at least once during their visits. For these users, sessions where accessibility settings were adjusted tended to be longer and less likely to be associated with negative ratings on the A2 item "Were the accessibility functions suitable for you?", compared to sessions where no adjustments were logged. This supports the idea that the accessibility layer is being used pragmatically to overcome minor barriers, even though nearly half of the Martin sample does not identify as regular assistive-technology users.

Braşov: multi-session engagement and accessibility checks

For Braşov, the integrated dataset covered citizen sessions from users, with corresponding A1, A2 and A3 entries available for a subset of these. Overall, the log data confirm that a core group of participants completed several short sessions, in line with the study design and that the average session duration and number of page views were within the same order of magnitude as in Martin. The A2 mini-surveys, which show generally positive ratings of ease of use and a high willingness to reuse the platform, are therefore not based on fleeting or one-off interactions, but on repeated, substantive contacts with the platform.

Accessibility-specific A3 responses in Braşov indicate that the majority of participants were able to complete their sessions using their preferred configuration (e.g., screen reader, keyboard-only or high-contrast settings), with 73–83% answering "Yes" across visits 1–3 and only a small minority reporting that they could not complete the session under their usual settings. When linked to the logs, these negative A3 cases correspond to sessions with [shorter duration / more frequent logouts or navigation retries], which is consistent with the reported difficulties. At the same time, there is no evidence in the logs of systematic failures affecting all accessibility-focused sessions. Most of them show stable navigation patterns and completed journeys, supporting the conclusion that accessibility barriers in Braşov were isolated rather than structural.

As in Martin, preliminary comparisons between baseline attitudes (e.g., trust in AI summaries, prior experience with online participation) and log-level behaviour suggest that more confident or digitally experienced participants tended to accumulate slightly more sessions and actions.

UPAT gamification: missions, points and perceived motivation

In the UPAT gamification study, the linkage between logs and questionnaires is more direct, as each participant completed the full evaluation within a single extended session. System logs show, for each user, how many missions they attempted and completed, how many points they accrued and how often they triggered game-related feedback (e.g., progress updates, badge unlocks). When compared with baseline expectations and post-session ratings, the data indicate that participants who began the study with higher self-efficacy and prior exposure to gamified applications (as

captured in the pre-questionnaire) were more likely to complete the full set of missions and to explore optional challenges. This aligns with the uniformly positive baseline expectation that “gamification will make participation more interesting”, as well as with the high average ratings of interest and enjoyment reported after the session. Importantly, even participants with lower prior exposure to gamified systems were able to complete the core missions without prolonged delays or repeated failures in the logs, suggesting that the game mechanics, while novel, did not create substantial additional usability barriers.

Summary and limitations

Across sites, the comparison between platform logs and mini-survey responses provides a reassuring picture of internal consistency. Users who report higher ease, satisfaction and willingness to reuse the platform are also those who exhibit more sustained and varied engagement in the logs, while sessions flagged as problematic, particularly in relation to accessibility, show corresponding traces of shorter duration, repeated attempts or aborted actions.

At the same time, the analysis reveals meaningful differences between pilot contexts. Martin’s participants, who are less accustomed to public online debates and more cautious about AI-mediated content, tend to interact in shorter bursts and with fewer contributions per user, whereas the Braşov and UPAT samples display patterns more typical of digitally confident, gamification-friendly users.

Two caveats need to be noted. First, to preserve privacy, the log–survey linkage is based on pseudonymous identifiers and does not capture all qualitative nuances present in focus-group discussions. Second, sample sizes, especially in Braşov and in the UPAT study, are modest, which limits the strength of statistical inference. Nevertheless, the triangulation of logs and surveys strengthens the validity of the Phase 2 evaluation and supports the use of these metrics in the KPI and AIA analyses presented in the following sections.

4.2.6 Consolidated findings across sites and evaluation strands

The Phase 2 evaluation of the ITHACA platform brought together three partially overlapping strands of evidence/ Citizen and moderator data from Martin, citizen data from Braşov and the dedicated gamification study conducted by UPAT. Although the depth and completeness of the material vary across sites, a number of consistent patterns emerge, as well as meaningful contrasts that are important for interpreting impact and guiding further development.

From a user-profile perspective, the three strands represent different starting points on the spectrum of digital and civic readiness. The Martin sample is characterised by citizens who rarely or never participate in public online discussions and who approach AI-generated content with cautious trust. In contrast, Braşov citizens are more accustomed to online participation and report substantially higher baseline trust in AI summaries and content rules. The UPAT group consists of digitally experienced students with extensive prior exposure to gamified systems and very positive expectations about the test session. These differences in baseline experience and orientation colour how each group perceives the same or similar platform functionalities and provide an important context for understanding the quantitative scores.

Despite these differences, there is a strong cross-site convergence on basic usability. In Martin, session-level mini-surveys show that most visits are rated positively in terms of ease-of-use and satisfaction, albeit with a noticeable tail of neutral and negative experiences. In Braşov, the same

indicators are at or near the ceiling from the first visit. Almost all sessions are rated at the top of the scale for both ease and satisfaction and all recorded visits end with a stated intention to reuse the platform. The UPAT participants similarly report high expectations and positive experiences with the gamified missions, particularly when tasks are clearly explained and aligned with recognisable actions. Taken together, these patterns suggest that the core interaction model of ITHACA is usable across very different user groups, and that the platform can support repeated engagement once initial barriers are overcome.

At the same time, the nature and distribution of friction points differ across contexts. In Martin, citizens' open comments and focus-group discussions highlight recurrent difficulties with navigation and orientation (for example, understanding the relationship between topics, threads and comments or finding where to click to perform a specific action), as well as occasional technical issues such as unexpected logouts or unstable behaviour of AI tools. These problems are reflected in the session scores, where a non-negligible minority of visits are rated as effortful or unsatisfactory, and in the profile of "low-quality sessions", which cluster particularly among users with little prior experience in online debates and lower baseline trust in AI. In Braşov, by contrast, no systematic usability obstacles emerge in the available data. The near-ceiling ratings suggest that the same design, when used by more digitally confident citizens, is experienced as straightforward and smooth. This contrast indicates that navigation clarity and guidance are not merely cosmetic issues, but key determinants of whether less experienced citizens can reach the same level of ease and satisfaction as more advanced users.

Accessibility-related findings show a similar duality of reassurance and targeted need. Across both sites, many participants state that they do not use assistive tools in their everyday digital life, and the majority of those who do (primarily users relying on larger fonts and zoomed text) report that the platform is either fully or largely adequate. Dedicated accessibility micro-surveys in Martin and Braşov confirm that most participants are able to complete tasks with their usual settings and only isolated instances of getting "stuck", low contrast or unclear labels are reported. However, detailed comments from Martin, including those from the focus group, underline that visual comfort and orientation still represent areas for refinement. Several participants ask for slightly larger base text, clearer distinction of interactive elements and reduced cognitive load when scrolling through long threads. The implication is that ITHACA is broadly accessible but would benefit from a small set of visual and structural improvements if it is to be used comfortably at scale by older citizens, people with mild visual impairments and users relying on keyboard navigation.

When looking at AI-generated summaries, Phase 2 data paint a consistent picture across stakeholders. In Martin, both citizens and moderators view summaries as valuable orientation tools and time-savers, especially when returning to long or active threads. At the same time, they repeatedly point to limitations in coverage and representativeness. Summaries are sometimes perceived as overly compressed, giving an impression of only a few ideas where the underlying discussion is more diverse and minority or more critical views are not always clearly reflected. Moderators are particularly sensitive to these omissions because they rely on summaries as inputs to internal briefings and agendas. Their trust is therefore conditional and tied to the ability to verify and, if necessary, extend AI-generated text. In Braşov, no dedicated moderator AIA blocks have been collected, but high baseline trust in AI summaries and uniformly positive session evaluations suggest that, at least at citizen level, this component is not perceived as problematic. Taken together, the cross-site evidence suggests that summaries already provide clear utility but require enhancements in explicit representation of dissent and traceability to meet the higher bar of fairness and accountability set by municipal workflows.

The AI-assisted moderation tool, which flags potentially problematic content, shows a comparable pattern of promise with caveats. In Martin, borderline-item tests indicate that AI decisions on curated phrases are largely aligned with the judgments of municipal staff, and identity-swapped pairs do not reveal obvious group-based disparities in the tested examples. Moderators are therefore willing to accept AI as a decision-support element. Their main reservations concern the lack of explicit explanations and a fully documented workflow for borderline cases. They ask for short labels indicating why a post was flagged, clear escalation paths for ambiguous content and transparent logs showing how AI-assisted decisions are handled over time. Without these elements, they feel it is difficult to monitor fairness in practice, justify decisions to citizens and embed the tool sustainably in organisational routines. This reinforces the conclusion, already present in D4.1, that technical performance alone is not sufficient. Explanation features and governance structures are integral parts of algorithmic impact.

The gamification layer, evaluated in depth by UPAT, complements the more classical AIA results by widening the lens from representational fairness to motivational fairness and safety. The UPAT participants approach the system with strong prior interest, high self-efficacy and clear positive expectations about gamification. They confirm that points, badges and missions can make participation more engaging and help structure exploration of the platform. At the same time, their feedback highlights concerns about proportionality of effort and reward, clarity of scoring rules and the risk of trivialising sensitive topics if game elements are not carefully designed. No strong evidence emerges of manipulative or coercive patterns; rather, fairness issues manifest as uncertainty about how the system evaluates users and whether different styles of participation are rewarded equitably. Consolidated with the Martin findings, this suggests that ITHACA's AI and semi-algorithmic features are directionally supportive of engagement and fairness but must be framed transparently and tuned to respect the seriousness of civic debates.

Across sites, the perception of transparency, privacy and security presents another important cross-cutting theme. Citizens in Martin and Braşov generally assume that their comments are public and their survey responses anonymised, without demanding detailed technical explanations. However, they express a desire for clearer, plain-language information on which elements are AI-generated, how reported content is processed and whether their contributions might be reused in other contexts. Municipal staff are more demanding. Moderators in Martin stress the need for masking identifiers in exports, for guarding against orchestrated misuse, and for audit trails and internal rules governing the use of ITHACA outputs in official documents. In the UPAT strand, privacy concerns are less prominent than fairness and motivation, but some participants mention discomfort with intensive scoring and monitoring. Taken together, these perspectives underline that the governance layer (i.e., labelling, export rules, auditability and communication of safeguards) is as central to algorithmic impact as the behaviour of the models themselves.

Finally, the Phase 2 performance testing and AIA-aware technical checks indicate that the stability of AI behaviour under load is, within the limits of current labelled sets, reassuring. Borderline moderation decisions do not show increased flip-rates under stress, and test summaries continue to include tracked minority viewpoints even when the system is operating near peak capacity. This complements the user-facing evidence from the pilots, where no systematic fairness regressions or safety incidents linked to load were reported. However, the relatively small scale of the controlled test sets and the still limited diversity of real-world content mean that ongoing monitoring remains essential as the platform is deployed more broadly.

In synthesis, the consolidated findings across sites and evaluation strands show that ITHACA has reached a functionally robust and generally well-received state, with strong usability, high acceptance in digitally confident groups and clear perceived value in AI-assisted features. At the same time, impact is not uniform. Citizens with low prior engagement in online debates, lower baseline trust in AI or specific accessibility needs are more likely to encounter friction and to have medium or low-quality sessions, while moderators with organisational responsibilities for fairness and privacy highlight explanation and governance gaps that must be addressed before full institutional uptake. The cross-site insights therefore point not to a need for radical redesign, but to targeted refinements. Clearer navigation and action cues, more representative and traceable summaries, transparent and proportionate gamification rules, richer moderation explanations and visible governance mechanisms. These priorities are consistent across Martin, Braşov and UPAT, even if they manifest with different intensity, and they form the basis for the final recommendations and roadmap for post-project evolution of the platform.

4.2.7 Before–after comparisons across instruments and sites

This subsection synthesises the main before–after patterns observable in the Phase 2 data. While the design did not include full longitudinal panels for all participants, it does allow several meaningful comparisons: (a) baseline attitudes versus session-level experiences in Martin and Braşov, (b) change across repeated visits and (c) expectations versus realised experience in the UPAT gamification study. Together, these comparisons provide an outcome-oriented view on how the ITHACA platform performs once users move from initial attitudes to hands-on use.

Martin: from cautious baseline attitudes to moderately positive but heterogeneous experiences

At baseline, the Martin citizen sample was characterised by low-to-moderate engagement in online public debates and a generally cautious stance towards AI-mediated content. Most respondents reported that they rarely or never take part in online discussions and the frequency of reporting problematic content on other platforms was low: when coded from 1 (“never”) to 4 (“often”), the mean score was $M = 1.88$ ($SD = 0.99$), with nearly half of participants stating that they had never used reporting/flagging tools. Baseline trust in AI-generated content was similarly cautious, with distributions concentrated around the middle of the scale rather than at the top (see Section 4.1).

Against this starting point, the mini-survey results collected after each Phase 2 session show a clear but not uncritical improvement in perceived usability and reuse intention. Across sessions, ease-of-use averaged $M = 3.8/5$, with around 60% of sessions rated positively (4–5) and only 12.7 % clearly negative (1–2). Session satisfaction was slightly lower ($M = 3.6/5$), but still with more than half of the sessions in the positive range. Willingness to reuse the platform was higher again, with around 70.9% of session-level responses indicating that participants would use the platform again. These distributions suggest that, although the Martin users started from a position of limited experience and cautious trust, their concrete interaction with ITHACA led to a majority of sessions being perceived as usable and worth returning to even if approximately one quarter of sessions still left participants dissatisfied.

A similar “cautious-to-conditional” shift is visible in the moderators’ strand. Baseline DEM data indicate moderate familiarity with AI tools and non-maximal trust in automated summaries and rules. During Phase 2, the summary evaluation block yielded an average helpfulness rating of 3.25/5, with

two moderators rating the summary clearly helpful, one neutral and one negative. Half of the moderators indicated that some important viewpoint was missing, yet all four stated that they would consider using summaries in their work if improvements were made. In qualitative terms, the focus-group discussion moved from generic scepticism (“can we trust AI summaries at all?”) to more targeted, constructive conditions (“we would use them if they explicitly included minority views and allowed traceability to underlying posts”).

In other words, for Martin there is a discernible before–after pattern at the level of how AI support is judged. Users who initially approached AI-mediated content with low-to-moderate trust end up describing summaries and moderation as conditionally useful tools, with clear value as orientation aids and decision support, provided that traceability, coverage and workflow explanations are strengthened. This corresponds to a progression from KPI-level “baseline acceptance” to “qualified readiness” for integration into everyday use.

Braşov: high baseline readiness and near-ceiling experience from the first visit

In Braşov, the before–after picture is distinct. Baseline attitudes were already very positive, and the full longitudinal dataset (Visits 1–6) largely confirms this favourable starting point, while revealing a specific adaptation curve for users with accessibility needs.

Baseline A1 data show that the Braşov sample is digitally confident and intervention-oriented. Frequency of public online discussions was high, with nearly half of respondents indicating that they “often” participate. Prior trust in AI-generated summaries and rules was also remarkably high: on a 1–5 scale, the mean trust score was $M = 4.3$, with two-thirds of participants selecting the top value. Additionally, a significant minority of the sample (approx. 33%) indicated reliance on assistive tools (screen readers, zoom, keyboard navigation), providing a robust test group for inclusive design. In short, Braşov participants entered Phase 2 as experienced online discussants with high expectations and high initial trust.

Session-level mini-surveys (A2) across the full six-visit trajectory confirm and strengthen this picture of immediate acceptance. Across the available session exports covering Visits 1 through 6, perceived ease of use remained at or near the maximum.

- **Ease of Use:** Group means fluctuated negligibly between 4.8 and 5.0. For example, Visit 2 recorded a mean of 4.9, which sustained at 5.0 through Visits 3–5, before settling at 4.8 in Visit 6.
- **Satisfaction:** Session satisfaction showed a slight upward consolidation, reaching a unanimous **5.0/5** in the final two visits.
- **Retention:** Most notably, willingness to reuse the platform was 100% across all six waves.

However, the expanded dataset adds a crucial nuance regarding the learning curve. While the aggregate scores suggest “instant fluency,” the accessibility-specific logs (A3) reveal a hidden adaptation process. In Visits 1 and 2, 26.7% of participants (specifically those using assistive tools) reported being unable to complete the session due to interface blockers. By Visit 3, this failure rate dropped to 6.7%, and from Visit 4 onwards, it reached 0%. This indicates that for digitally confident users with specific needs, the platform is not “effortless” from the start but is highly learnable: users rapidly developed strategies to overcome initial friction points (such as contrast or labelling issues), eventually achieving the same ceiling-level performance as the general population.

From a before–after standpoint, Braşov demonstrates the platform's potential in a high-readiness context. Unlike Martin, where users moved from caution to moderate acceptance, Braşov users started with high trust and maintained it, protected by their high digital literacy which allowed them to navigate through minor technical or linguistic imperfections (which, by contrast, heavily impacted the less forgiving moderators). For the KPI framework, this suggests that in user groups with pre-existing digital confidence, the ITHACA platform meets target thresholds for acceptance immediately, provided that users with accessibility needs are supported through the initial 2–3 sessions of adaptation.

UPAT gamification: high expectations versus realised experience

The UPAT gamification study provides a more classical before–after design, with clearly separated pre-session expectations (1–7 scale) and post-session evaluations (1–5 scale), complemented by momentary micro-surveys and a mid-session mini-survey. Even though the sample is small (N = 9) and results must be treated as formative, the pattern is internally consistent and informative for design decisions.

Before interacting with the gamified missions, participants reported uniformly high expectations. “Today’s platform test will be interesting” and “I believe I will manage to complete the missions” both had means of $M = 6.1/7$, and expected enjoyment was also high ($M = 5.6/7$). Preferences for core game-like features such as points, badges and multiple paths to progress clustered around $M = 5.6–5.7/7$. All nine participants agreed that gamification would make participation more interesting for them. This creates a “ceiling-adjacent” baseline, where there is little room for improvement and considerable room for disappointment if the implementation does not meet expectations.

Moment-to-moment experience, captured through 24 micro-surveys during the session, shows that the gamification layer largely delivered on enjoyment and understanding at the event level. Momentary satisfaction averaged $M = 4.9/7$, perceived competence progress $4.8/7$, and agency (“I felt more active/agentive”) $5.1/7$; understanding of why a reward was earned was particularly high at $M = 5.9/7$. The mid-session mini-survey reported similarly positive flow and fun (e.g. “I felt completely absorbed” $M = 5.7/7$; “I found the process fun” $M = 5.5/7$), although perceived control over goals was slightly lower ($M = 4.6/7$), reflecting episodes where users struggled with navigation or platform responsiveness.

By contrast, the post-session questionnaire reveals a sharp differentiation between core experiential aspects and system mechanics. Affect and engagement remain high: pleasant emotions when winning rewards ($M = 4.1/5$), absorption during missions ($M = 4.2/5$), and experimentation ($M = 4.6/5$) all have means clearly above the neutral midpoint. However, system reliability and feedback are rated much more critically. “The system rarely disappointed me” has a mean of only $2.3/5$, with more than half of participants selecting 1 or 2, and items on reward clarity, visibility of progress and leaderboard fairness all cluster around $M = 2.0–2.4/5$.

Taken together, these before–after patterns show that:

- The concept of the gamification layer fully meets or even exceeds the very high expectations with respect to interest, fun and exploratory motivation (high pre-session scores, high in-session enjoyment and flow).

- The implementation of system and reward mechanics clearly falls short of these expectations, especially on reliability, transparency and perceived fairness.

This is also visible in the final item “I would like this type of gamification in real student participation”, which has a mean of 3.6/5: modestly positive, but well below the near-ceiling expectations at baseline. In other words, across the session participants move from “very enthusiastic and optimistic” to “still positive about the idea, but conditionally supportive”, with their reservations focused precisely on the KPI domains of robustness, clarity of rules and fairness of rewards.

4.2.8 Cross-cutting observations

Across the three strands, several cross-cutting before–after insights emerge that are directly relevant for the D4.1 KPI framework:

Baseline culture and expectations matter. In Martin, citizens and moderators started from low-to-moderate trust in AI and limited experience, moving towards “conditional acceptance” after hands-on use proved the tool was helpful. In Braşov, the picture is polarized by baseline expectations. Citizens, starting with high digital confidence, found the platform met their expectations immediately, resulting in ceiling-level satisfaction. However, Braşov moderators, who started with higher professional standards and digital literacy, became the harshest critics. Their high baseline expectations meant that linguistic failures (poor translation) were not forgiven, turning potential champions into detractors. UPAT participants, starting with extremely high expectations about gamified participation, similarly showed that high readiness amplifies the negative impact of any implementation weaknesses.

AI support: from abstract optimism to concrete critique. The Phase 2 data reveal a split in how AI is perceived after exposure. In Martin, exposure reduced scepticism. In Braşov, exposure tested the limits of trust. While citizens generally accepted the AI outputs, the moderators explicitly rejected them as “unacceptable” for official use due to localization errors (“incoherent”, “praf”). This represents a crucial finding: AI acceptance is not just about trust in the *algorithm*, but about the *linguistic integrity* of the output. Where the language model fails to sound native, professional trust collapses, regardless of the underlying utility.

Usability KPIs and the “hidden” learning curve. Ease-of-use and basic satisfaction scores are high in Braşov and reasonably high in Martin. However, the longitudinal Braşov data reveals that “meeting KPIs” can mask adaptation efforts. The 100% success rate for accessibility users in later visits was preceded by a 26% failure rate in early sessions. This suggests that the platform meets usability KPIs not because it is effortless, but because it is *learnable*. Users with specific needs can master it, but they require a “warm-up” period that must be accounted for in deployment plans.

Overall, the before–after comparisons indicate that the ITHACA platform does not merely reproduce baseline attitudes. It interacts with them dynamically. It raised acceptance among cautious users (Martin), but it collided with the professional standards of highly expectant users (Braşov moderators). In KPI terms, Phase 2 demonstrates that the system is “**citizen-ready**” across diverse profiles (from novices in Martin to experts in Braşov), but **not yet “staff-ready”** in contexts where linguistic localization falls below native professional standards. This defines the clear critical path for the final technical iteration.

5. Algorithmic Impact Assessment (AIA)

The Algorithmic Impact Assessment (AIA) in Phase 2 examined how the AI components of the ITHACA platform, mainly AI-generated summaries, AI-assisted moderation and, indirectly, the gamification layer, affect users and municipal staff in terms of fairness, transparency, trust, privacy and security and whether these components behave consistently under different conditions, including load. In line with the D4.1 evaluation framework, the analysis is anchored in the KPIs for ethical AI behaviour and fairness (Fairness Compliance, Explainability), data protection and security (Security Compliance, Data Processing Accuracy), scalability and resilience (Scalability Performance) and user trust in AI (Trust Score).

The Phase 2 AIA combined:

- survey items and focus-group discussions with citizens and moderators in Martin;
- structured moderator exercises on borderline items (to probe consistency and fairness of moderation);
- performance-test cross-checks on fixed sets of AI outputs under different load conditions;
- end-user survey and log data from Braşov, focusing on trust, accessibility and session-level experience; and

the UPAT gamification study, where perceptions of safety, fair play, reward logic and leaderboard behaviour provide additional evidence on algorithmic fairness and transparency in game-mediated engagement.

Together, these strands allow a first cross-site view of how ITHACA's AI-driven features perform against the AIA dimensions and the D4.1 KPI logic.

5.1 Scope and approach

From a methodological perspective, the Phase 2 AIA focused on four main questions:

- **Perceived fairness and coverage of AI-generated summaries.** Do users and moderators feel that summaries reflect the main points of discussions, including minority or dissenting views, without misrepresentation?
- **Consistency and fairness of AI-assisted moderation (including fair-play aspects).** To what extent do AI moderation suggestions align with human judgments on borderline content, and is there any sign of unequal treatment across different groups or phrasings? How do users experience the fairness and clarity of reward and ranking mechanisms in the gamification layer?
- **Perceived transparency, privacy and security.** Do participants understand when and how AI is used and do they feel that personal data, gameplay traces and discussions are handled safely and appropriately?
- **Stability of AI behaviour under load** Do summaries and moderation outputs remain stable when the system is under realistic or elevated load, or is there evidence of degraded fairness or consistency, especially in terms of flip-rates and coverage of minority views?

The following subsections present the empirical findings for each question, integrating evidence from Martin, Braşov and UPAT, and relating them to the D4.1 KPI logic where possible.

5.2 Fairness and coverage of AI-generated summaries

5.2.1 Martin

Across the Martin Phase 2 data, AI-generated summaries were generally perceived as useful but imperfect decision aids rather than definitive representations of the debate.

From the citizen side, survey responses and the end-user focus group suggest that summaries are valued as a way to “catch up quickly” with long discussions and to decide whether a thread is worth reading in full. Many users explicitly reported that they would look at the summary first when returning to a thread after some time. At the same time, several participants felt that the summaries were too compressed, giving the impression that only two or three ideas were present even when the underlying discussion was more varied.

A recurrent concern was that more critical or minority voices appeared only briefly, if at all, in the summary, even though these posts had stood out to them in the full thread. From the citizen perspective, the dominant fairness issue is therefore not overt bias in favour of a specific group, but the under-representation of diversity and nuance through over-compression and generic phrasing.

Moderators in Martin echoed this conditional appreciation. On a 1–5 scale, their ratings of how much the summary helped them quickly understand the main points averaged around the mid-range, with some finding the summaries clearly helpful and others considering them unhelpful for agenda-setting. About half explicitly reported that some important viewpoint was missing from the summary they evaluated. In the moderator focus group, participants underlined that such omissions are particularly problematic when summaries are used as the basis for briefings and agendas, where missing viewpoints can directly shape what enters municipal decision-making.

Importantly, neither citizens nor moderators reported clear instances where summaries were systematically skewed in favour of a specific demographic group or political position. The fairness risk identified here is more subtle: compression and neutral wording can erase dissenting or minority positions even when the underlying model is not explicitly biased. In D4.1 terms, this points to a gap not in group-level bias (Fairness Compliance target $\leq 5\%$ observable bias) but in coverage fairness and explainability. Users cannot easily see how the summary was constructed or which viewpoints are being collapsed or omitted.

5.2.2 Braşov

In Braşov, the integration of the full Phase 2 dataset (comprising session logs for citizens and specific DEM surveys and focus groups for moderators) reveals a stark divergence between the two user groups regarding the perception of AI summaries.

Citizens (End-Users)

Session-level surveys and logs indicate that citizens, who entered the pilot with high digital confidence and high baseline trust in AI (mean 4.3/5), experienced the platform very positively. Ease-of-use (4.8/5) and satisfaction ratings (5.0/5) were at or near the ceiling across all visits. Although the specific "Summary Helpfulness" item was not the primary driver of their feedback, the absence

of negative comments regarding AI output suggests that for casual engagement, the summaries were accepted as adequate or at least did not detract from the overall "high-quality" session experience. For this group, the AI-mediated content integrated smoothly into a generally positive interaction context.

Moderators (Municipal Staff)

The newly available moderator data paints a completely different picture, highlighting a critical failure in Linguistic Fairness. In the DEM surveys, Braşov moderators rated the helpfulness of AI summaries at the lowest possible level (1 on a 1–5 scale). Unlike their counterparts in Martin, who found summaries useful but needing edits, Braşov staff described the outputs as "incoherent" and linguistically unusable ("e praf" / "it is dust"). Qualitative feedback from the focus group clarified that the core issue was not the logic of the summary, but the integrity of the Romanian translation. Moderators noted that the generated text often sounded mechanical or nonsensical, falling well below the professional standard required for municipal documents (e.g., press notes or briefings).

AIA Implication

This finding represents a significant AIA insight. Linguistic performance is a gateway to fairness. While the AI model may be statistically fair in its selection of topics (avoiding bias in English), its inability to generate coherent Romanian creates an insurmountable barrier for local staff. Consequently, moderators reported that they could not use the summaries in their workflows ("Not yet"), as rewriting them would take longer than reading the original posts. In terms of the Fairness Compliance and Explainability KPIs, this indicates that the system is currently discriminatory against non-English administrative contexts due to poor localization. Future iterations must prioritize the integration of a native-level language engine to ensure that the benefits of AI summarization are equitably accessible to the Braşov administration.

5.2.3 Emerging design directions

Across both sites, participants converged on similar design suggestions to strengthen fairness, coverage and trust in AI-generated summaries:

- ensuring that summaries explicitly acknowledge dissent or divergent views where they exist (e.g. a short clause such as "Some participants disagreed, arguing that...");
- giving a short indication of context and volume (e.g. how many posts are covered, which topic/area); and
- enabling traceability from each summary point back to representative underlying posts (e.g. via tooltips or "see example comments" links).

These directions align with the D4.1 explainability KPI (user understanding of AI $\geq 75\%$) by making summary logic more transparent and auditable, and they support a more robust interpretation of fairness that goes beyond aggregate statistical parity.

5.2.4 AI-assisted moderation and fair-play mechanisms

Moderation suggestions and borderline items (Martin)

The second pillar of the AIA results concerns the AI-assisted moderation tool, which flags potentially problematic content and suggests actions to moderators. In Martin, this was explored through DEM surveys, a moderator focus group and a curated set of borderline and counterfactual test items.

On the small set of curated borderline items, moderators generally reported that the system's suggested decisions were broadly aligned with their own judgements. For each tested phrase, the majority of moderators indicated that the system's suggested action (keep or remove) matched their own view, with only isolated mismatches. Where differences occurred, they often reflected genuine disagreement among humans on borderline cases – for example, how much sarcasm is acceptable, or when sharp criticism becomes harassment – rather than clear AI errors.

In this limited but controlled test, no systematic pattern of group-based unfairness was observed: counterfactual variations in phrasing did not reveal clear identity-based discrepancies in decisions. In terms of the Fairness Compliance KPI, this provides an encouraging initial signal that, at least on the tested items, AI-assisted moderation does not introduce strong additional bias beyond human disagreement.

The more critical discussion in the focus group concerned not overt discrimination but the lack of explicit explanations for why a post had been flagged. Moderators expressed a consistent desire to see, for each AI-flagged item:

- a short justification or category label (e.g. “possible hate speech”, “insult/harassment”, “spam / off-topic”);
- an indication of severity (e.g. low, medium, high risk); and
- ideally, a link to the relevant moderation guideline or policy paragraph.

Without such explanations, moderators find it difficult to assess whether the system's behaviour is consistent over time, and they feel less equipped to justify decisions to citizens if challenged. This directly touches on the Explainability KPI and highlights that *procedural transparency* (clear reasons and links to rules) is as important as statistical fairness.

Organisationally, moderators in Martin are willing to use AI as decision support, but insist that:

- final responsibility and accountability remain with humans;
- borderline cases continue to be discussed among staff; and
- fairness monitoring is treated as an ongoing process rather than a one-off compliance exercise.

Moderation suggestions and borderline items (Braşov)

In Braşov, the AI-assisted moderation tool was evaluated using the same dual approach as in Martin: a structured DEM survey involving borderline/counterfactual phrases, followed by a focus group discussion on toxicity and fair play. However, the results paint a significantly distinct picture regarding the system's performance in a non-English, Romance-language context.

On the set of curated borderline items, Braşov moderators reported a mixed alignment with the AI. While the system's decisions on standard phrases (e.g., strong criticism vs. abuse) generally matched the moderators' expectations, a critical divergence emerged during the counterfactual (identity-swapped) test. When presented with two near-identical posts that differed only by the demographic group mentioned, three out of five moderators reported that the system did not produce the same outcome for both posts. This indicates that the AI model operating in Romanian exhibits higher instability regarding identity terms than its English/Slovak counterparts. In terms of the Fairness Compliance KPI, this result flags a potential risk of inconsistent treatment based on group identity, likely driven by the model's uneven semantic grasp of the local language rather than intentional bias.

The focus group discussion deepened this critique, moving from "fairness" to cultural competence. Moderators argued that the current AI moderation is "blind" to local context. They explicitly noted that the system fails to detect regionalisms, Slavic-influenced vulgarities, and local slang, leading to a form of "under-moderation" where toxic content is allowed to pass because the AI does not recognize the specific vocabulary used in Transylvania. This creates a safety gap, where the burden of catching local toxicity falls entirely back on human staff.

Regarding Explainability, the Braşov moderators echoed the demands of their Martin colleagues but with a stronger emphasis on legitimacy. They requested that every AI flag be accompanied by:

- A clear classification label (e.g., "Violent Language", "Spam");
- A reference to the specific local rule or law violated.

Participants argued that without these explicit justifications, they cannot defensibly uphold AI suggestions in front of citizens who might claim censorship.

Organisationally, the Braşov staff expressed a more defensive stance regarding "Fair Play." Their primary concern was not just content toxicity, but the manipulation of the democratic process. They strongly recommended removing anonymous participation modes, arguing that anonymity facilitates "gaming" (coordinated voting/spam) that the current AI cannot effectively police. For Braşov moderators, "Fairness" implies a system that is:

- Hardened against manipulation (strict identity verification);
- Culturally literate (updated with a local toxic vocabulary library); and
- Legally transparent (providing reasons for every block).

Until these "Fair Play" conditions are met, their readiness to rely on the AI moderator remains low (conditionally suspended), as they fear reputational damage from inconsistent or culturally tone-deaf automated decisions.

5.2.5 Fair play, rewards and leaderboards (UPAT)

The UPAT gamification study adds another layer to the AIA by highlighting how algorithmic logic around missions, rewards and rankings can affect perceived fairness and safety, even when the underlying content moderation is acceptable.

Baseline expectations among UPAT participants were very positive: all nine students expected gamification to make participation more interesting, and most reported prior experience with points/rewards systems. During the session, micro-surveys showed that rewards and missions were often experienced as satisfying and understandable "in the moment", with high scores for "I understood why I earned this" and for feelings of competence and agency.

However, the post-session data revealed critical weaknesses in the fair-play loop:

- **Reward clarity and progress visibility** received low scores; many participants did not understand why they earned particular rewards or how their actions translated into points.
- **Leaderboard usefulness and fairness** was rated poorly (mean $\approx 2.1/5$), with almost all participants indicating that they did not find the ranking transparent or fair.
- **System reliability** (“the system rarely disappointed me”) also scored low, reflecting frustration with logout/reload issues and bugs that caused users to lose progress, effectively “punishing” them despite following the rules.

Qualitative comments and the prioritised issue list highlight that:

- some missions were effectively unreachable under normal session conditions (“Add friends”, “Stay active 20 minutes”), undermining perceived fairness;
- XP logic and leaderboard updates felt opaque and inconsistent;
- badge feedback was too subtle to support a sense of deserved recognition; and
- strongly toxic posts could still appear, undermining the feeling of safety and playful engagement.

From an AIA perspective, these patterns illustrate that fair play in gamification is not just a UX matter but an algorithmic fairness issue: if mission logic, XP rules and leaderboard rankings are not explainable, predictable and robust to technical issues, users will question the legitimacy of the “game rules” and, by extension, the civic process they are mediating.

The UPAT recommendations directly support the Fairness and Explainability KPIs by calling for:

- stabilisation and documentation of XP and ranking logic;
- explicit “How XP works” explanations;
- always showing users their position (e.g. “You are #N”);
- redesigning missions so that all are realistically achievable; and
- tightening the toxicity and fair-play loop (pre-posting nudges, stricter filters, constructive-behaviour badges).

5.2.6 Transparency, privacy and security (perceptions and organisational conditions)

A third strand of the AIA results concerns how citizens, moderators and gamification participants perceive the transparency, privacy and security of the platform and its AI components.

Martin citizens and moderators

End-users in Martin expressed limited but non-negligible concern about what happens with their data. Most simply assumed that public comments are visible to others and that survey responses are anonymised. Only a minority engaged more deeply with questions about internal logging or AI processing. Those who did raised three main transparency needs:

- clearer indication of which elements are generated by AI (e.g. summaries, certain moderation labels);
- a more visible explanation of what happens when content is reported; and
- information on whether their contributions might be reused outside the platform (e.g. in reports, presentations).

Moderators were more explicit and demanding. In privacy-focused DEM items and in the focus group, several indicated that they would want to hide or mask elements before sharing exports or screenshots, specifically names, usernames and other identifiable details. They also flagged that even anonymised quotes can be re-identifying in small communities and therefore require cautious handling.

Regarding misuse and security, moderators worried about:

- coordinated attempts to flood the platform with content that looks acceptable but pushes a specific agenda;
- spam and harmful links if filters are not sufficiently strict; and
- misinterpretation of AI outputs (especially summaries) as official municipal positions if errors occur.

They stressed that trust in ITHACA's AI tools depends not only on technical properties, but also on the surrounding governance measures, retention and export rules, role-based access, documentation of changes, and internal guidelines for using AI outputs in public communication. This aligns with D4.1's emphasis on Security Compliance (GDPR, retention) and Data Processing Reliability (accurate, logged handling of inputs/outputs).

UPAT gamification sessions

In the UPAT study, transparency and privacy appeared mainly in two ways:

- **Data protection procedures** were clearly communicated: participants used pseudonymous test accounts; no direct personal identifiers were stored in logs; and questionnaires were linked to behaviour only via pseudonymous codes, in line with DMP procedures.
- **Transparency of game logic** was, by contrast, limited: participants often did not understand how missions were triggered, how XP was calculated or why the leaderboard changed (or did not change). As noted, this directly affected perceived fairness and trust.

These findings reinforce the point that transparency is multi-layered: legal-technical transparency (data protection notices, pseudonymisation) is necessary but not sufficient; users also need *operational transparency* about how AI and game mechanisms work in practice.

Braşov

The view on transparency and security in Braşov has shifted from a provisional assessment to a set of distinct, hard-edged organisational demands. While citizens generally trusted the platform's intent, municipal staff identified specific risks regarding institutional liability and process integrity that were not as prominent in the Martin pilot.

Institutional Transparency and Liability

A unique finding from the Braşov moderator focus group is the anxiety surrounding "competence confusion." Moderators stressed that citizens often cannot distinguish between problems solvable by the City Hall and those belonging to the central Government or other agencies. They argued that for the platform to be safe for the municipality to use, it must include transparent labelling of responsibility. If the AI or the interface implies that the City is listening to a proposal it has no legal power to implement, it creates a reputational security risk. Therefore, transparency in Braşov is interpreted as the strict, visible delineation of institutional roles.

Security against "Gaming" and Manipulation

In the privacy and security survey blocks (DEM-6), Braşov moderators expressed a strong distrust of anonymous participation modes. Unlike Martin, where the focus was on protecting user identity, Braşov staff prioritized the protection of the voting process. They explicitly requested the removal of anonymous options and the implementation of stricter identity verification to prevent "gaming" (coordinated voting, bots, or spam). For them, a "secure" system is one that guarantees one-person-one-vote and blocks malicious actors from skewing the debate, even if this requires collecting more user data.

Data Hygiene in Exports

Regarding internal data handling, Braşov moderators echoed the need for masking but with a specific focus on vulgarity. When generating reports for upper management, they requested that the AI automatically mask or redact highly toxic language from examples. The concern here is not just privacy, but professional safety: preventing offensive content from entering formal administrative documents while still reporting on the *fact* that such content was detected.

Explainability as a Security Feature

Finally, Braşov moderators linked security directly to legal transparency. They indicated that they would feel "safe" relying on the AI moderator only if it could cite the specific local law or administrative rule that justifies a block. This suggests that for Braşov staff, "security" encompasses legal defensibility: the system must provide the evidence needed to protect the municipality against claims of censorship.

5.2.7 Stability under load and performance-related AIA checks

The technical dimension of the AIA focused on whether AI behaviour remains stable under load, particularly during pre-pilot performance tests (Gate A and subsequent runs). Here the emphasis was on measurable *flip-rates* in moderation decisions and on the stability of coverage of minority views in summaries.

Using a fixed labelled set of borderline moderation items and a small set of representative long threads with identifiable minority viewpoints, the development team compared AI outputs at idle and under peak-load scenarios reflecting expected Phase 2 usage plus a safety margin. Within this limited but realistic scope:

- **Moderation decisions** for the borderline items remained within the predefined flip-rate threshold. No systematic pattern of changes was observed that would suggest a fairness regression under load (e.g. more aggressive removals for specific phrasings only when the

system was stressed).

- **Summaries** of the test threads continued to include the previously identified minority-view sentences when the system was under stress. Differences occurred mainly in phrasing or ordering, not in the presence/absence of particular viewpoints.

These results indicate that, for the tested items and load profiles, performance optimisation and elevated load did not introduce new, overt biases or instabilities in AI outputs. In D4.1 terms, the Scalability Performance KPI (“full performance under peak load”) appears to be met at this preliminary stage for the AI components inspected, at least with respect to stability of decision logic.

That said, the labelled sets are small and the platform continues to evolve. As the pilots scale up and content diversity increases, continuous monitoring of flip-rates, summary coverage and potential drift in moderation behaviour will be necessary to ensure that fairness and stability are maintained.

5.2.8 Synthesis of AIA impact and implications for KPIs

Overall, the Phase 2 AIA paints a picture of AI components that are promising and functionally useful but not yet “frictionless” in terms of impact and KPI fulfilment across sites.

- **Fairness and coverage of summaries**
 - Citizens and moderators in Martin consider summaries useful orientation tools but are concerned about compressed coverage and under-representation of minority views.
 - In Braşov, there is no evidence of dissatisfaction in the small end-user sample, but moderator-level AIA remains a gap.
 - Design implications include explicit signalling of dissent, context indicators and traceability to underlying posts – all of which support both fairness and explainability.
- **Moderation consistency and fair play**
 - Borderline-item tests in Martin show broad alignment between AI and human decisions, with no clear identity-based discrepancies in the tested pairs; fairness appears acceptable as decision support.
 - The main weaknesses relate to explanation and workflow: a lack of explicit reasoning for flags and incomplete documentation for borderline categories.
 - UPAT results add that opaque reward and ranking logic, unreachable missions and technical instabilities can significantly undermine perceived fairness, even when content moderation itself is acceptable.
- **Transparency, privacy and security**
 - Users and moderators endorse basic assumptions about anonymisation and accept that some data processing is necessary, but they call for clearer labelling of AI-generated elements, better explanations of reporting workflows and stricter rules for masking identifiers in exports.
 - Gamification participants appreciate the pseudonymous, protected setup but find the internal game logic insufficiently transparent.
 - These patterns underline that governance measures (export rules, audit trails, internal guidelines) are as important as technical protections for meeting Security Compliance, Trust and Explainability KPIs.
- **Stability under load**
 - Gate A tests suggest that AI summaries and moderation decisions remain stable under elevated load in the scenarios tested, with no significant flip-rate increase or loss of minority-view sentences. This provides a technically reassuring baseline but must be maintained and re-checked as usage scales.

Taken together, the Phase 2 AIA shows that the ITHACA AI components already provide tangible value (helping users and staff navigate complex debates, supporting moderation and enabling playful

engagement) and, in the tested cases, do not exhibit obvious unfairness or instability under load. At the same time, the cross-site evidence highlights coverage, traceability, explanation and governance (including fair-play rules for gamification) as the key levers for making these AI tools trustworthy enough to be embedded in everyday civic and administrative workflows and for systematically progressing towards the D4.1 KPI targets in subsequent phases.

6. Revision of the concept

This chapter deals with the revised ITHACA concept based on the phase 0 and phase 1 formative evaluations (see D4.2), the more summative evaluations in phase 2 (see this report D4.3), as well as the experiences and developments since the initial formulation of the ITHACA concept and reflects the work performed in T4.5 (led by UniGraz). The initial ITHACA concept was completed at the end of the first project year as part of Task 1.7 and revised by September 2024 (see D1.3). It synthesized the main findings, key messages, best-practices, metrics, functionalities and outcomes of all Tasks in WP1 into a holistic concept for ITHACA. In other words: it aimed to extract the 'quintessence' of the State-of-the-art analysis (SOTA) and research carried out in the context of WP1. Similar to the original ITHACA concept, we have also structured the content of the revised concept according to a number of key questions. The individual subchapters, i.e. the individual facets and sub-research areas of ITHACA, have been adopted directly in order to ensure better comparability. The set of key-questions is as follows:

(I) Aspects of the 'quintessence', i.e. aims and contributions beyond SOTA as identified in D1.3:

(Ia) What has been done or implemented?

(Ib) What were the deviations (including positive ones) from the initial concept? In the case of deviations from the initial concept, what are the justifications for this?

(II) Contributions about ITHACA's goal of including vulnerable and marginalised individuals and groups as much as possible:

(IIa) What has been done or implemented?

(IIb) What were the deviations (including positive ones) from the initial concept? In the case of deviations from the initial concept, what are the justifications for this?

(III) Future directions / policy recommendations, and recommendations for similar initiatives / projects / platforms as well as potential ITHACA extensions?

While the key questions **Ia**, **Ib** & **III** for a revised concept are quite obvious, we also addressed central questions of Civic Engagement Platforms (CEPs) in the age of AI as represented by key questions **IIa** & **IIb**: to what extent vulnerable and marginalised groups and individuals can be included, what challenges this poses, and what solutions, methods and approaches ITHACA has applied.

6.1 Social Context, participatory democracy and inclusion of vulnerable groups

Ad (Ia)

During WP2, we asked participants of the AI Citizens' Juries from the two pilot sites, Martin and Brasov, about their experiences, expectations, must-haves, and suggestions. In doing so ITHACA fulfilled (i) 'inclusion by design' and (ii) inclusion by co-creation.

A set of AI features was briefly described and introduced to the AI Citizens' Juries and evaluated if they would like to have the respective feature or not. Based on the evaluations of AI Citizens' Jury members, we sorted the features according to preference (see D2.1, p.75). The following list retains this order and also indicates whether, and how, these features were implemented in the ITHACA platform.

- *Security measure* (see 'Trust and Security Infrastructure' chapters in D3.1 & D3.3 as well as 'AI Cybersecurity Tool' chapters in D5.1 & D5.3)
- *Text translation into another language* (see 'Civic Forum and Topic Discussions' section in D3.1; translation capabilities are from Romanian and Slovak to English and from English to Romanian and Slovak)
- *Event calendar* (not done)
- *Toxicity sensor* (see AI Toxicity Detection Model via the AI Fairness Tool as described in D5.1 & D5.3 as well as D3.1 & D3.3 for details on ITHACAs implementation and integration into the System Architecture)
- *Text translation into simpler language* (ITHACA provides to possibilities, either to summarize long discussions and/or quick, handy translations menu entries on the main page in the smart tools card)
- *Recommendations* (not done)
- *Automated reporting and analysis* (ITHACA is auditing all api calls, i.e. interactions with the backend, and user's login logs. Opening and completing or creating new proposal, as well as adding new comments are analysed)
- *Informed decision making* (Users can read the proposals as well as comments and discussions by others and can actively contribute to the discussion. Going through pros and cons should help to make an informed decision)
- *Text-to-speech / Speech-to-Text* (The ITHACA platform is compatible with voice Speech-to-Text that allows users to interact with the platform hands-free. Text-to-Speech ensures that visually impaired users can navigate and understand the content. To implement these functionalities was challenging because of a lack of good existing models in Slovak and Romanian)
- *Multimedia posts* (Users can add written texts for commenting proposals by others and can also add pictures when suggesting a proposal)

- *Local News Page* (A dedicated web crawler has been implemented to retrieve article text and associated metadata from official municipal websites of the pilot cities at regular intervals; see section 4.2.1 in D3.1)
- *Auto-correction* (not done)
- *Chatbot* (To assist platform navigation, providing information to users, and enhancing user experience, a self-hosted chatbot has been implemented; see D3.1 to D3.4)
- *Collaboration Tools* (not done)
- *Auto Tagging* (not done)
- *Integration of Social Media* (see ‘Social Media Data Acquisition’ as described in section 4.2.2 in D3.1 and 4.1.1 in D3.2; the content of the social media pages of the two pilot sites are collected to gain a deeper understanding of citizens' opinions with sentiment analysis)
- *Group/private chat between users* (user to user chat is implemented)
- *Sentiment Analysis* (not in the narrower sense, but by clicking in the community forums on proposals, at the bottom of the screen we can see which proposal had most positive / negative votes)
- *Gamification* (A Gamification Tool has been implemented based on psychologically-sound considerations, i.e. based on the Self-determination Theory as described by Ryan & Deci (2002) and Ryan et al. (2021) to fulfill basic needs of autonomy, competence and social relatedness, aiming to transfer extrinsic into more sustainable intrinsic motivation; see chapter 7 in D3.1 and section 4.1.5 in D3.2)

Ad (Ib)

The initial idea to increase the ‘weights’ (of their comments, feedback, etc.) from underrepresented groups in the platform’s training data has not been implemented - as it would have contradicted the principle of equality, even if the underlying goal would have been to mitigate biases. This imbalance (i.e. different weights), in the sense of positive discrimination would have been associated with further biases and underlying ethical questions, which would not have solved any problems but created new ones - just as an example: What weight should a single adult with a low socio-economic background receive compared to someone with limited access to infrastructure/mobility?

Ad (IIa)

As mentioned above, from the very beginning, the ITHACA consortium was keen to be as inclusive of vulnerable and marginalized groups and individuals as possible. Based on predefined vulnerability criteria, such as low socioeconomic background, as the foundation for the AI Citizen Juries selection process, we aimed to gather as many user-centric requirements, perspectives and expectations as possible. For both pilots, we conducted: (a) online introductory workshops and Delphi studies with representatives of vulnerable and marginalized groups and communities, (b) onsite introductory and algorithmic risk assessment AI Citizen Juries, as well as (c) online focus groups with experts (on IT, AI, HCI, and Ethics). In the course of WP4, i.e., in the evaluation studies in all phases (phase 0 and phase 1; see D4.2) an attempt was also made to obtain as diverse a sample as possible in order to

verify that the platform is indeed inclusive and accessible for a range of vulnerable groups and individuals (e.g. low digital literacy or visual impairment, etc.). This fundamental concern of the ITHACA consortium was also carried through in the final mixed-method phase 2 evaluations; a description of the participants' backgrounds can be found in section 4.2.1 of this report. Successively, a series of shortcomings were identified by mixed-method approaches, combining qualitative and quantitative data, prioritized and remedied as best as possible from phase 0 to phase 1 and finally to the final phase 2 evaluation (see for example section 5.6.7 and chapter 6 in D4.2 as well as section 4.2.6 in this D4.3).

Ad (IIb)

In our opinion it is fair to state that the highly sophisticated method for the AI Citizens' Juries process is a positive deviation from the initial plans.

This selection process included the following steps:

1. A more 'sociological' analysis of the residents in Brasov and Martin from the pilot partners.
2. An overall identification of vulnerable groups and communities from the literature encompassed by experiences from the colleagues who are residents of Brasov and Martin.
3. Identification of variables and criteria that constitute vulnerability and marginalization.
4. Contextualized (i.e., based on the concrete situations in Brasov and Martin) binary mapping between vulnerable social groups and communities and vulnerability criteria.
5. Assignments of numeric weights to the items of the vulnerability selection questionnaire.

Through this process, we obtained a diverse sample for the Citizens' Juries.

Ad (III)

Future directions

- Future studies could assess the quality of group discussions on the ITHACA platform and compare it with social media discussions and traditional/ in-person discussions. For instance, to explore the impacts of the inclusion-by-design approach on the variety of opinions and occurrence of friendly or toxic statements.

Policy Recommendations

- Using simple language for all legal requirements to include vulnerable and marginalized groups and individuals - example: a GDPR agreement is required. That's fine. However, the GDPR agreements are quite complex themselves (e.g. brief description of procedure of data processing, data retention, etc.). Therefore, a brief summary in simple language might be helpful to explain what the respective sections are actually about and why they are described in such detail (e.g., why is there a paragraph on data retention?). In other words, the aim of the GDPR was to be transparent to users about what data is collected and stored, why, for how long and for what purpose, and to obtain their explicit consent. However, the text itself

cannot be described as absolutely inclusive, as there are technical terms that must be included. Therefore, a summary in simple/ plain language might be helpful.

- Developing a course/curriculum to train municipal employees so they can use AI tools for civic participation correctly, make better choices and give better advice to citizens.

Recommendations for similar initiatives / projects / platforms / ITHACA extensions

- People who ‘have to accept’ a GDPR agreement online or in printed form as part of their studies may, in the first case (online), be unable to ask anyone if they do not understand something, and in the second case, they may feel reluctant to ask the researcher, as it may be considered embarrassing to ask for clarifications. An AI-supported chatbot that could be consulted anonymously for precisely such questions (relating to regulations, legal texts, etc.) would therefore be a relevant contribution to increasing inclusion in society.
- Similar future platforms, that are large enough, could potentially form a partnership with private companies, in economic or product development sectors. For example, companies could use a similar platform to better assess the actual desires of their potential customers, thereby incorporating civic participation into product improvements.
- Future similar initiatives could focus on how the needs of vulnerable groups are evolving. Meaning, what features should be added after perhaps months of use and what features turn out to be a waste of resources after all and therefore can be omitted. Such initiatives could help develop a spreadsheet that specifies the criteria that an inclusivity feature should definitely fulfil to be considered worthwhile adding/developing.

6.2 Legal Frameworks and Policies

Ad (Ia)

In the context of ITHACA, all legal and regulatory ‘must-haves’ identified at the beginning of the project have been addressed. Privacy, transparency, fairness and security were operationalised through concrete tasks and deliverables. The following table below maps each requirement to what has been implemented in the project.

Table 14. Summary of legal requirements and their implementation in ITHACA

Requirement (Must-have)	Implementation in ITHACA
Privacy by Design & Default	Conducted a dedicated Privacy-by-Design / Privacy-by-Default report and a detailed Data Protection Impact Assessment (DPIA).
Transparency in Data Processing	A detailed Record of Processing Activities (RoPA) covering purposes, categories, recipients and retention as well as a Privacy Notice has been drafted,
Transparency in Automated Decision-Making	Platform avoids fully automated decisions. The content-moderation workflow for toxic-speech detection includes human-in-the-loop oversight; a fairness/explainability tool (see D5.1 - D5.4) supports transparency of AI-assisted components.
Data Minimisation & Purpose Limitation	Only essential user data is collected. Continuous legal checks through technical/PM meetings, and the DPIA. A PPML tool (see D5.1 - D5.4) was created to enhance privacy-preserving processing.

Requirement (Must-have)	Implementation in ITHACA
Security Measures	Security measures implemented by technical partners. An AI cybersecurity tool (D5.1 - D5.4) was designed to harden components and support secure operation.
Explicit Consent (Cookies & Tracking)	Informed-consent form prepared for participants. Cookies policy template and cookies instructions produced; no cookies/trackers have been added to the platform so far. A cookie banner will be enabled if/when trackers are introduced.
Human Oversight	Platform design avoids AI-only decisions; the moderation dashboard and audit log ensure human review and accountability across workflows.
Transparency, Explainability & Fairness	Privacy Notice and AI Notice drafted. Fairness tool designed under WP5. See also the synthesis of the Algorithmic Impact Assessment in section 5.2.8 for an overview of the main conclusions from the phase 2 evaluation.
Avoidance of Unacceptable-Risk AI Features	No prohibited functions (e.g., social-scoring) are included in the platform design.
Regular Compliance Audits & Updates	Ongoing internal compliance checks via the DPIA process and WP8 (Ethics) oversight.

Ad (Ib)

- Early compliance groundwork: The Record of Processing Activities (RoPA), along with the Privacy by Design/Default report, were prepared right at the start of the project. This gave the technical teams a clear compliance baseline to work from during development and pilot phases.
- Draft ePrivacy Regulation not implemented: Since the draft ePrivacy Regulation never came into force, the consortium wasn't required to adopt any of its provisions - no action was necessary on that front.
- No cookie obligations: The platform was intentionally designed without cookies or similar tracking tools. As a result, the key requirements under the current ePrivacy Directive regarding cookies didn't apply. This decision also meant there was no need for a cookie banner or related compliance steps at this point.

Ad (IIa)

In meeting GDPR requirements (see EP&CEU, 2016), the consortium placed particular emphasis on protecting vulnerable and marginalised individuals. A key part of this effort was limiting data collection to what was genuinely necessary, following the principle of data minimisation. Technical and organisational safeguards were also put in place to strengthen data protection. From early on, the team made a deliberate choice to avoid collecting sensitive personal information unless there was a compelling reason to do so. This approach helped lower the risks of profiling, exclusion, or discrimination, and aimed to ensure fair and equal access for all users engaging with the platform.

Ad (IIb)

An important positive outcome was the consortium's alignment with the European Accessibility Act (see EP&CEU, 2019), which came fully into effect in June 2025. While this wasn't part of the original plan, accessibility-by-design principles were incorporated into the platform to make sure people with disabilities — along with other vulnerable users — could take part fully. This adjustment went beyond the project's initial scope and ultimately made the ITHACA platform more inclusive and accessible for everyone.

Ad (III)

(1) Recommendations for ITHACA-like future calls, projects and initiatives

Future calls and projects should no longer frame CEPs as one-off, short-term experiments and pilots. Instead, they should explicitly require from project promoters that digital participation tools are designed and governed as elements of the permanent democratic infrastructure of the municipality or public authority that deploys them. This should entail, at the very least:

Making CEP-based consultations a recurring element of formal decision-making cycles (e.g. for legislation, planning, budgeting).

Requiring long-term evaluation of their impact on democratic participation and deliberative quality, using indicators such as inclusiveness of participation, diversity of contributors, or the actual influence of online input on decisions.

In other words, funding schemes should ask project promoters to define at the outset how the platform will be sustained (financially, technically, politically) beyond the project's lifetime, as well as how its democratic effects will be periodically assessed.

(2) "Inclusion by default" as a core design and compliance requirement

A key added value of CEPs for democracy is their potential to make it possible for people to participate who have traditionally been excluded from such processes. At the same time, purely digital tools have the potential to reinforce exclusion for those with low digital skills or limited access to technology.

For this reason, "inclusion by default" should become a design and compliance requirement for similar initiatives in the future:

Projects should be obliged to adopt a formal inclusion policy, describing how they will reach under-represented groups (e.g. low-income communities, migrants, people with disabilities, older persons, digitally illiterate citizens) through targeted outreach, partnerships with local organisations and hybrid offline/online channels. As described in section 6.1, the ITHACA consortium applied a fairly comprehensive methodology to include underrepresented, marginalized, and vulnerable groups and individuals as much as possible (see also D1.3 and D2.1).

CEPs should be required to comply with relevant accessibility and inclusion standards (e.g. accessibility-by-design, plain language, multi-device access), and to document how their design choices help reduce, rather than exacerbate, the risk of exclusion or discrimination.

Inclusion should thus be treated both as a legal-compliance issue (non-discrimination, accessibility) and as a democratic-quality objective.

(3) High-risk-level safeguards and Fundamental Rights Impact Assessments for AI use

Even where AI components in CEPs do not formally fall under the "high-risk" category of the AI Act (see EP&CEU, 2024), their potential impact on democratic participation and fundamental rights is not trivial. Therefore, future projects and platforms should be required, by default, to apply the core

safeguards associated with high-risk AI systems to any AI used in civic participation contexts (e.g. clustering, summarisation, toxicity detection, recommendation tools).

This includes at the very least:

- Systematic risk management, documentation and logging of AI-assisted functionalities;
- Clear human-in-the-loop oversight and the prohibition of AI-only decisions affecting participation rights; and
- Meaningful transparency and explainability towards users (see also section 5.2.8 for the main conclusions based on the phase 2 evaluations in ITHACA).

In addition, the deployers of CEPs (i.e. municipalities, public bodies, etc.) should be obliged to carry out a Fundamental Rights Impact Assessment (FRIA) for the use of such AI systems, ideally combined with data protection impact assessments where personal data are involved. This should be an explicit requirement in future calls and in procurement/contracting of CEP solutions.

(4) Ethical committees for content moderation – balancing harm prevention and freedom of expression

Finally, future CEP projects should, by default, establish an independent or multi-stakeholder ethics/oversight committee with a clear mandate to review and monitor content-moderation policies and practices.

The role of such a body would be to ensure that moderation effectively addresses illegal content and toxic speech that could harm individuals, chill participation or undermine the integrity of the deliberative process, while at the same time safeguarding freedom of expression and avoiding disproportionate interference with legitimate opinions, political speech, criticism of authorities and the pluralism of public debate.

The committee should regularly review criteria, workflows and impact of moderation (including appeals statistics and patterns of removals), and issue recommendations where the balance between harm prevention and free expression is not adequately maintained. In this way, content moderation in CEPs becomes an accountable, transparent and value-aligned process, rather than simply a risk-avoidance mechanism.

(5) Enrich CEPs with contextual information from external sources, not only direct user input

Future CEP projects should move beyond the traditional model of a “blank” participation platform that relies exclusively on users coming in and posting their views. Stand-alone CEPs with no broader informational context are often not attractive enough to sustain engagement, especially when citizens are already interacting on other channels (social media, news websites, NGO pages, municipal portals, etc.).

To address this, future initiatives should:

- *Systematically enrich the CEP environment with contextual information* about the municipality, the topic under discussion and the wider public debate, by integrating feeds or dashboards from trustworthy external sources (e.g. municipal and regional websites, official social media profiles of municipalities and NGOs, local media coverage, reactions and engagement metrics around relevant articles).
- *Do so in a transparent and rights-respecting way*, clearly labelling sources and methods of aggregation, and ensuring compliance with data protection, platform and copyright rules.

6.3 Ethical Frameworks and Guidelines

Ad (Ia)

As outlined in D1.3, no new ethical guidelines have been developed or suggested. We chose to stick with the existing guidelines gathered through the research presented in D1.3 and tried to adhere to a handful of key guidelines, as it is advisable to focus on a smaller number of mutually exclusive principles that are clearly defined and have consequences in case of non-compliance. We applied the Responsible AI principles of Akbarighatar et al. (2023), which build on justice as fairness and follow a pyramid structure where lower-level principles are prerequisites for higher ones. Accordingly, we prioritized these prerequisites, such as privacy and security, benevolence, non-maleficence, and safety and reliability, over transparency, accountability, interpretability, explainability, liberty, inclusiveness, and fairness. The evaluation of the extent to which ITHACA has implemented key ethical principles, as well as the conclusions to be drawn from this, are described in detail in chapter 5 on Algorithmic Impact Assessment.

Ad (Ib)

There were no deviations from the initial concept.

Ad (IIa)

In her PhD thesis, Maria Zangl (UniGraz) investigated themes related to how humans perceive, conceptualize (and possibly misconceptualize) GAI and its corresponding applications. The aim of one of her studies was to identify gaps between technical experts and users' perceptions and to investigate if and to what extent a common understanding of trustworthy AI and ethical principles exists. This user study was conducted with lay participants that have not received formal education about AI or certain GAI applications. We asked them about their understanding of privacy and security and mechanisms of AI in general and put an emphasis on capturing participants' explanations regarding AI tools that are accessible to citizens without prior knowledge. The skills enabling joint ethical AI discussions are closely tied to the Explainability principle (see also section 6.6 for a more in-depth discussion), requiring systems to be not only technically explainable but also interpretable for users regardless of expertise. This demands simple, accessible language and regular comprehension checks to ensure true inclusiveness and a human-centric approach. Such competencies empower users to critically and constantly evaluate prevailing definitions and descriptions of AI attributes, decisions, and features. It allows them to uncover contradictions or flawed reasoning. In turn, this reduces dependence on marketing claims or principles set by product owners and fosters a more participatory and accountable discourse on trustworthy AI. What is important to recognize is that, in the case of understanding AI and having the competencies to use AI applications safely and efficiently, not only marginalized groups but the majority of the public can be seen as vulnerable and at the mercy of tech giants.

Ad (IIb)

One suggested aim was to take into account methods that help reduce the consumption of resources and energy of an AI system, which belongs to ITHACA's challenges and tasks. For these reasons, no new Large Language Model was created for the ITHACA platform, as the training of such a chatbot in the Slovakian and Romanian languages, keeping it running, and constantly improving it would be i) very resource intensive (also from an ecological perspective) and ii) time consuming. As a solution, ITHACA decided to make use of existing LLMs used as chatbots (see D3.1 to D3.4 for

technical details). Although we cannot control the ecological impact anymore, no new data must be collected to train a new model and existing models can be used.

Ad (III)

Ethical principles of AI can be defined differently by technical developers and by laypersons / average users. In order to reach a consensus on what a trustworthy AI is, these differences need to be addressed in a way that satisfies the requirements of both perspectives.

Numerous adjectives are used to describe ethical principles, such as secure, robust, fair, unbiased, inclusive, etc. In some cases, it is mentioned that an ethical principle A is part of / a prerequisite of an ethical principle B. But then it is mentioned that it is the opposite, that B is a prerequisite for A. In other words, definitions are sometimes tautological and / or contradictory. Future secondary research could rigorously examine these definitions and draw attention to any contradictions.

We identified an important future research challenge and opportunity: A focus on fundamental skills and competencies that enable individuals in different contexts to critically engage with AI tools and outputs, i.e. extend media literacy in the 'traditional sense' to include challenges posed by GAI. The field of technical development is rapidly evolving, which causes the problem that competency frameworks (applicable in formal, non-formal and informal educational contexts, e.g., in schools, adult education institutions, etc.) lag behind these technical developments. Competency frameworks must therefore not be static but flexible enough to remain effective.

Future research and applications could focus on adaptive educational interfaces that detect and address misconceptions that citizens might have when using an AI tool in real time. For example, if a user asks the integrated chatbot in ITHACA for personal or political opinions or assumes it uses individual input to 'think', the system could briefly explain its actual functioning and limits. Early in the interaction, short diagnostic prompts (e.g., asking how users think the AI tool generates answers) could reveal mental models and trigger tailored guidance or tooltips. Interactive visuals such as simple flowcharts or analogies could clarify generative processes and counter false assumptions. Gamified elements like quizzes or 'myth vs. fact' challenges based on identified misconception categories could further enhance understanding by a motivating and incidental learning approach (compared to intentional learning).

6.4 Good practices of citizen engagement and participatory democracy

Ad (Ia)

ITHACA activities are aimed at moving from theory to practice by deploying its platform in real cities (Braşov, Romania and Martin, Slovakia) with actual citizens participating. The project utilized and targeted vulnerability criteria (e.g. low income, low digital literacy, social exclusion) and engaged NGOs and multipliers for the recruitment of participants from underrepresented communities. During pilot sessions, participants engage in think-aloud tasks and provide immediate feedback during platform use. This kind of qualitative user testing helps uncover real usability, accessibility, and transparency problems. The project has given live demos of the front-end interface, showing that the interface is functional, at least in demonstration mode. User testing through think-aloud protocols and immediate feedback during platform interaction reflects the good-practice requirement for intuitive and user-friendly design, as stressed in the evaluation criteria used in D1.1. This approach aligns with the study's findings that best-practice platforms (e.g., Decidim, Adhocracy+, EngagementHQ) succeed when interfaces are accessible, navigable, and transparent about how

input is processed. Furthermore, ITHACA demonstrated functional prototypes of the interface and conducted live demonstrations of the front end, showing an implemented system ready for user testing; similar to the hands-on, iterative deployment process used in many of the evaluated best practices.

Additionally, ITHACA organised early focus groups to explore participants' perceptions, expectations, and concerns regarding AI in decision-making, which mirrors the emphasis in the study on transparency, citizen feedback, and ethical/ legal considerations as foundational for trustworthy democratic technologies. Platforms considered in the scope of ITHACA as "best practice" embed structured mechanisms for identifying risks (e.g., bias, privacy concerns, fairness) and for communicating system functioning to users. ITHACA's early engagement approach is consistent with this requirement. The collected insights were subsequently used to refine system requirements, ensuring that citizen perspectives shape the development of the platform from the outset. A key principle in responsible and participatory urban technology design, as demonstrated in all of the applications analysed (e.g., Better Reykjavik, ManaBalss, Consul) which iteratively incorporate public input into the design and governance processes.

Ad (Ib)

The feedback loops and tracing from citizens' contributions to municipal decision-making have not been implemented to show concrete policy impact yet. The metrics for deliberative quality (balanced argumentation, inclusion indices) until validated, cannot prove their concept goal and the tools engaged for this reason. Same accounts for the audit / governance tools. The emphasis on vulnerable groups and offline/hybrid participation formats has been strengthened early on, reflecting an adaptive response to local partner needs and COVID/post-COVID conditions.

Ad (IIa)

ITHACA's inclusion efforts reflect concrete implementation of the good-practice principles identified in D1.1. The project adopted the study's emphasis on inclusive processes, engaging vulnerable groups early in the design cycle through targeted recruitment supported by NGOs and local actors, and mirroring approaches used by other platforms. This early engagement generated detailed insights into barriers and facilitators; digital literacy, language, disability, socio-economic factors echoing the addressed recommendations to assess accessibility needs before deploying technological solutions. The platform design integrates these findings through accessible and intuitive interfaces, consistent with the usability and design criteria identified and reinforced by the mixed-methods user-requirements documented and the diverse participant profiles involved in Phase 2 evaluation.

Ad (IIb)

ITHACA deviated from the initial concept in several ways, primarily due to ethical, organisational, and temporal constraints. While the D1.1 analysis highlighted a broad set of AI features used by mature platforms including predictive analytics, recommendation engines, and automated summarisation, ITHACA implemented foundational AI components during the pilot phase. This deviation is justified by GDPR requirements, the need to minimise data processing during early testing and the limited dataset sizes available. Similarly, although the study reviewed platforms already integrated into municipal governance processes, ITHACA pilots did not yet achieve formal institutional integration, as such integration requires long-term political and administrative commitments beyond the scope of the pilot cycle.

The project placed stronger emphasis on hybrid and offline-inclusion formats than the original online-centric concept implied. For example, the pilot in the city of Martin focused specifically on vulnerable group members such as seniors, people with disabilities, and members of marginalized communities. This adjustment recognises that purely online tools may exclude digitally disadvantaged participants, so in-person sessions ensured higher accessibility and credibility among vulnerable groups. The selection process for the “AI Citizens’ Juries” has been more nuanced and context-sensitive than originally sketched: the project developed a multifaceted methodology for identifying vulnerability variables (age, income, education, mobility, etc.), and tailored the criteria differently in Martin vs Braşov.

Conversely, ITHACA went beyond the initial concept in positive ways. The project implemented a far more intensive co-design process than described in the study, engaging vulnerable groups early through focus groups, think-aloud sessions, and NGO-supported recruitment. The selection of two socio-economically distinct pilot sites (Romania and Slovakia) operationalised the recommendation for inclusiveness by testing the platform under varied social conditions. Pilot results have confirmed the value of these deviations are expected to generate valuable technical usage data, such as interaction patterns, feature performance, error points, and user-journey bottlenecks which will be used to further enhance platform functionality after the pilot phase. This iterative learning process aligns with the recommendation to refine digital participatory tools based on real-world behavioural data, ensuring that ITHACA evolves into a more robust, effective, and user-centred system.

Ad (III)

Early results indicate positive engagement, improved inclusivity, and a willingness among previously under-represented groups to participate. Ethical and governance considerations have been addressed through design deliverables, and Personal Information Management Systems (PIMS) have been conceptualised to give participants control over their data. Despite advances, several planned features remain questionable in terms of their performance. Full implementation of AI modules including argument clustering, sentiment analysis, summarisation, moderation support, and transparency dashboards have not yet been documented in the pilot operation which is the tricky part. Similarly, systematic metrics for deliberative quality, quantitative outcomes for vulnerable groups, and clear feedback loops linking citizens’ input to municipal decision-making are still pending. These gaps indicate opportunities for refinement in subsequent project phases and for future platforms seeking to advance inclusive, AI-mediated citizen engagement. Based on current experience and emerging lessons, future directions emphasise continued prioritization of inclusion, transparency, and ethical governance. Platforms should maintain hybrid engagement channels to reach digitally excluded populations and adopt context-sensitive criteria for participant recruitment, reflecting local socio-cultural conditions. Iterative co-design, mixed-method evaluation, and comparative studies should be conducted to assess deliberative quality, policy impact, and usability. Policy recommendations include encouraging municipalities and EU-level bodies to adopt scalable and adaptable civic engagement frameworks, strengthen administrative capacity for AI-mediated platforms, and integrate citizen contributions into decision-making processes. To that end, potential extensions for ITHACA include expanding pilot deployments to additional municipalities with diverse governance contexts, enhancing AI modules to support other-lingual and cross-cultural deliberation, integrating real-time feedback mechanisms for deliberation quality, and exploring interoperability with other civic-tech platforms to increase reach and impact. These extensions would enable the platform to scale effectively, strengthen inclusion, and continuously improve alignment between citizen voices, deliberation quality, and local policy outcomes.

6.5 Compliance, Trust and Privacy preserving tools, methods and approaches

Ad (Ia)

In the context of Privacy Preserving Machine Learning (PPML) tool development (Task 5.1), we evaluated the different PPML methods covered in D1.3 (Chapter 4.3), and we concluded that Differential Privacy has many advantages compared to the rest of the methods, such as advanced data concealment, being computationally cheap and fast, as well as complying with relevant legal guidelines as mentioned in deliverable D5.1 (Chapter 2.3). More specifically, we exploited the Differentially Private Adam since it outperformed other methods, and it also comes from a reputable source, i.e. the Tensorflow (Google) research team.

In the case of GAI, models being utilized in the context of the ITHACA platform (e.g. chatbot), the PPML tool implemented during task T5.1, can be used to evaluate their privacy. Moreover, partners from the consortium delved into more details regarding diversity, non-discrimination, human oversight, robustness, safety etc., in all AI models (not limited to GAI), through their recent publication (Zangl et al., 2025). Key evaluation results and conclusions on compliance, trust and privacy are included in chapter 5 as part of a holistic algorithmic impact assessment.

Ad (Ib)

During the conceptualization of the ITHACA platform (deliverable D1.3 - Chapter 6.3.2), the employment of a tool/tools to preserve the compliance and trust, as well as for governance and risk management in the AI models running in the background of the platform, was recommended. After evaluating the tools analysed and compared in D1.3 (Table 8), we deduced that none of these tools comply with the relevant AI regulations (e.g., AI Act) or meet the project's needs. Hence, we focused on developing tools to evaluate the trustworthiness of the AI models of the ITHACA platform (T5.1), which adhere to the regulations as described in deliverables D5.1 and D5.3 (Chapters 2.2 & 2.3), and are designed and developed to be easily integrated into the ITHACA or any other civic engagement platform. These tools can be utilized to assess the conformity of both traditional AI and GAI models to fairness, privacy, and security principles.

Ad (IIa)

The AI Fairness tool developed in the context of task T5.1, guarantees conformity with the fairness principle, to promote equality, inclusivity, and oppose discrimination as well as render the evaluated AI system more trustworthy and unbiased. The tool employs objective fairness metrics (e.g., treatment equality, disparate impact, between-group generalized entropy, etc.) to evaluate whether an AI model is biased and unfair towards people from vulnerable and marginalized groups.

Ad (IIb)

We refrained from incorporating any of the compliance and trustworthy tools reviewed in D1.3 (Table 8), which would aid in avoiding discrimination and exclusion, for the reasons described in Question 1.2. Instead, we ensured that the AI Fairness tool integrates a combination of fairness metrics to produce accurate results (during the fairness evaluation of AI models) and thus offer a safe, equal, and inclusive space for all users regardless of which group they belong to.

Ad (III)

During the lifetime of the ITHACA project, we worked on a publication that focuses on a single aspect of the evaluation of trustworthiness in AI models. This paper regards explainability and was recently published (Zangl et al., 2025). As future work, we aim to work on a review paper as well, covering tools to maintain compliance, governance, risk management, and trust in both traditional and GAI models.

6.6 XAI Open source tools

Ad (Ia)

ITHACA identified a set of essential requirements for Explainable Artificial Intelligence (XAI) tools in the context of democratic participation, as well as areas where innovation beyond the current SOTA is required. These requirements and gaps are grounded in a systematic analysis of existing open-source XAI frameworks, academic literature, and governance-oriented AI transparency needs. Core requirements include the transparency of AI reasoning, accessible natural-language explanations, interpretability of input-output, interactive explanations and contestability, bias detection and mitigation supported by XAI, visual explainability dashboards, customisable explanation depth (users able to choose the level of summaries and receive detailed technical explanations) and comprehensive transparency for generative AI.

Contributions beyond SOTA identified in D1.3 highlight the main research and development gaps which require innovation beyond currently available tools. These areas include the development of model-agnostic explainability methods that can operate consistently across diverse NLP architectures, standardised interpretability metrics for enabling systematic comparison and auditing of XAI methods and last, explainability frameworks for generative AI, particularly in settings where accountability and traceability are essential.

ITHACA has already laid the groundwork for integrating Explainable AI (XAI) into its civic participation platform by conducting a systematic assessment of open-source AI tools and NLP technologies capable of improving transparency, clarity, and accountability in democratic decision-making. The project identifies multiple open-source solutions that can provide XAI functions essential to the project, including automated bias detection, transparency of content moderation, argument classification and intelligible summarisation to support informed participation. Rather than adopting specific pre-existing XAI libraries, the project implemented an explainability-by-design, ensuring that outputs such as argument extraction, summarisation, topic structuring and sentiment interpretation are accompanied by transparent and interpretable reasoning pathways, in alignment with the XAI operational framework described in D1.3. To ensure alignment with democratic values and public accountability, the evaluation framework applied explainability through inclusiveness, transparency, privacy, fairness, and accountability criteria, positioning ITHACA as a responsible, user-centered innovator in AI-supported participation. Importantly, the deliverable underscores that no existing civic participation platform has yet combined GAI with XAI tailored to democratic processes, marking this as a clear beyond-SOTA innovation. The platform incorporates mechanisms that reflect both local and global interpretability needs, including the ability to clarify how specific inputs contributed to model outputs and how model behaviour functions on a broader level. Additionally, the project references EU regulatory obligations, notably the EU AI Act, emphasising that explainability is a legal requirement for GAI in sensitive social domains such as public participation, thereby ensuring ITHACA remains at the forefront of compliant and trustworthy AI deployment in democratic governance. ITHACA addressed the gaps identified concerning the lack of XAI frameworks for

Generative AI with the implementation of explainability safeguards such as content support and chatbot-style interactions. These safeguards include transparency about how generated text is produced and how data sources and prompt structures influence outputs, in line with the EU AI Act's explainability obligations.

Ad (Ib)

In the area of XAI, the ITHACA project has deviated, shifting from a limited set of pre-identified technical enablers toward a more adaptive open-source ecosystem strategy evaluated for transparency, accountability, and interpretability requirements applicable to democratic processes. This deviation was justified as necessary to ensure regulatory alignment, which imposes explicit explainability obligations for GAI in socially sensitive applications such as civic participation, a legal and ethical dimension more strongly emphasised than in the original technical concept. Additionally, the initial plan assumed that existing civic-tech examples would provide transferable explainability models. However, since no current platform successfully integrates GAI with full XAI safeguards for democratic decision-making support, meaning ITHACA had to shift from technology adoption towards leading innovation in developing human-comprehensible AI-supported argument analysis and content transparency. Deviations to date seem to be based only on the ability to fully integrate these tools and provide concrete demonstration of them.

Testing shows that AI summaries, moderation aids and analytic tools function reliably under load, without systematic unfairness or discriminatory patterns. However, users highlighted weaknesses in explanation, coverage, and traceability. Moderators found AI support helpful but requested clearer justification paths and improved workflow integration. Gamification testing identified issues with opaque reward logic and technical instabilities, prompting revisions in platform governance, fairness rules and usability improvements.

Ad (IIa)

The XAI tools and approaches identified have been selected with a strong focus on protecting and empowering vulnerable and marginalised participants in democratic processes. They emphasise that transparency, understandability and accountability are essential conditions for inclusive civic participation, especially for individuals with lower digital literacy or limited trust in digital systems. By using natural language processing models deliberately chosen for their ability to reduce complexity and provide clear, contextual explanations of AI-generated insights, ITHACA aims to offer features such as comprehensible argument summarisation, accessible sentiment interpretation and supportive content structuring so that individuals with fewer technical skills can engage confidently in online deliberation. The evaluation framework applied to these tools requires inclusiveness, transparency, fairness, and privacy as core metrics, explicitly linking explainability to the representation of minority voices and equitable treatment of all user inputs. ITHACA also further highlights that GAI, when used without explainability safeguards, may reduce trust and exacerbate exclusion, especially for groups already distant from institutional participation and therefore must comply with explainability obligations in the EU AI Act to maintain democratic legitimacy in sensitive civic contexts.

Ad (IIb)

There is a big issue which has not been tackled either by ITHACA or by any other current civic-tech platform before, concerning the inclusion of an explainable GAI-supported participation or identifying a clear innovation need to ensure that advanced AI tools not only assist engagement but do so in ways that marginalised groups can understand, scrutinise, and benefit from equally.

Ad (III)

Looking ahead, XAI in civic-participation platforms should move from ad-hoc explainers to a full governance stack that couples open-source NLP tooling with user-facing transparency and legal compliance. In practice, that means to deliver explanations ordinary people can actually use, while retaining the nine-criterion lens (inclusion, transparency, fairness, privacy and accountability,..) so marginalised groups can understand and contest automated summaries, rankings or clustering that shape deliberation. Policy direction from the EU now makes this imperative rather than optional. The AI Act (Reg. 2024/1689) requires GPAI providers to publish training-data summaries, disclose AI-generated content and implement safeguards, setting a floor for explainability and documentation that civic platforms should meet or exceed with model cards and datasheets for datasets and with audit-ready technical documentation (and clear notices to users) built into the platforms. Beyond EU law, the OECD AI Principles (OECD, 2019) and the Council of Europe's (2024) Framework Convention on AI anchor explainability in democratic values and human-rights impact assessment, similar initiatives should therefore adopt participatory risk/impact reviews with civil-society input, publish accessibility-tested explanations, and create "contestability" flows so people especially those with low digital literacy can challenge or override AI-assisted moderation or topic grouping. For ITHACA specifically, recommended extensions include: operationalising a real-time panel and provenance/attribution cues across the UI, adding plain-language rationales and counterexamples for model outputs, releasing open model/dataset documentation for any reused components from platforms and piloting transparency dashboards and participatory audits to evidence inclusion gains for vulnerable users.

6.7 Personal Information Management Systems

Ad (Ia)

Personal Information Management Systems (PIMS) aim to deal with user data in a responsible way that enables users to have access and control over their data. **Trust** and **accessibility** are key requirements from an end-user perspective when users access their own data using PIMS. The need for data access and control in a secure and responsible way applies particularly to ITHACA, ensuring that all users, especially vulnerable groups, manage the data they provide for democratic and citizen participation.

In the first phase of the project a model has been developed that provides a holistic description of the aforementioned requirements and the user data to be stored. This model outlines an ideal case by fulfilling as many requirements as possible, in order to increase trust and accessibility to a maximum. Furthermore, a set of user data is listed that might be relevant for the ITHACA platform, in order to exemplify the need for the requirements. Requirements and user data are listed and described in the deliverable D1.2, along with an analysis of existing participation platforms dealing with user data. The requirements are listed here again:

- Personal data storage. A record or storage of personal user data needs to hold all personal user data, such as own contributions in form of various media, reactions on contributions of others, personal settings, as well as permissions and consents.
- Data security. The storage and transfer of personal data must adhere to the highest security standards. Specifically, it is imperative to guarantee the implementation of appropriate technical and organisational measures, such as encryption and access management systems. Additionally, the transfer of personal data outside the European Economic Area

(EEA) should either be avoided or be conducted in compliance with the guarantees described in Articles 44-50 of the GDPR.

- Interoperability. Personal data saved in a data storage needs to be accessible by the various modules of the ITHACA platform, which requires a well-defined and documented API, as well as a documented data model with sufficient semantics.
- Privacy. It is essential to uphold the principles of 'Privacy by Design' and 'Privacy by Default' in the handling of personal data. This entails implementing robust technical and organisational measures to ensure data security, in alignment with the 'Privacy by Design' principle. Concurrently, 'Privacy by Default' measures must be in place to safeguard user privacy. These measures include collecting only the data that is strictly necessary (data minimization), restricting any further use of the data, and employing techniques such as anonymization and pseudonymization to protect the collected data.
- Data analytics. Personal data should be anonymously analysed, in order to draw conclusions that can be used by an application to support the user. If data analytics is not performed anonymously, then additional consent is needed.
- Data management. Only necessary user data should be stored needed for the system's functions and no data should be stored without purpose (data minimisation). Furthermore, users need to be able to access and delete their own data.
- Data integrity. Personal data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures (data integrity principle). This can be achieved by appropriate backup mechanisms, which ensures that data remains intact.
- Permission management. The users should be enabled to perform fine grained settings about which application or function is allowed to access which data.
- Consent management. Consent management is needed to put people in control of which data are used for which purposes. This includes that users can deny or revoke the use of personal data.
- Transparency. It must be shown to the user, which personal data is collected, stored, and how it is processed. This includes a complete list of data and the purposes of use thereof along with, where, and why the data is used. This requirement is strongly related to understandability and explainability of the presented information.
- Sharing visibility. It should be shown which data is shown to other users, user groups, or the public.
- Traceability. Actions and processes should be recorded and presented in a way that users can understand past activities and the development of a process.
- Understandability. All information presented to the end-user (e.g., transparent presentation of stored data, stored permissions and consents, and explanations) should be understandable by end-users. This includes the use of simple language, easy to understand explanations of technical concepts, and understandable presentation of graphical data.
- Explainability. The processing and presentation of data, consent, and permissions has to be explained in a way that is understandable by the end users. Within the scope of the ITHACA Project, which incorporates advanced artificial intelligence (AI) techniques for data processing, the principle of explainability is of paramount importance. Given that AI methods can often function as black boxes, it becomes crucial to elucidate how data is processed and how specific outcomes are derived. This is not just a matter of ethical transparency but also a legal requirement for compliance with GDPR provisions on automated decision-making,

- including profiling, as outlined in Article 22. Ensuring explainability in the processing of personal data is therefore essential for both user trust and regulatory compliance.
- End-user needs. A great part of ITHACA end-users is intended to be vulnerable groups with various special needs that have to be taken into account. The design of data access and interfaces need to be guided by the end-user needs and limitations. This includes accessibility features of the user interface that lowers the barrier to use the ITHACA platform and thus enables the participation of vulnerable groups.
 - Issue reporting. Users should have the possibility to report and give feedback if there are any issues or problems with the data management.
 - Identity authentication. The identity of a user should be authenticated by the system, so that one user acts under the same account and is known to the system. A full verification process of a person would be preferable, in order to avoid fake accounts and misuses through anonymous accounts (see also the updated version of D1.2 that integrated an analysis of the EU Digital Identity Wallet). However, the validation process needs to be in line with capabilities of the end users considering the vulnerability restrictions as identified in the requirements analysis of WP2

On request, further investigation has been undertaken regarding newer advancements in this field incorporated by the concept of the **EU Digital Identity Wallet**. This concept strongly refers to similar user requirements, like user control, usability, transparency, ease of use, and smooth onboarding. Overall requirements include user-centricity, interoperability, privacy-by-design, and security-by-design. While most features and requirements of such EU Digital Identity Wallets strongly overlap with these requirements outlined above, there are also differences. A key feature of Wallets is the authentication of persons, often performed with two-factor authentication and registration at the local government. However, ITHACA focuses on anonymous participation and vulnerable groups. Some users lack the digital skills for Wallet authentication, while others prefer to interact without disclosing their real identity. Thus, recent developments in the field of Digital Identity Wallets cannot be applied in ITHACA.

The implementation of the ITHACA platform took into account parts of the user data and requirements included in the PIMS model mentioned above. Regarding the user data, the platform focused on taking up **personal data** (e.g. name, username, password, role, and language), **participation data** (e.g. created comments, votes, likes, and ratings), and **technical data** (e.g. system settings or log data). Furthermore, requirements have been taken up related to user **data management and storage**. Users can authenticate and log in, and all personal and participation data is stored securely and protected. The system's underlying technology also provides robust data interoperability features. Moreover, all data processes are designed for ease of use, ensuring a high level of usability for the user. Features related to **user control, permission management and transparency** are implemented on a basic level. Details of the implementation are described in D3.3 and D3.4.

Ad (Ib)

There is a perceived opportunity to further optimize the initial scope of the research and development efforts. Specifically, an exploration of enhanced granular control over permission management and data transparency could yield significant benefits. By developing the capability to individually align data types with platform functionalities, a more detailed understanding of data processing could be achieved. Additionally, incorporating further iterative engagement with end-users through

supplementary pilot studies would offer richer insights into the overall attainment of the trust and accessibility objectives.

Ad (IIa)

ITHACA puts a focus on supporting vulnerable and marginalised groups. The PIMS model reflects this requirement by addressing (a) data protection and security of user and participation data, and (b) accessibility and usability. The first aspect refers to the fact that personal information should be safe and not be disclosed without the agreement of the user. The second aspect refers to the fact that people with a lack of skills (e.g. digital skills, language skills, etc.) should not be excluded from using the platform.

The ITHACA implementation addresses both security and accessibility. Data security is ensured through the utilization of dedicated technological components, while accessibility is achieved by offering intuitive login procedures and native language support. A simplified level of data control has been implemented, in order not to overwhelm vulnerable people with too much complexity and too many (micro-)decisions.

Ad (IIb)

On the one hand, mechanisms should be in place to allow especially vulnerable groups to decide on the purposeful use and external visibility of their information. On the other hand, too much control and micro-decisions might overwhelm people, especially if they have reduced digital or language skills. Hence, a good balance between a simple and easy-to-use user interface and possibilities for data control is important. However, as noted above, a final pilot study to investigate this case could not be undertaken after the implementation phase.

Ad (III)

The PIMS concept is a human centric approach to personal data (EDPS, 2020). Users should be empowered to easily get insight and control of their data stored in a PIMS where their data should be stored in a secure way. However, simply giving vulnerable people more control over their data isn't helpful and could even be harmful. It often forces them to spend extra mental energy figuring out complex settings, adding a burden to people already struggling with systemic unfairness (see D2.1). For individuals facing disadvantages, the lack of understanding could, for example, contribute to decision fatigue and heightened anxiety.

For vulnerable users, a PIMS must assume higher baseline risk (beside the requirements listed above):

- Default to minimal data collection and sharing ('data minimization by design').
- Use strict, conservative default permissions: no third-party access unless explicitly and clearly granted.
- Users should be able to opt-in or customize the purpose and visibility of their data elements—such as controlling comment usage for summaries or the public display of ratings—if they choose.
- To ensure transparent data use, the PIMS must provide users with a human-readable overview of their data, logically organized into meaningful categories such as health, housing, and employment.
- For explainability, the PIMS must ensure users can easily understand the data, its significance, and how the system utilizes it. The PIMS should explain personal data in user terms (everyday language), sharing decisions and consequences (e.g., what the recipient will be able to see and do with the data), automated processing and decisions (e.g., your

comment was flagged), and controls and options (e.g., clearly mark options for high-risk situations).

- In addition, people with disabilities face systematic barriers in authentication, consent flows, and understanding privacy risks. Therefore, PIMS design requirements also need to support accessible, inclusive interaction (disability, language, low digital literacy). For example, PIMS must comply with WCAG 2.x accessibility standards like screen-reader friendliness, keyboard navigation, and high contrast (W3C Web Accessibility Initiative¹).
- With respect to interoperability, the PIMS should store and exchange personal data using open, well-documented formats and schemas, connect to external services via standard, well-audited protocols, and should interoperate with digital identity wallets like the EU Digital Identity Wallets.

6.8 Ethical AI metrics and gaps tackled by ITHACA

Ad (Ia)

Fairness, PPML, Cybersecurity and Explainability are important aspects of the evaluation of the Trustworthiness of AI Tools. Fairness and PPML tools were developed from scratch, based on legal requirements set by the EU AI Act. An open-source Cybersecurity tool has been employed and is ready to be integrated into the ITHACA platform. Double degree of explainability has been implemented in the form of the two different visualization tools. A more detailed one aimed at the moderator, who is assumed to have at least some technical knowledge and a much simpler one aimed at the user providing visual cues about the fulfilment of the different Trustworthiness criteria (see also chapter 5 for a holistic algorithmic impact assessment during ITHACAs phase 2 evaluations).

Ad (Ib)

The fairness metrics may not be perfectly understood by end users, since it is especially difficult to explain the full concept and underlying equations used for the computation of the metrics.

Ad (IIa)

Multiple fairness metrics are employed to ensure that marginalized groups are included. Impartiality, equity and equality are guaranteed and tracked throughout the life of the platform, by the developed AI Fairness Tool.

Ad (IIb)

Currently, the platform considers only two sets of users—vulnerable and not-vulnerable—due to data-retention policies and limited availability of detailed user information. While this binary categorization is sufficient for the majority of Trustworthiness metrics, it restricts the applicability of more granular fairness metrics described in D1.3 that rely on richer subgroup distinctions. This deviation emerged as a necessary compromise to ensure compliance with privacy constraints and ethical data-handling requirements. Pilot studies confirmed that, despite this limitation, the binary approach still provided meaningful insights into the inclusion of marginalized users but also highlighted the potential value of finer-grained group analysis. As a result, future iterations of the platform may explore privacy-preserving methods that would enable a more detailed fairness evaluation without compromising user data protection.

¹ <https://www.w3.org/WAI/standards-guidelines/wcag/>

Ad (III)

The explainability of the models is enhanced via the visualization tools developed, however it is not objectively assessed via specific metrics. This was beyond the scope of the current project, however we deem it an important aspect of AI evaluation tools and it should be included in a future version of the platform.

7. Recommendations

This deliverable closes the ITHACA evaluation loop and constitutes the last systematic assessment of the platform within the project lifetime. No further pilot iterations are foreseen; instead, the focus shifts to consolidating the Phase 2 outcomes and identifying concrete improvements that can realistically be implemented in the short term, alongside medium-term guidance for future deployments and extensions of the platform. The recommendations below build on the KPI framework of D4.1, the Phase 0 and Phase 1 findings reported in D4.2, and the Phase 2 evaluation and Algorithmic Impact Assessment (AIA) presented in this report.

Wherever possible, recommendations are explicitly tied to the KPIs and thresholds defined in D4.1 (e.g. $\geq 99.5\%$ uptime, ≤ 2 seconds response time, zero safety incidents, $\geq 80\text{--}90\%$ satisfaction for key UX and AI indicators), and to the concrete gaps identified in Phase 2 (e.g. explainability, coverage of minority views, gamification fairness, governance procedures).

7.1 Technical robustness and performance

From a technical standpoint, the Phase 2 results indicate that the platform meets its core stability and performance targets: uptime during the pilot windows remained within the $\geq 99.5\%$ threshold, response times for core journeys stayed near or below the ≤ 2 s target, and Gate A tests showed that the platform can technically sustain the target concurrent load with stable latency and error rates. At the same time, recurring minor issues (e.g., sporadic logout behaviour, non-functional refresh in specific views, technical errors in certain social features) show that reliability still requires deliberate attention.

Short-term technical recommendations are therefore as follows:

- Stabilise critical paths first. Before deployment at larger scale, the consortium should treat the Phase 2 “bug list” as a closure checklist, prioritising: (a) session continuity (minimising unexpected logouts), (b) correctness of post, reaction and reporting flows, and (c) functional integrity of social and gamification actions (e.g., refresh, “Unfriend”). These improvements directly support the System Uptime, Error Rate and Task Completion Rate KPIs by reducing recoverable but trust-eroding failures.
- Operationalise performance monitoring. The performance profiles established in Gate A tests should be translated into a lightweight operational monitoring setup (e.g. dashboards tracking uptime, p50/p95 latency, error counts and RPS against the existing SLOs). This will allow host cities or operating organisations to ensure that the $\geq 99.5\%$ uptime and ≤ 2 s response targets remain realistic under real civic loads, and to detect regressions early.
- Preserve AIA-aware performance checks. The combined monitoring of flip-rates and minority-view coverage under load, introduced in Phase 2, should be retained as part of any

substantial update or scaling. Specifically, each major release should be verified against a small, fixed AIA test set to ensure that latency optimisations do not introduce new fairness or stability issues in AI outputs.

These steps can be implemented without further pilot iterations and will help to lock in the current positive performance profile as the baseline for exploitation.

7.2 User experience, accessibility and inclusiveness

The Phase 2 evaluation confirms that the platform is usable in practice: the majority of participants in both pilots managed to complete the core journeys, and the overall usability perception is clearly above negative. However, it does not yet fully reach the ambitious D4.1 threshold of $\geq 80\%$ positive usability and $\geq 85\%$ “easy to learn” ratings, particularly for older and less digitally confident users, and not all accessibility issues identified in earlier phases have been fully resolved.

On this basis, the main UX and accessibility recommendations are:

- **Simplify first-time onboarding and navigation.** The platform should offer a more guided first login for citizens, ideally in the form of a short, optional walkthrough explaining the three or four essential actions (find a topic, read a thread, consult the summary, post and react). This can be implemented with lightweight overlays or “first-time tooltips”. For moderators, the back-office console should provide a clear entry point to key dashboards (topics, flags, AI tools), reducing cognitive load when switching tasks. These improvements directly support the Ease of Learning and Usability Score KPIs.
- **Reduce visual and cognitive overload.** In line with Phase 2 feedback, the platform should aim for fewer competing elements per screen, clearer grouping of controls and more consistent feedback messages (success, error, pending review). For vulnerable users in particular, shorter text blocks, clearer headings and a more obvious distinction between “mandatory” and “optional” actions will make journeys less demanding. This contributes to better Accessibility Compliance and Inclusiveness Compliance without requiring structural re-design.
- **Complete and surface accessibility features.** The existing accessibility controls (font size, contrast, possibly keyboard navigation cues) should be made more prominent and consistent across all views, with brief, plain-language labels (e.g. “Larger text”, “Higher contrast”). Future versions should aim for full WCAG 2.1 AA compliance on the main participation flows; in the shorter term, prioritising colour-contrast, focus indicators and screen-reader labels for interactive elements will already reduce barriers. Crucially, the Braşov pilot revealed a distinct ‘adaptation curve’ for assistive-technology users, who initially faced blocking issues (26% failure rate) but achieved 100% success after 2–3 sessions. To flatten this curve, future deployments should include a dedicated ‘Accessibility Warm-up’ module, a simplified, sandbox environment where users can test their screen readers or keyboard setups against the platform’s navigation logic before entering a live debate. This would bridge the gap between technical compliance and practical fluency.
- **Maintain inclusive recruitment and communication practices.** The citizens’ jury methodology and vulnerability-aware sampling used in WP2 and WP4 should be retained as a template for future deployments, ensuring that people from low-income, low-literacy, migrant or disability groups are actively invited and supported, rather than assumed to appear “organically”. Linked to this, all key user-facing texts (registration, consent, instructions) should offer a brief, plain-language summary alongside the full legal or technical wording.

Collectively, these recommendations aim at raising the proportion of users who can use the platform smoothly, with minimal assistance, towards the intended 80–85% threshold, while embedding accessibility and inclusiveness as non-negotiable defaults.

7.3 AI components, fairness and explainability

The AIA conducted in Phase 2 shows that the platform’s AI components (summarisation, toxicity detection and fairness/risk indicators) are technically promising and do not exhibit obvious group-based unfairness in the tested cases. At the same time, they fall short of the project’s ambitions around explainability, coverage fairness and organisational readiness for AI-supported decision-making.

Three areas are particularly important:

a) Summaries as traceable, pluralistic views

- The Phase 2 results suggest that AI summaries are widely used as orientation tools but are sometimes perceived as too compressed and prone to under-represent minority or dissenting views. Short-term improvements should therefore focus on:
 - adding explicit markers of dissent (e.g. a short clause such as “Some participants disagreed, arguing that ...” when such content is present);
 - displaying simple context indicators (e.g. “Summary of 24 posts in this thread”); and
 - providing one-click links from each summary bullet to representative underlying comments.
- Ensure Linguistic Equity in Generative Models. The sharp rejection of AI summaries by Braşov moderators due to poor translation quality ('incoherent', 'dust') highlights that statistical fairness is irrelevant if the output language is broken. For all non-English deployments, the system must enforce a 'Native-Level Quality Gate'. This means fine-tuning the underlying Large Language Model (LLM) on local administrative corpora or switching to region-specific commercial models that guarantee professional-grade syntax. Without this, AI tools are effectively discriminatory, offering automation benefits only to English-speaking staff while burdening local teams with manual correction tasks.

These changes will strengthen coverage fairness and support the AI Explainability KPI by making summarisation logic more visible and auditable to both citizens and moderators.

b) Moderation assistance with explicit justifications

- For AI-assisted moderation, Phase 2 shows acceptable alignment with human decisions in English/Slovak contexts, but significant inconsistency in Romanian, particularly regarding identity-swapped fairness tests (where 60% of moderators reported unequal treatment). To move towards the $\geq 80\%$ AI Explainability target, each AI-flagged item should be accompanied by:
 - a short category label (e.g. “possible hate speech”, “insult/harassment”, “spam / off-topic”);
 - an indication of severity (low/medium/high risk); and
 - a link or tooltip to the relevant moderation guideline or policy section.

- **Localised Toxic Vocabulary:** The moderation engine must be updated with a curated library of regional slurs, dialect-specific vulgarities, and context-dependent keywords (e.g., Slavic-influenced terms in Romania) to prevent 'under-moderation' safety gaps where local toxicity evades detection.

This “explanation stub” can be implemented with minimal UI changes, but it will make a substantial difference to moderators’ capacity to evaluate and contest AI suggestions and to explain decisions to citizens.

- Human-in-the-loop principles should be formalised in documentation: AI suggestions must remain advisory, with clear signals that final responsibility lies with human staff, and with configurable thresholds for “auto-hide” or “needs review” states, in line with the AI Trustworthiness KPI (zero safety incidents and explicit human oversight).

c) Making fairness and risk tools usable

- The fairness and PPML tools developed in Task 5.1, and the associated visualisation dashboards, provide a rich basis for checking equality of treatment and privacy risk. However, Phase 2 feedback indicates that non-expert users may find the underlying metrics difficult to interpret. For municipal staff, these tools should therefore be presented in a layered way:
 - a high-level “traffic light” summary (e.g. green/amber/red for fairness and privacy risk);
 - a small number of key metrics with simple descriptions; and
 - optional access to the full metric set for expert users.
- As ITHACA will not have another evaluation cycle, it is important that the consortium documents a minimal AIA checklist for future instances of the platform: for example, a short protocol covering (a) fairness checks on representative datasets, (b) flip-rate monitoring under load, and (c) periodic moderator feedback on AI behaviour. This provides a practical route to maintaining AIA discipline beyond the project.

7.4 Gamification and meaningful engagement

The UPAT evaluation shows that the gamification module has a high motivational potential but that current mission, points and leaderboard logic fall short of the intended fairness and transparency standards. Participants enjoyed the playful elements but expressed confusion about why and when they earned points, and they questioned the fairness and usefulness of the leaderboard.

Given that this is the last evaluation cycle, the recommendations focus on making gamification safe and interpretable enough to be retained as an optional engagement layer in future deployments:

- Stabilise and document scoring rules. XP and ranking logic should be fixed and documented so that: (a) all missions are realistically achievable within typical participation windows; (b) users can see a simple “How XP works” explanation; and (c) the leaderboard is updated predictably and reflects meaningful contribution rather than artefacts of bugs or edge cases. This directly addresses perceived fairness and supports broader trust in the platform.
- Prioritise “fair play” over competition. As recommended by T4.5, gamification in civic contexts should encourage constructive, respectful contributions, not merely volume. Badges and missions should therefore privilege behaviours such as reading diverse viewpoints, reporting toxic content responsibly, or up-voting constructive ideas, rather than only counting raw post numbers. Toxic content should never be rewarded, even indirectly.

- Avoid over-gamification of sensitive topics. Municipalities using the gamification layer should receive guidance on when and how to enable it (e.g. suitable for exploratory agenda-setting, less appropriate for highly polarised or legally sensitive consultations). A simple configuration profile (“light”, “standard”, “off”) can give cities practical control without redesigning the module.

These measures can be implemented without rerunning pilots and will help ensure that gamification increases engagement without undermining inclusiveness or perceived legitimacy.

7.5 Governance, policy and capacity-building

A recurrent message from Phase 2, particularly from municipal staff and the T4.5 concept revision work, is that technical features alone are not sufficient. Sustainable, trustworthy deployment of ITHACA depends on clear governance rules, plain-language communication with users and targeted capacity-building for municipal employees and other key actors.

Key recommendations in this area are:

- Formalise internal data governance rules. Each deploying city should adopt a short internal policy note specifying: (a) who can access which logs and exports; (b) when and how identifiers must be masked or pseudonymised before sharing screenshots or quotes; (c) retention periods and deletion procedures; and (d) acceptable use of platform outputs in public communication. This will operationalise the Security Compliance and Data Processing Reliability KPIs and respond directly to moderators’ concerns about re-identification in small communities.
- Offer plain-language consent and privacy explanations. In line with T4.5 policy recommendations, all legal and GDPR-related texts should be accompanied by a brief, non-technical summary explaining what data is collected, for which purpose, for how long and with which safeguards. This is essential for inclusiveness and for enabling vulnerable and marginalised groups to give truly informed consent.
- Develop AI literacy and moderation training. The project should leave behind a concise training package (slide deck and short guide) for municipal staff on: (a) how the AI tools used in ITHACA work conceptually; (b) how to interpret fairness and risk indicators; (c) how to maintain human oversight; and (d) how to explain AI-supported decisions to citizens. This responds directly to the need identified in T4.5 to build municipal capacity to “use AI tools for civic participation correctly, make better choices and give better advice to citizens”.
- Keep the AIA perspective alive. Finally, the AIA templates and findings developed in WP4 and WP5 should be integrated into any future governance or procurement documentation around the platform (e.g. as annexed checklists for new AI features or third-party tools). This ensures that fairness, transparency and robustness remain core criteria as the platform evolves beyond the scope of the current Grant Agreement.
- Secure the 'Fair Play' of the Democratic Process. Braşov moderators identified the risk of 'gaming' (coordinated voting/spam) as a critical barrier to trust. Governance rules must therefore explicitly define the identity verification level required for different types of participation (e.g., open reading vs. verified voting). Technical safeguards should be implemented to detect and flag coordinated behaviour (bot-like patterns), ensuring that the platform adheres to the 'one-person-one-vote' principle essential for municipal legitimacy.

7.6 Recommendations for future research and transfer

While this deliverable marks the end of formal evaluation activities within ITHACA, the work done in WP4 and WP5 opens several avenues for further research and transfer:

1. Comparative studies of deliberation quality. As suggested in T4.5, future work could compare the quality of group discussions on ITHACA with other online platforms and with traditional town-hall or jury processes, focusing on inclusiveness of viewpoints, civility and perceived legitimacy.
2. Privacy-preserving fairness evaluation. Given the current binary “vulnerable / non-vulnerable” distinction adopted for ethical and legal reasons, future research should explore privacy-preserving techniques (e.g. secure aggregation, synthetic data, differential privacy) that would allow more fine-grained fairness analysis without compromising GDPR compliance.
3. Transferability to other civic contexts. The evaluation protocols, KPI framework and AIA methods developed for ITHACA can be adapted to other civic engagement platforms and domains (e.g. mobility planning, climate assemblies). For future projects, we recommend reusing: (a) the KPI table from D4.1 as a starting point; (b) the phased evaluation approach (prototype → controlled testing → pilots); and (c) the integration of AIA cross-checks into performance testing.
4. Cross-Cultural AI Auditing. The divergence between Martin (high alignment) and Braşov (low alignment) moderators suggests that AI models behave differently across languages. Future research should prioritize multilingual audit protocols, specifically testing how 'fairness' degrades when models process low-resource or complex-morphology languages compared to English.

In summary, the Phase 2 evaluation shows that ITHACA has achieved a solid level of technical maturity, usability and ethical grounding, in line with most of the KPIs defined in D4.1, while transparently documenting where targets have only been partially met. The recommendations above provide a concrete roadmap for consolidating these gains and for guiding municipalities and future initiatives in deploying, governing and extending AI-enabled civic participation platforms in a safe, inclusive and trustworthy manner.

8. Conclusions

Phase 2 constituted the final evaluation cycle of the ITHACA platform and therefore the last opportunity within the project to test its technical robustness, user experience and AI components against the KPI set defined in D4.1. The analysis combined controlled stress tests, multi-site pilot data (Martin and Braşov), and the focused gamification study at UPAT, and examined each KPI in terms of its target/threshold, the available evidence, and whether this target was met, partially met or could not be meaningfully assessed. Table 15 below provides a consolidated overview of all D4.1 KPIs, the corresponding thresholds and how they were addressed in Phase 2. It should be read as the formal summary of the project's KPI-based performance at closure, rather than as a planning tool for future iterations.

Table 15. KPIs, thresholds and Phase 2 evaluation outcome

KPI	D4.1 target / threshold	How it was addressed in Phase 2 (final status)
System Uptime	Maintain high availability; D4.1 performance section specifies $\geq 99.5\%$ uptime during operation.	During all Phase 2 windows (Martin, Braşov, UPAT sessions) there were no prolonged outages or incidents where participants could not access the platform beyond short, transient glitches and scheduled maintenance. From the evaluation point of view, uptime stayed within the intended 99.5% threshold for the pilot periods. Status: Met for the observed period (full-year monitoring still outside project scope).
Response Time	Core interactions processed within ≤ 2 seconds (user actions should feel responsive).	Gate-A and subsequent performance tests showed typical response times within the 2 s budget for core journeys (login, dashboard, thread view, posting). In Martin and Braşov, no systematic complaints about “slowness” appeared in surveys or focus groups. Status: Met in tested scenarios; continued monitoring recommended.
Error Rate	D4.1 qualitative target + performance criteria: ≤ 3 critical and ≤ 10 minor errors per month; no blocking failures in core tasks.	No critical failures of core journeys were reported in Phase 2. Recurrent minor issues were observed (e.g. non-functional page refresh in gamification module, “Unfriend” action returning technical error, occasional unexpected logout). This affected experience but were recoverable and logged for fixes. Status: Critical error threshold met; minor errors above ideal target in some modules (gamification), but contained and documented.
Load Capacity	Platform must sustain ≥ 500 concurrent users (synthetic load) while keeping performance within targets.	Gate-A stress tests simulated 500+ concurrent users posting and voting; latency and error rates remained within acceptable bounds and AI outputs stayed stable on fixed test sets. Real pilot usage never reached this concurrency (small, controlled samples). Status: Met in synthetic tests; not yet field-validated at true city-scale.
AI Trustworthiness	Zero safety incidents attributable to AI (no harmful outcomes caused by summaries/moderation/gamification logic).	Across Martin, Braşov and UPAT, no cases were reported where AI outputs led to harm or unsafe outcomes. Moderators remained in control of final decisions, and problematic content was still handled by humans. Some trust-eroding bugs (esp. in gamification) were reported, but not safety incidents in the strict sense. Status: Target met (0 safety incidents).
AI Accuracy	AI-driven features should achieve high task-level accuracy (alignment with expected outcomes); D4.1 does not specify a numeric percentage.	In Martin, AI suggestions matched human choices. In Braşov, 60% of moderators reported that the AI did not detect local/regional toxic vocabulary. Status: Met in English/Slovak; Not in Romanian (inconsistent fairness).

KPI	D4.1 target / threshold	How it was addressed in Phase 2 (final status)
AI Performance	AI operations (summaries, moderation suggestions) should respect the < 2 s response time budget for user-perceived interactions.	During performance tests and pilots, AI-related responses (summary display, moderation flags) appeared within the general UI response window; there were no systematic reports of “AI taking too long”. Any delays users noticed were instead linked to navigation, not to AI computation. Status: Met within the limits of the monitoring performed.
AI Accessibility	High user satisfaction with AI features; AI should be usable and helpful to non-expert users. (No numeric value set in D4.1.)	Citizens in Martin and Braşov treated AI summaries as useful orientation aids. However, Braşov moderators were not positive due to poor linguistic quality in Romanian. While usable for citizens in English/Slovak contexts, the feature was deemed professionally inaccessible for non-English staff. Status: Mixed, it is met for citizens; Not met for professional staff in non-English contexts.
AI Explainability	Users (and especially moderators) should understand why AI made a decision; D4.1 implicitly aims for a high majority reporting that they understand AI outputs.	Phase 2 shows a clear gap: moderators explicitly asked for short explanations and category labels on flagged posts, and UPAT participants considered XP and leaderboard logic opaque. Users could tell that AI is being used but not how or why specific outputs appear. Status: Not met, explainability is the main weakness identified for AI features.
User Engagement	Sustained participation (e.g. percentage of registered users active per month). No fixed percentage is defined in D4.1, but higher is better.	Phase 2 ran as short, controlled studies with small, invited samples, not as an open, long-running deployment. Many participants logged in multiple times within the 1–2-week protocol, but this is not representative of organic monthly engagement at scale. Status: Not meaningfully assessable in Phase 2; KPI left open for future deployments.
Retention Rate	Continued user engagement over time (e.g. users returning over weeks/months). No numeric threshold fixed.	Within the Phase 2 windows, a sizeable share of participants returned at least once; some completed all missions. However, the observation period was too short to measure true retention (e.g. 3–6 months). Status: Only short-term revisit behaviour observed; long-term retention not assessed.
Feature Adoption Rate	High adoption of key features (reading, posting, reacting, reporting, summaries, accessibility controls). Often interpreted as ~70%+ for “key” features.	In Martin and Braşov, reading, posting and reacting were used by almost all participants. Summaries were consulted by many, especially when returning to threads. Reporting and accessibility controls were used mainly when prompted in missions. Gamification features (missions, XP, badges) were heavily used in the UPAT lab but are not yet integrated into everyday pilots. Status: Core features close to target; advanced/accessibility features below target and require stronger visibility and onboarding.

KPI	D4.1 target / threshold	How it was addressed in Phase 2 (final status)
Usability Score	Overall usability rating should be predominantly positive (e.g. majority of users rate usability above neutral). D4.1 leaves this qualitative.	In Martin, usability was acceptable but not frictionless. In Braşov, citizens reported near-ceiling usability scores consistently across all six visits, with 100% retention intent. This indicates the platform is highly usable for digitally confident populations, while remaining 'learnable' for others. Status: Met (High) for high digital profiles; Partially met for low-digital profiles.
Ease of Learning	Most users should rate the platform as easy to learn, especially within a short familiarisation period.	After minimal instruction, most participants in Martin and Braşov were able to log in, find topics and post. Less digitally confident users and older adults sometimes struggled with the first login, unfamiliar vocabulary and discovering features like reporting or filters. Status: Partially met, core concepts learnt quickly; first-time onboarding still too fragile for some profiles.
Task Completion Rate	High task success rate (D4.1 examples indicate targets like ≥ 80% of tasks completed without issues).	In guided Phase 2 missions, nearly all participants completed the scripted tasks (login, navigate to assigned topic, read summary, post, react, report, etc.). Drop-offs occurred mainly in optional or free exploration tasks (e.g. trying filters or advanced features without prompts). Status: Largely met for core scripted tasks; not systematically quantified but qualitatively high.
Accessibility Compliance	WCAG 2.1 Level AA compliance on core journeys.	Main flows are technically accessible. In Braşov, a distinct adaptation curve was measured: assistive-technology users initially failed (26% block rate) but achieved 100% success by Visit 4. This confirms the platform is compliant but requires a 'warm-up' period for users to master navigation. Status: Met (with learning curve condition).
Inclusiveness Compliance	Ability to adapt to different cognitive abilities, digital literacy levels and preferences, avoiding systematic exclusion.	Recruitment in Martin included older adults and less digitally confident users, who managed to participate but reported cognitive load (long texts, complex pages). UPAT results suggest that current gamification design mainly suits digitally fluent students, risking exclusion of other profiles if transferred "as is". Status: Partially met – inclusiveness considered and tested, but simplification and better tailoring still needed.
Security Compliance	100% compliance with GDPR and security standards; zero critical security vulnerabilities.	The platform operated under TLS, with pseudonymous accounts, role-based access control and a GDPR-compliant logging schema. There were no reported security incidents in Phase 2. Moderators did, however, ask for clearer organisational rules for exports, screenshots and identifier masking. Status: Technical compliance target met; organisational policy and documentation need further strengthening.

KPI	D4.1 target / threshold	How it was addressed in Phase 2 (final status)
Data Processing Accuracy	High accuracy of data handling (no systematic mismatch between what users do and what is stored/processed).	Evaluation work used platform logs, Keycloak exports and survey datasets. There were no detected systematic inconsistencies between observed user actions and recorded data. Some missing questionnaire entries (e.g. incomplete Typeform responses) were due to external tool behaviour, not platform mis-logging. Status: Acceptable; no formal numeric audit, but no accuracy problems surfaced in Phase 2.
Fairness Compliance	AI should behave fairly and ethically, with no systematic group-based bias or discriminatory patterns.	The AIA (Martin, Braşov) did not identify identity-based discrimination in the tested moderation items or summaries. The main fairness issues observed were coverage fairness (under-representation of minority/dissenting views in summaries) and perceived unfairness in gamification scoring and leaderboards. Status: Qualitatively acceptable for tested cases; coverage and game-logic fairness remain open improvement areas.
Explainability (general)	Users should understand platform outputs (not only AI but also, e.g., how scores and rankings are built).	Beyond AI-specific explainability, many participants could not tell how or why certain outputs were produced: why a post was flagged, how XP was calculated, why the leaderboard looked as it did. This reduced trust especially in the gamification module and in AI-assisted moderation. Status: Clearly below desired level; explainability is a cross-cutting gap.
Scalability Performance	Maintain full performance under peak load (latency, error rates and AI behaviour stable).	Gate-A stress tests with gradually increasing RPS showed that the platform maintained target latencies and error rates and that AI outputs on fixed AIA test sets remained stable (no increase in flip-rates, no loss of minority-view sentences). Real-life traffic in pilots stayed below these test levels. Status: Met in test environment; must be safeguarded as deployments scale.
Trust Score	Users should report high trust in the platform and its AI components (majority above neutral).	Citizens and moderators in Martin expressed conditional trust: they trust the platform more when they can check summaries against the full thread and when humans remain visibly in control of moderation. Baseline trust in institutions is relatively high, bugs and opacity in gamification slightly undermined trust in that module. No single numeric “trust index” was calculated. Status: Partially met – trust present but contingent on transparency and reliability.
Data Processing Reliability	Reliable capture and storage of data inputs/outputs, with secure, usable logs.	For Phase 2 analysis, logs and exports were sufficiently complete and reliable to reconstruct user sessions and conduct the planned evaluations. Minor gaps (missing records due to connectivity or external survey tools) did not undermine the overall dataset. Status: Largely met – data handling reliable for evaluation purposes.

Taken as a whole, the KPI review in Table 13 shows that the platform has reached a solid level of **technical maturity** by the end of the project. System uptime during all evaluation windows met the $\geq 99.5\%$ target, response times for core journeys stayed within the ≤ 2 s budget, and synthetic Gate-A tests demonstrated that the platform can sustain the specified concurrent load while keeping latency and error rates under control. No critical failures of core tasks were reported, and scalability performance and data processing reliability were confirmed both through stress tests and through the successful use of logs and exports in the analysis. In KPI terms, the technical performance and reliability targets are therefore largely met for the scope and duration of the pilots.

At the same time, Phase 2 underlines that **user experience and inclusiveness targets are only partially achieved**. Usability is clearly above a negative threshold (most users in both pilots completed their core tasks and described the platform as overall usable), but it does not yet reach the ambition of a consistently “easy to learn, easy to use” system for all profiles. Ease of learning and task completion are high for scripted journeys, yet first-time onboarding remains fragile for less digitally confident users, and some advanced and accessibility features (reporting, filters, accessibility controls) have adoption rates below the implicit 70% target. However, the Braşov data adds a positive counterpoint. For digitally confident citizens, the platform achieved ceiling-level usability scores and 100% retention intent, proving that the design is highly effective when the basic digital literacy barrier is crossed. Accessibility and inclusiveness KPIs are therefore partially met. The platform can be used by older and vulnerable participants with support, but cognitive and visual load, as well as the current design of gamification and some UI elements, still pose barriers for certain groups.

For the **AI-related KPIs**, the picture is similarly mixed. On the positive side, AI trustworthiness and safety targets are fully met: no safety incidents attributable to AI were reported, AI-assisted moderation on curated borderline items broadly aligned with human judgments, and performance tests showed that AI behaviour remained stable under load. Fairness KPIs are qualitatively acceptable in English and Slovak, but Phase 2 identified a critical fairness gap in linguistic localization. The Braşov pilot revealed that the AI moderator applied inconsistent standards to identity-swapped pairs in Romanian, and the summarizer was rejected by staff as linguistically incoherent. This highlights that 'Accuracy' and 'Fairness' are language-dependent properties that were not met for the Romanian administrative context. AI accessibility and trust are therefore best described as “moderate and conditional” rather than fully achieved, and explainability, both for AI and for game logic, remains the principal unmet KPI at project end.

In terms of **engagement KPIs**, Phase 2 could only provide partial evidence. Short-term engagement and revisit behaviour within the 1–2-week study windows were satisfactory, and key participation features (reading, posting, reacting) were widely adopted in both cities. However, by design, the pilots did not run as long, open-ended deployments, and the project could not reliably measure long-term retention, organic monthly active users, or sustained engagement at city scale. These KPIs must therefore be considered as open for future deployments: the platform has demonstrated that it can support repeated participation in structured scenarios, but its long-term engagement profile in a real municipal context remains to be established.

Finally, the evaluation confirms that **security and data-related KPIs are largely fulfilled** within the scope of the project. The platform operated with TLS, pseudonymous accounts, role-based access control and GDPR-compliant logging; no security incidents were reported, and no systematic data processing inaccuracies were detected during the analyses. At the same time, moderators' concerns about re-identification risks in small communities and about the use of exports and screenshots point

to the need for clearer organisational rules and documentation around data governance. Technically, the security and data processing KPIs are met; institutionally, they will need to be underpinned by robust local policies and procedures as the platform moves into exploitation.

In conclusion, the KPI-based assessment in Table 13 confirms that ITHACA exits the project stage with a **technically robust, functionally rich and ethically aware platform** that already meets most of its core performance, security and safety targets. The remaining gaps are concentrated in four areas: (i) usability and inclusiveness for less confident users, (ii) explainability and transparency of AI and gamification logic, (iii) fair representation of dissenting and minority views in AI summaries and game mechanics, and (iv) empirical evidence on long-term engagement beyond controlled pilots. Addressing these points through targeted refinements and clear governance frameworks, rather than major redesign, will be key for municipalities and other stakeholders who choose to deploy ITHACA as a sustainable, trustworthy civic participation infrastructure beyond the lifetime of the project.

References

- Akbarighatar, P., Pappas, I. O., & Vassilakopoulou, P. (2023). Justice as fairness: A Hierarchical framework of responsible AI principles. *ECIS 2023 Research-in-Progress Papers*, 79. https://aisel.aisnet.org/ecis2023_rip/79
- Council of Europe. (2024) Framework Convention on Artificial Intelligence. <https://rm.coe.int/1680afae3c>
- Deci, E. L., & Ryan, R. M. (2002). *Handbook of self-determination research*. University Rochester Press.
- EDPS (European Data Protection Supervisor) (2020). Personal Information Management Systems. TechDispatch 2020(3). https://edps.europa.eu/sites/default/files/publication/21-01-06_techdispatch-pims_en_0.pdf
- EP&CEU (European Parliament & Council of the European Union, 2016). Regulation (EU) 2016/679 (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
- EP&CEU (European Parliament & Council of the European Union, 2019). Directive (EU) 2019/882 on the accessibility requirements for products and services (European Accessibility Act). Official Journal of the European Union, L 151, 70–115. <https://eur-lex.europa.eu/eli/dir/2019/882/oj/eng>
- EP&CEU (European Parliament & Council of the European Union, 2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- Koutavelis, E., Rigas, E., Dragota, A., & Loi, I. (2025). Technical requirements, architecture design and 1st release: Prototype (D3.2). ITHACA Project, European Union Horizon Europe Framework Programme (Grant Agreement No. 101094364).
- OECD (2019). *Recommendation of the Council on Artificial Intelligence*. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Ryan, R. M., Deci, E. L., Vansteenkiste, M., & Soenens, B. (2021). Building a science of motivated persons: Self-determination theory's empirical approach to human experience and the regulation of behavior. *Motivation Science*, 7(2), 97-110.
- Touliou et al. (2024). Platform test and evaluation plan (D4.1). ITHACA Project, European Union Horizon Europe Framework Programme (Grant Agreement No. 101094364).
- Touliou et al. (2025). Platform tests, pilot evaluation and recommendations report (D4.2). ITHACA Project, European Union Horizon Europe Framework Programme (Grant Agreement No. 101094364).
- Zangl, M., Loi, I., Zachos, P., Bedek, M., Dimogerontakis, E., Nikolaou, C. E., Albert, D. & Moustakas, K. (2025). A multidisciplinary analysis of transparent AI-driven toxicity detection tools for civic engagement platforms. *AI & SOCIETY*, 1-18. <https://doi.org/10.1007/s00146-025-02424-5>

Annex 1: Online mini-surveys

Annex 1.1: Online mini-surveys (Participants)

A1. Baseline (pre-Visit-1; ~2 min)

Welcome! This quick 2-minute form helps us understand your starting point (experience with online discussions/AI and any accessibility needs) so we can improve the platform for everyone. There are no right or wrong answers.

Data & privacy: *Your answers are used for research to improve ITHACA and are reported in aggregate (no names in public reports). Please avoid sharing sensitive personal details. You can skip any question or stop at any time. Questions: [Contact email]. Data rights (access/withdraw): [DPO email].*

User ID (provided by researcher):

- **How often do you take part in online public discussions?**
Never / Rarely / Sometimes / Often
- **Before today, how much did you trust AI-generated summaries or content rules on online platforms?**
1 Not at all – 2 – 3 – 4 – 5 A lot or I haven't used them
- **Today, what is your main goal? (choose one)**
 - o Share my view
 - o Read what others think
 - o Understand the key points quickly (summary)
 - o Report problematic content
 - o Other: _____
- **Do you use any of these when browsing? (tick all that apply)**
 - o Screen reader (e.g., NVDA, JAWS, VoiceOver)
 - o Keyboard only (Tab/Shift+Tab)
 - o Switch device, trackball, joystick, or other pointer alternative
 - o Voice control or speech-to-text
 - o Touch only or one-hand use
 - o Larger text / zoom / magnifier
 - o High contrast or colour filter
 - o Reduced motion / animations off
 - o Captions or transcripts for audio/video
 - o Simpler layout / reading help / plain-language mode

- None of these
 - Other: _____
- **In the past, how often have you reported problematic content on online platforms?**
Never / Rarely / Sometimes / Often

A2. Post-Visit (end of each visit; 2–3 min)

One-minute check-in. After this session, tell us how it went—was it easy, satisfactory, and did anything slow you down? This helps us fix issues quickly.

Data & privacy: Responses are linked only to a study ID and used to improve the platform; no names appear in reports. Participation is voluntary—you may skip any question. Support: [Contact email]. Data rights: [DPO email].

User ID (provided by researcher):

- It was easy to do what I wanted.**
1 Strongly disagree – 2 – 3 – 4 – 5 Strongly agree
- Overall, I'm satisfied with this visit.**
1 – 2 – 3 – 4 – 5
- What slowed you down? (one sentence)**
- Would you return to use this again?**
Yes / No
- Did the platform work with your accessibility needs?**
Yes, fully / Partly / No / I don't use accessibility tools

A3. Quick Accessibility Check (targeted; ~2 min)

Accessibility check. In about 1–2 minutes, tell us if you could complete the step with your setup (e.g., screen reader, keyboard only) and what, if anything, got in the way. This helps us remove barriers fast.

Data & privacy: We collect only what we need to fix accessibility issues (no medical details). Answers are stored with a study ID and reported in aggregate. You can skip any question. Support: [Contact email]. Data rights: [DPO email].

User ID (provided by researcher):

- Could you complete [Journey X] with your setup (e.g., screen reader, keyboard only)?**
Yes / No
- What kind of accessibility problem did you meet? (tick all that apply)**
 - Hard to see (low contrast)
 - Keyboard focus jumped or got stuck
 - Button or link didn't have a clear label
 - Screen reader didn't read something
 - Motion/animation made it harder

- I couldn't continue (stuck)
 - Other: _____
- Comment** (optional short text)

A4 · End-of-Study (participants; ~3–4 min)

Final check-in (3–4 minutes). Thanks for completing your sessions! This short form asks about your overall experience across the last two weeks: ease, summaries, fairness of rules, accessibility, and what we should fix first.

Data & privacy: Your answers are linked only to a study ID and reported in aggregate (no names in public reports). Please avoid sharing sensitive personal details. Participation is voluntary—you can skip any question or stop at any time.

Questions/support: [Contact email] · Data rights (access/withdraw/delete): [DPO email].

User ID (provided by researcher):

1. **Compared to your first session, using the platform now feels...**
Much harder / A bit harder / About the same / A bit easier / Much easier
2. **Overall, I can do what I want on the platform.**
1 Strongly disagree – 2 – 3 – 4 – 5 Strongly agree
3. **I would use this platform again for local issues.**
Yes / Maybe / No
4. **The AI summaries helped me keep up with long discussions.**
1 – 2 – 3 – 4 – 5 or I didn't use summaries
5. **The summaries reflected different viewpoints, including minority ones.**
1 – 2 – 3 – 4 – 5 or Not sure / I didn't use summaries
6. **The content rules (what's allowed/not allowed) felt fair and consistent.**
1 – 2 – 3 – 4 – 5 or Not sure
7. **Compared to before the study, my trust in AI summaries/moderation is...**
Lower / About the same / Higher / I'm not sure
8. **In the last two weeks, did you report any content?**
Yes / No
If Yes: *Most decisions matched my expectation / Some did / Rarely did / Not sure*
9. **Accessibility — did the platform work with your setup?**
Yes, fully / Partly / No / I don't use accessibility tools
If Partly/No (optional): *What got in the way? (short text)*
10. **What most slowed you down overall? (one sentence)**
11. **What should we fix first? (list up to 3 short bullets)**
12. **Anything else you want us to know? (optional short text)**

Branching tips:

Q8 follow-up only if “Yes”.

Q9 follow-up only if “Partly/No”.

Treat “I didn’t use summaries” / “Not sure” as **missing** for Q4–Q6 averages.

Final mini-survey (closing statement for the participant’s last A2)

Thank you! Your feedback will directly guide what we fix first and how the city uses ITHACA.

What happens next: *we’ll summarise the results and, where possible, share updates on improvements.*

Your choices & rights: *you can withdraw or request access/deletion of your responses—email [DPO email].*

If you’d like to join a short online discussion group about next steps, please contact [Contact email].

Annex 1.2: Mini-surveys for municipal demonstrators/moderators

(short, plain-language, and aligned with the AIA + uptake goals).

DEM-1 · Baseline (before the session, ~2 min)

Welcome! This 2-minute form helps us understand your role and today’s focus so we can tailor the session. There are no right or wrong answers.

Data & privacy: *Your responses are used to improve ITHACA and reported in aggregate (no names in reports). Please avoid sharing sensitive personal data. You can skip any question or stop at any time. Questions? [Contact email]. Data rights (access/withdraw): [DPO email].*

User ID (provided by researcher):

- **Your role**
 - Policy / Strategy
 - Communications / Press
 - Community engagement
 - Moderation / Support
 - IT / Admin
 - Other: _____
- **Have you used AI features like summaries or content rules in your work before?**
 - Yes, often
 - Sometimes
 - Rarely
 - Never
- **Today’s focus in your workflow (choose one main aim):**
 - Prepare/brief a decision or meeting

- Draft a public message/press note
 - Moderate a discussion
 - Scan a long thread quickly
 - Other: _____
- **What would make ITHACA most useful for you today?** (one sentence)

DEM-2 · Summary check (during J2, ~2–3 min)

Quick check on the AI summary. After reading the summary and skimming a few posts, please answer these 2–3 questions (≈2 minutes) about usefulness and coverage.

Data & privacy: We record your answers with a study ID only. No sensitive data is requested. You may skip any question. Support: [Contact email]. Data rights: [DPO email].

(Shown after opening the long thread & reading the AI summary)

User ID (provided by researcher):

- Did the summary help you grasp the main points quickly?**
 - 1 Not at all – 2 – 3 – 4 – 5 Very much
- Was any important viewpoint missing?**
 - No, it covered the key views
 - Yes → Which one? (one sentence)
- Would you use this summary in your work right now?**
 - Yes (as-is)
 - Yes, with a small edit
 - Not yet
- (If “small edit” or “Not yet”) **What edit would make it ready?** (one sentence)

(Feeds AIA: summary usefulness + coverage, incl. minority view)

DEM-3 · Moderation check (during J3, ~2–4 min)

(Repeat this block for each of the 3–6 “tricky” posts you review)

About these examples. You’ll review a few borderline posts and say how you’d handle them. Some items may contain strong opinions. Please skip any item you’d rather not view.

Data & privacy: Your decisions are stored with a study ID and used to assess consistency and fairness. No names will appear in reports. Skip anything you prefer not to answer. Support: [Contact email]. Data rights: [DPO email].

User ID (provided by researcher):

- Item #[X] decision you would take:**
 - Keep (allowed)
 - Remove (breaks rules)

- Not sure
- Did the system’s decision match yours?**
 - Yes
 - No
 - System decision not shown
- Was anything about the wording (e.g., mentions of a group) influencing the decision unfairly?**
 - No
 - Yes → *Briefly say why:* _____

(Optional quick pair test for fairness)

- **Two near-identical posts with different group words should get the same outcome. Did they?**
 - Yes / No / Not sure

(Feeds AIA: moderation accuracy + consistency across identity mentions / counterfactual pairs)

DEM-4 · Privacy & security quick check (~1 min)

Your expert view. This 1-minute check asks whether anything should be masked before sharing and if there are security concerns in routine use.

Data & privacy: *We collect your professional opinion only; no personal data needed. Aggregated findings may appear in the project report (no individual attribution). Support: [Contact email]. Data rights: [DPO email].*

User ID (provided by researcher):

13. Anything here you’d want hidden or masked before sharing internally/externally?

- 13.1** No
- 13.2** Yes → *What? (short text)*

14. Any worry that the system could be tricked or misused in your context?

- 14.1** No
- 14.2** Maybe / Yes → *What safeguard would reassure you? (short text)*

(Feeds AIA: privacy & security notes)

DEM-5 · Organisational uptake (~2 min)

How would this help your workflow? In ~2 minutes, tell us where ITHACA would fit (agenda, comms, moderation, briefings) and what would make it “ready now”.

Data & privacy: *Responses are analysed in aggregate to inform adoption plans; no names in reports. Participation is voluntary. Support: [Contact email]. Data rights: [DPO email].*

User ID (provided by researcher):

Where would this help your workflow today? (tick all that apply)

- a. Agenda-setting
- b. Drafting a message/press note
- c. Moderating a thread
- d. Preparing a briefing for leadership
- e. Public-facing summary or dashboard
- f. Other: _____

How ready is this for routine use in your team?

- g. 1 Not ready – 2 – 3 – 4 – 5 Ready now

The one change that would make it “ready now” is... (one sentence)

Would you act as an internal demonstrator/advocate once those changes are made?

- h. Yes / Maybe / No

DEM-6 · Wrap-up (final 1 min)

Final minute. Please share your Top-3 fixes and any last notes. This helps us prioritise changes.

Data & privacy: *Collected with a study ID and reported in aggregate. You may skip any question. Support: [Contact email]. Data rights: [DPO email].*

User ID (provided by researcher):

- **Top 3 fixes we should do first** (write 3 bullets or rank if your form supports it)
- **Anything else we should know?** (optional short text)

Suggested branching

1. Show **DEM-2** only after the summary is viewed.
2. Repeat **DEM-3** block for each tricky post.
3. **DEM-4** always shown (quick).
4. **DEM-5** always shown (uptake).
5. **DEM-6** at the end.

DEM-6 · Closing statement (show on “Thank you” screen)

Thank you! Your input will directly shape what we fix first and how the city can use ITHACA.

What happens next: we’ll compile a short summary of today’s session and share it with you for confirmation. If you’d like to add anything, reply to [Contact email].

Your choices & rights: you can withdraw or request access/deletion of your responses by emailing [DPO email].

Annex 2: Focus group material

Annex 2.1: Phase 2 Focus group guides

Participants (End-Users)

Format: online (60–75 min) • **Platform under discussion:** <https://ithaca.simavi.ro/>

Size: 6–8 people (mix of Guided/Free-mode users; include at least 1–2 who use accessibility tools)

Goal: learn what worked/failed in real use; assess perceived usefulness/fairness of summaries & rules; capture accessibility barriers; prioritize fixes.

1) Prep (1–2 days prior)

- Recruit & confirm 6–8 who completed ≥3 sessions (any mode mix).
- Send calendar invite with video link + reminder to have their login handy (no need to log in during FG).
- Prepare:
 - o Slides (or one-pager) with agenda and ground rules.
 - o A short **demo thread** (Pilot-Eval) + 2–3 **borderline** items (for show-and-tell only; no personal data).
 - o Dot-voting tool (Mentimeter/Zoom poll/“type 1–5 in chat”).
 - o Note-taking template (provided in §6).

2) Roles

- Facilitator:** runs session; keeps time; asks probes.
- Note-taker:** captures verbatim quotes, friction points, decisions.
- Tech helper:** admits participants; manages chat/polls; handles AV issues.

3) Consent & ground rules (script, 3 min)

“Thanks for joining. Today we’ll discuss your experience using ITHACA. We’ll keep **anonymous notes**—no names in reports. You can **skip** any question and **leave anytime**. Please be respectful; one person at a time; focus on your own experience.”

(Ask for verbal “OK”.)

4) Agenda (60–75 min)

0–5 min — Warm-up

- Round-robin: name (first name or alias) + one thing you tried on the platform.

5–20 min — Ease & friction (core journeys)

- “What did you try most often (post, read a summary, react, report)?”
- “Where did you **get stuck**? Walk us through the step.”
- Probe: “If you had 10 seconds to tell a friend how to do that step, what would you say?”

20–35 min — AI summaries (usefulness & fairness)

15. Show a **demo thread** + its **AI summary** (screen share).
16. “If you’d seen this summary first, would it have helped you catch up?” (1–5 quick poll)
17. “What’s **missing** from the summary—especially a minority or less popular view?”
18. “One sentence you’d add to make it fairer/clearer?”

35–45 min — Content rules & reporting

Show **2–3 borderline items** (neutralized examples).

“Would you report this? Why/why not?”

“Do the rules feel **consistent** across different kinds of speech?”

45–55 min — Accessibility

- “Who used **screen readers/keyboard/zoom/contrast**? What **barrier** did you hit?”
- “What **one change** would remove the biggest barrier for you?”

55–70 min — Priorities (Top-3 fixes)

6. Ask everyone for **one must-fix**; create a shortlist.
7. **Dot-vote** (each gets 3 votes).
8. Read out Top-3; confirm they match the room’s view.

70–75 min — Close

- 1) “Anything we missed?”
- 2) Explain next steps & thank them.

5) Question bank (use as probes)

- “What slowed you down the most?”
- “What felt unfair about a summary or decision?”
- “What would make you come back to use it again next month?”
- Accessibility: “Did keyboard focus/labels/contrast ever block you?”

6) Note-taking template (copy/paste)

- **Participants:** P1...P8 (list tools: SR/KB/Zoom/Contrast)
- **Friction points:** journey • step • description • example text/URL
- **Summaries:** usefulness poll avg; missing view(s); suggested sentence
- **Moderation:** item → report/ignore rationale; consistency comments
- **Accessibility:** barrier type; page/element; suggested fix
- **Top-3 fixes (with short why):** 1) ... 2) ... 3) ...
- **Quotes:** “[short verbatim] — P#”

7) After the session (same day)

- Clean up notes (anonymized) → 2-page summary.
- Share Top-3 fixes with owners (internal).
- File any accessibility barriers to devs (short ticket with page/element).

Demonstrators (Municipal Staff)

Format: online (60–75 min) • **Audience:** policy, comms, moderators, IT/admin

Goal: capture **organisational uptake** (where/how they would use ITHACA) and collect **AIA evidence** in their words: summary coverage (incl. minority views), moderation fairness/consistency, privacy/security expectations.

1) Prep (1–2 days prior)

- Recruit 4–6 staff covering **policy/comms/moderation/IT**.
- Ensure everyone can sign in (though session can run from facilitator’s screen).
- Prepare links: **one long thread** (Pilot-Eval) + **3–6 borderline posts** + **2–3 counterfactual pairs** (identity token swap).
- Have AIA Sheet placeholders open (for quick capture).

2) Roles

- **Facilitator:** workflow framing, AIA probes, timekeeping.
- **Scribe:** fills **uptake examples** and **AIA notes** live.
- **Tech helper:** screen share, polls, chat.

3) Consent & framing (script, 3 min)

“Thanks for helping us test **how ITHACA fits your workflow**. We’ll take anonymous notes; skip anything you prefer. Today we’ll check a summary, a few borderline items, and discuss privacy/security. We’re looking for **practical adoption** and any fairness/safety concerns.”

4) Agenda (60–75 min)

0–5 min — Intros & workflow pick

7. Round-robin: role + one task where public input matters (agenda, briefing, comms, moderation). Pick **one scenario** to anchor discussion.

5–25 min — Summaries in workflow (AIA: coverage)

- Open **long thread** + **AI summary** (screen share).
- **Utility:** “Would this save you time for this scenario?” (1–5 quick poll)
- **Coverage:** “What **key view** (esp. minority) is **missing**? Quote one post you’d expect represented.”
- **Adoption:** “Where would you paste this? (briefing line, agenda note, public update) What small edit would make it ‘ready now’?”

25–45 min — Moderation consistency (AIA: fairness & stability)

- Show **3–6 borderline items** (one by one).
 - “Your decision under your policy (Keep/Remove/Unsure)? Why?”

- “Does the system’s decision match what you’d expect?”
- **Counterfactual check (2–3 pairs):** near-identical sentences differing only by group word.
 - “Should these have the **same** outcome? Did they?”
- Capture any **system-human mismatches**, especially where **identity terms** appear.

45–60 min — Privacy & security (AIA)

- **Privacy:** “What would you **mask** before sharing (e.g., names, IDs, quotes)?”
- **Security:** “Any risks that would stop routine use (tricking the system, spam, prompt/format injection)?”
- “What safeguard/policy would reassure you?” (e.g., audit trail, export watermark, rate limits)

60–70 min — Organisational uptake

- “Where exactly would this help **next month?** (agenda, comms, moderation, leadership briefing)”
- “What one change would make it **ready for routine use?**”

70–75 min — Priorities & close

- **Top-3 fixes** for demonstrators (dot-vote if time).
- Thank you; explain next steps (short write-up for confirmation).

5) Question bank (use as probes)

- **Uptake:** “Which artefact (summary line, chart, export) would be most useful? To whom?”
- **Coverage:** “Name the minority view at risk of being missed here.”
- **Moderation:** “What **policy clause** is decisive for this item?”
- **Counterfactual:** “If we swap [group] with [group], should anything change?”
- **Privacy/Security:** “What governance note would you attach to an export?” “Any scenario where a bad actor could game the system?”

6) Live capture template (copy/paste)

- **Scenario:** task/team
- **Uptake example #1–#3:** *feature* → *step helped* → *how used*
- **Summary (AIA):** utility (1–5); coverage: missing view (Y/N + which); example sentence to add
- **Moderation (AIA):** per item — staff decision vs system; fairness notes; **counterfactual: same/different**
- **Privacy/Security (AIA):** mask this → ... ; risks → ... ; safeguards → ...
- **Top-3 fixes:** item • owner • target date

7) After the session (same day)

- Produce a **2-page memo** with: participants/roles; scenario; ≥ 3 uptake examples; AIA findings (summary coverage, moderation fairness, privacy/security); Top-3 fixes with owners/dates.
- Copy AIA entries into the **AIA Sheet** for the site (v1).
- Send memo to demonstrators for factual check; incorporate minor edits.

8) Accessibility & inclusion tips (both FGs)

- Offer live-caption toggle and a dial-in alternative.
- Avoid rapid screen scrolling; describe what's on screen.
- Read poll options aloud; allow chat answers.
- Pause after each question (5–7 seconds) for slower processing.
- If distressing content appears, offer to skip and move on.

9) Time-boxed versions (if you only have 45 min)

- **Participants FG (45')**: Warm-up (3) → Ease (10) → Summaries (12) → Moderation (10) → Accessibility (5) → Top-3 (5).
- **Demonstrators FG (45')**: Intros+Scenario (5) → Summaries (12) → Moderation incl. 1 counterfactual (15) → Privacy/Sec (8) → Top-3 (5).

Annex 2.2: Focus Group Report Templates

Participants (End-Users)

Site: [Braşov / Martin] **Date:** [DD Mon YYYY] **Time:** [HH:MM–HH:MM]

Moderator: [Name] **Note-taker:** [Name] **Language:** [RO / SK / EN]

Session mode: [Online / In-person] **Platform:** <https://ithaca.simavi.ro/>

1) Purpose

Briefly restate aims (usability, accessibility, perceived fairness/usefulness of summaries & content rules) and that the session prioritises fixes.

2) Participants

- **N:** [6–8] **Codes:** P1...PN
- **Notable accessibility setups (if any):** [screen reader / keyboard-only / zoom / contrast / other]
- **Consent:** obtained from all participants.

3) Agenda (actual)

- 0–5 Warm-up
- 5–20 Ease & friction (core journeys)
- 20–35 AI summaries (usefulness & coverage)
- 35–45 Content rules & reporting
- 45–55 Accessibility

- 55–70 Top-3 fixes & wrap-up

4) Methods & Materials

- Demo thread (Pilot-Eval) shown? [Yes/No]
- Borderline items shown? [Yes/No] (count: [])
- Any polls used? [Yes/No] (briefly note)

5) Findings

5.1 Ease & Friction (core journeys)

Journey/Step	What happened (problem or success)	Evidence (URL/screenshot/quote)
[]	[]	[]
[]	[]	[]

5.2 AI Summaries — Usefulness & Coverage

- Usefulness (1–5 poll mean): []
- Missing/under-represented viewpoints (esp. minority view):
 - []
- Suggested sentence(s) to add:
 - “[quote]”

5.3 Content Rules & Reporting

Item (ID/link)	Group decision (Report/Ignore)	Rationale (short quote)
[]	[]	“[]”
[]	[]	“[]”

Perceived consistency of rules across different speech types: [summary]

5.4 Accessibility

Barrier type (contrast/labels/focus/keyboard/reader/motion/other)	Location (page/element)	Proposed fix
[]	[]	[]
[]	[]	[]

6) Top-3 Fixes (voted)

- Σχήμα 2. — owner: [] target: []
- Σχήμα 3. — owner: [] target: []
- Σχήμα 4. — owner: [] target: []

7) Notable Quotes

- “[quote]” — P#
- “[quote]” — P#

8) Action Log

Issue Owner Due Status (Open / In progress / Done)

[] [] [] []

9) Data & Ethics

Anonymised notes; no sensitive data captured. Storage: [location]. Access: study team only.

Appendices

- A. Materials shown (links or screenshots)
- B. Poll results
- C. Full notes (if attached)

Demonstrators (Municipal Staff)

Site: [Braşov / Martin] **Date:** [DD Mon YYYY] **Time:** [HH:MM–HH:MM]
Roles present: [Policy / Comms / Moderation / IT-Admin / Other] **N:** [4–6]
Moderator: [Name] **Scribe:** [Name] **Language:** [RO / SK / EN]
Session mode: [Online / In-person] **Platform:** <https://ithaca.simavi.ro/>

1) Purpose

Capture **organisational uptake** (where/how ITHACA supports real tasks) and collect **AIA** evidence: summary coverage (incl. minority view), moderation fairness/consistency (incl. counterfactuals), and privacy/security expectations; agree Top-3 fixes.

2) Scenario

Chosen workflow scenario (agenda-setting / briefing / comms / moderation):

- **Description:** [one paragraph]
- **Teams involved:** [] **Outputs needed:** []

3) Agenda (actual)

- 0–5 Intros & scenario pick
- 5–25 Summary in workflow (AIA: coverage)
- 25–45 Moderation consistency (AIA: fairness & stability)
- 45–60 Privacy & security (AIA)
- 60–70 Organisational uptake
- 70–75 Top-3 fixes & close

4) Methods & Materials

- Long thread (Pilot-Eval) used? [Yes/No]
- Borderline items shown? [Yes/No] (count: [])

- Counterfactual pairs (identity token-swap) shown? [Yes/No] (count: [])

5) Findings

5.1 Summary in Workflow (AIA)

- **Utility (1–5 poll mean):** []
- **Coverage:** was a minority/less popular view present? [Yes/No]
- **If missing, which view?** [text]
- **Suggested sentence/edit to make it “ready now”:**
 - “[proposed line]”
- **Planned use in workflow:** [briefing line / agenda note / public message / dashboard]

5.2 Moderation Consistency (AIA)

Item (ID/link)	Staff decision (Keep/Remove/Unsure)	System decision (if shown)	Fairness notes
[]	[]	[]	[]
[]	[]	[]	[]

19. **Counterfactual pairs:** treated the same? [Yes/No] — Notes: []

20. **Key policy clauses referenced:** []

5.3 Privacy & Security (AIA)

Data to mask before sharing (names/IDs/quotes/other): []

Risks/abuse scenarios raised (spam, prompt/format injection, misuse): []

Requested safeguards (audit trail, export watermark, rate limits, moderation-queue rules): []

5.4 Organisational Uptake

- **Where it helps next month (tick):** Agenda / Comms / Moderation / Leadership briefing / Public-facing
- **Readiness for routine use (1–5):** []
- **One change that makes it “ready now”:** []

6) Top-3 Fixes (agreed)

- 9. — *owner:* [] *target:* []
- 10. — *owner:* [] *target:* []
- 11. — *owner:* [] *target:* []

7) Uptake Examples (≥3)

- 3) **Feature** → **Step helped** → **How used** (e.g., “AI summary → briefing → 2-line brief for cabinet”)
- 4) []

5) []

6) []

8) Notable Quotes

- “[quote]” — Role
- “[quote]” — Role

9) Action Log

Action Owner Due Status (Open / In progress / Done)

[] [] [] []

10) Data & Ethics

Anonymised notes; no sensitive data captured. Storage: [location]. Access: study team only.

Appendices

- A. Materials shown (links or screenshots)
- B. Poll results
- C. Full notes (if attached)

Annex 3: Field Protocols

Citizens

A: Content snapshots**0) Snapshot**

- **Goal:** Each participant completes **≥6 logged-in visits in 2 weeks** (≈10–15' each), across **Guided** and **Free** modes, completing **A1 once, A2 every visit, A3 if applicable, A4 at the end.**
- **Where:** <https://ithaca.simavi.ro/>
- **What counts as a “visit”:** logged-in + **≥1 action** (read summary / react / post / report) or active for **≥5 minutes.**

1) Study Pack (prepare once before recruitment)

Checklist to assemble:

Platform URL: <https://ithaca.simavi.ro/>

Survey links (short URLs):

- A1 Baseline • A2 Post-Visit • A3 Quick Accessibility • A4 End-of-Study

Roster Sheet (spreadsheet): columns = PID, Name (internal only), Email/Phone, City (Braşov/Martin), Mode Order (first 3 = Guided/Free; next 3 = switch), Sessions done (0–6+), A1/A2/A3/A4 status, Notes.

Message templates: recruitment SMS/Email, consent text, reminders (Day 3/7/10).

Accessibility note: quick instructions for larger text, high contrast, keyboard-only, screen reader.

Support contacts: researcher email/phone; local tech contact; password reset link.

Data folder: [Add links]

2) Recruit & Invite (copy-paste messaging)

Message (45–60 words):

Hi [Name]! We're testing the city's ITHACA platform. Over **2 weeks**, please log in **6 times** for **10–15 min** at <https://ithaca.simavi.ro/> and do a **1-min survey** after each visit. Voluntary; anonymous in reports. Want quick start + your survey links?

Email:

Subject: ITHACA — 6 short visits from home (2 weeks)

Hello [Name],

Thanks for helping us test ITHACA. Over **two weeks**, please log in to <https://ithaca.simavi.ro/> at least **six times** (≈10–15 min each). After every visit you'll complete a **1-minute survey**. It's voluntary; responses are anonymous in reports.

Your first 3 visits: **[Guided/Free]** → next 3 visits: **[Free/Guided]**.

Please always **log in with your credentials** so your visits count.

If you need larger text, high contrast, keyboard-only or screen reader, we support that—just tell us.

Best,

[Researcher • Contact]

3) Consent & Baseline (first login)

Consent script (read verbatim):

“Participation is voluntary; you can stop anytime. We'll collect short surveys; your answers are reported in aggregate with no names. Please avoid sharing personal or sensitive details in public comments. OK to proceed?”

Information sheet and consent form to be sent/ provided and signed prior participation.

Action: Send **A1 Baseline**. Confirm submission in the Roster (A1=Done).

4) Mode Plan & Definition of a Visit

- Assign **Mode Order** per PID:
 - **Order A:** Visits 1–3 = **Guided**, 4–6 = **Free**
 - **Order B:** Visits 1–3 = **Free**, 4–6 = **Guided**
- Tell participant:

“A visit counts if you are logged in and do at least one action or stay active for 5+ minutes. Aim for 10–15 minutes each visit.”

5) Guided Missions (send before each Guided visit)

Mission J1 — Join & contribute (5–7')

1. Sign in: <https://ithaca.simavi.ro/>
2. Open **[Topic link: content related to pilot and in native language]**
3. Post **one short comment** (1–2 lines)

4. Add **one reaction** to any post

Mission J2 — Read a summary (4–5')

1. Open [**Long thread link: content related to pilot and in native language**]
2. Click to view the **AI summary**
3. Skim **3–5** original posts under it
4. You will note if it helped / missed anything in the survey

Mission J3 — Use the Toxicity tool (3–4')

1. Check with three terms [Vangelis will provide the terms in English, Romanian and Slovak] (3 items)
2. For each item check the outcome
3. If the system shows an outcome, note if it matched your expectation

Accessibility nudge to include when relevant, similar to the following:

“Try **larger text** (Ctrl/Cmd + +), **high contrast**, **keyboard-only** (Tab/Shift+Tab), or a **screen reader**. If anything doesn’t work, mention it in the survey.”

6) Free Visits (send this one-liner)

“Explore topics you care about: read, react, or post. If you see something off, try **Report**. Finish with the **1-minute survey**.”

7) End-of-Visit Routine (every visit)

1. Participant finishes actions.
2. **Immediately send/open A2 Post-Visit** (1–2’).
3. In Roster: increment **Sessions done**; mark A2=Done (Visit #).
4. **If participant uses assistive tools** (from A1): send **A3 Quick Accessibility** once during the **first three visits** (and again anytime they report a blocker OR each time). Mark A3 date in Roster.

8) End-of-Study Routine (after ≥6 visits or end of week 2)

1. Send **A4 End-of-Study** (3–4’).
2. In Roster: mark A4=Done.
3. (Optional) Invite to **online participant focus group** (60–75’), send slot options.

9) Reminder Schedule (use local time)

- **Day 3:**

“Hi [Name]! Quick nudge—could you do **one 10–15-minute visit** today and the 1-minute check-in at the end?”

- **Day 7 (switch modes):**

“Halfway! For your next visits, switch to [**Guided/Free**]. Same 10 minutes + mini-survey.”

- **Day 10:**

“Almost there! If you can, please complete your remaining visits this week.”

10) Troubleshooting (talk-tracks to read)

- **Can't sign in:**

"Use 'Forgot password'; try Chrome/Edge. If it still fails, screenshot the exact message and send it to me."

- **Summary doesn't show:**

"Refresh; try another long thread; if still missing, continue your visit and tell us in the mini-survey."

- **Accessibility barrier:**

"Tell me the page and step (e.g., 'Post button not labelled'). You can skip it—complete A3 so we can fix it."

11) Quality & Close-out Checks (researcher actions)

- **Minimum complete case:** A1 + ≥ 6 visits with A2 each + (A3 if applicable) + A4.
- **If someone has only 4–5 visits by Day 12:** send a friendly "two more to go" message.
- **Data hygiene:** the roster total of A2 entries must equal the **number of visits** logged; spot-check 10% for consistency (session time vs A2 timestamp).
- **File any accessibility blockers** into /issues/ with page, element, and the participant's short description (no PII).

APPENDIX — LINK & FILE PLACEHOLDERS

A. Participants (paste into your docs):

- A1 Baseline: [URL]
- A2 Post-Visit: [URL]
- A3 Quick Accessibility: [URL]
- A4 End-of-Study: [URL]

B. Site-specific content:

- **Braşov:** [Topic], [Long thread 1], [Long thread 2], [Tricky items], [Counterfactual pairs]
- **Martin:** [Topic], [Long thread 1], [Long thread 2], [Tricky items], [Counterfactual pairs]

C. C) Participants (End-Users) — Suggestions for Tasks:

Overall rule: complete ≥ 6 visits in 2 weeks, 10–15 minutes each, logged in at <https://ithaca.simavi.ro/>. After **each** visit, complete **A2 Post-Visit** (1–2'). Do **A1** once (first visit) and **A4** once (last visit). If you use assistive tech, do **A3** once early (and again if you hit a barrier).

Visit 0 (first login, 5')

1. Go to <https://ithaca.simavi.ro/> → **Log in** (reset password if needed).
2. Open the **A1 Baseline** mini-survey and submit.
Success criteria: able to log in; A1 submitted.
Record: A1 status.

Visit 1 — Mission J1: Join & Contribute (10–15')

1. Open [Topic list link] → choose [Topic: e.g., Mobility].

2. Open [**Thread: e.g., Bike lanes in District 2**].
3. **Post 1 short comment** (1–2 lines).
4. **Add 1 reaction** to any post.
5. If you see something off, try **Report** (optional).
6. Finish by completing **A2 Post-Visit**.
Success criteria: 1 comment + 1 reaction posted while logged in; A2 submitted.
Record: A2 (ease, satisfaction, blocker, return intent, accessibility fit).

Visit 2 — Mission J2: Read a Long Thread with Summary (10–15')

1. Open [**Long thread link #1**].
2. Click to open the **AI Summary** panel; **read it fully**.
3. **Skim 3–5 original posts** below the summary.
4. If you disagree or see a missing view, optionally **write 1 comment**.
5. Finish by completing **A2 Post-Visit**.
Success criteria: AI summary opened and read; A2 submitted.
Record: A2; (if applicable) note “summary used” in comments.

Visit 3 — Mission J3: Check 3 Tricky Posts (8–12')

1. Open [**Tricky items list link**].
2. For **each** of 3 items: decide **Report** or **Ignore** (your judgment).
3. If the system shows an outcome, note mentally whether it **matched** your expectation.
4. Finish by completing **A2 Post-Visit**.
Success criteria: 3 items reviewed with an action; A2 submitted.
Record: A2 (note slowdown point if any).

If you use assistive tools: after any one of visits 1–3, also fill **A3 Quick Accessibility** (2').
Success criteria: A3 submitted once (and again if you later hit a barrier).

Visit 4 — Free Explore (10–15')

1. Pick **any topic you care about** from [**Topic list link**].
2. Do **at least two** of: read a summary / react to a post / write a short comment / report something off.
3. Finish by completing **A2 Post-Visit**.
Success criteria: ≥2 actions completed; A2 submitted.
Record: A2.

Visit 5 — Compare Two Threads (10–15')

1. Open [**Long thread link #2**] (different topic).
2. Read the **AI Summary** → skim **3–5** posts.
3. If time: open [**Shorter thread link**] and **react** or **comment** once.
4. Finish with **A2 Post-Visit**.
Success criteria: summary read on #2 + one action in #2 or the short thread; A2

submitted.

Record: A2.

Visit 6 — Your Choice + Wrap (10–15' + 3–4')

1. Return to **any** thread you engaged with or open a new one from **[Topic list link]**.
2. Do **two** actions (e.g., react + comment, or report + comment).
3. Complete **A2 Post-Visit**.
4. Complete **A4 End-of-Study** (final survey).

Success criteria: ≥2 actions; A2 + A4 submitted.

Record: A2, A4.

Accessibility add-ons (use whenever relevant)

- **Screen reader / keyboard only:** try navigating the thread list, opening a thread, and posting a short comment **without a mouse**. If blocked, describe briefly in **A3**.
- **High contrast / large text:** enable your browser's zoom (Ctrl/Cmd + +) or **[High-contrast toggle location if available]** and attempt steps above. Note any issues in **A2** or **A3**.

Demonstrators/ Moderators

0) Snapshot

- **Goal:** Run **two remote sessions (or per person) per site** (45–60', 3–6 or 1-on-1 staff: policy/comms/moderation/IT) to:
 - (a) collect **organisational uptake** examples;
 - (b) gather **AIA evidence** (summary coverage incl. minority view; moderation fairness/consistency incl. counterfactuals; privacy/security expectations);
 - (c) agree **Top-3 fixes**.

1) Study Pack (prepare 1–2 days prior)

Attendees confirmed (roles listed; login access checked).

Materials (tabs open):

- **Long thread 1** (Pilot-Eval) with diverse viewpoints
- **Borderline items** (3–6 neutralised examples)

DEM survey links: DEM-1 Baseline; DEM-2 Summary; DEM-3 Moderation (repeat per item); DEM-4 Privacy/Security; DEM-5 Uptake; DEM-6 Wrap.

Note template with sections: Scenario • Uptake examples (≥3) • Summary coverage • Moderation fairness • Privacy/Security • Top-3 fixes.

Video platform ready (screen-share on).

2) Run-of-Show (minute-by-minute, read-off script)

0–3' Welcome & consent

"Thanks for joining. You will **show how you would use ITHACA** in your work. We will look at **one real task**, a **summary** of a long thread, a few **borderline posts**, and **privacy/security**. We take anonymous notes; you can skip anything. OK to proceed?"

→ Send **DEM-1 Baseline** (role, focus, prior AI use) and wait 60–90".

3–8' Choose a real scenario

“Name one task where public input matters: **agenda item, briefing, communications, or moderating a thread**. We will use that.”

Researcher writes the chosen **Scenario** in notes.

8–25' Summary in workflow — AIA: coverage

- Facilitator screen-shares **Long thread + AI summary**.
- Prompts (ask one by one):
 - **Utility:** “For your scenario, does this summary save time capturing key points?” (*show quick 1–5 poll or thumbs up/down*)
 - **Coverage:** “Is any important—especially **minority**—view **missing**? Which?” (*ask for one sentence they expect*)
 - **Adoption:** “Where would you paste this **right now** (briefing line, agenda note, public message)? Any small edit needed?”
- → Send **DEM-2 Summary check**; wait ~2’.
- Researcher captures: utility score, missing view (Y/N + which), suggested sentence/edit, planned use.

25–45' Moderation consistency — AIA: fairness & stability

- For **each** borderline item (3–6 total):
 - Show item (neutralised example).
 - Ask:

“Under your policy, **Keep/Remove/Unsure**—why?”

“Does the system’s decision match what you expect?”

- → Send **DEM-3 block** for the item; wait ~45”.

45–55' Privacy & security — AIA

21. Prompt:

“Before routine use, what would you **mask** in an export (names, IDs, quotes)?”

“Any **risks** that would stop use (e.g., content that could trick the system)? What **safeguard** would reassure you (audit trail, export watermark, rate limits, moderation queue rules)?”

22. → Send **DEM-4**; wait ~60”.

23. Scribe notes: masking needs; risks; safeguards requested.

55–60' Organisational uptake & wrap

Send **DEM-5** (where it helps; readiness; one change for “ready now”) and **DEM-6** (Top-3 fixes).

Facilitate quick **Top-3** selection (raise hands or chat vote).

Close:

“We’ll send a 2-page note with **uptake examples, AIA observations, and Top-3 fixes** for your confirmation. Thank you!”

3) Live Capture — Exactly what to write (scribe)

Fill these headings during the session:

- **Scenario:** [task/team; when used; output needed]
- **Uptake examples (≥3):** Feature → Step helped → How used next month
 - o e.g., “AI summary → briefing → 2-line item for cabinet”
- **Summary (AIA):** utility [1–5]; missing view **Y/N** (which); **sentence/edit** to add; intended **destination** (briefing/agenda/public)
- **Moderation (AIA):** per item — ID • staff decision • system decision (if shown) • fairness note;
- **Privacy/Security (AIA):** mask → risks → safeguards → ...
- **Top-3 fixes:** item • owner • target date

4) After the Session (same day)

12. Produce a **2-page memo** from the notes:

- 12.1. Participants/roles; Scenario; **≥3 uptake examples**; AIA (summary coverage; moderation fairness incl. counterfactuals; privacy/security); **Top-3 fixes** with owner/target.

13. Copy AIA items into your **AIA sheet** (if used centrally).

14. Email the memo to demonstrators for factual confirmation (track-changes OK).

5) Time-boxed variant (45' total)

- 7) **0–3'** Consent & DEM-1
- 8) **3–15'** Summary (DEM-2) — limit to one long thread
- 9) **15–33'** Moderation (DEM-3) — 3 items + 1 pair
- 10) **33–40'** Privacy/Security (DEM-4)
- 11) **40–45'** Uptake (DEM-5) + Top-3 (DEM-6)

6) Troubleshooting

- **No logins available:** facilitator screen-shares and drives; still run DEM forms.
- **Lack of consensus on moderation:** record all rationales; note the policy clause each person cites.
- **Time overrun:** skip extra items; keep **one** pair for counterfactual fairness.

APPENDIX — LINK & FILE PLACEHOLDERS

- **Demonstrators (paste into your invites):**
- DEM-1: [URL] • DEM-2: [URL] • DEM-3: [URL] • DEM-4: [URL] • DEM-5: [URL] • DEM-6: [URL]
- **Site-specific content:**
- **Braşov:** [Topic], [Long thread 1], [Long thread 2], [Tricky items], [Counterfactual pairs]
- **Martin:** [Topic], [Long thread 1], [Long thread 2], [Tricky items], [Counterfactual pairs]
- **Demonstrators/ Moderators (Municipal Staff) — Suggested Tasks (Single 45–60' Session)**

Setup: Everyone logged in or the facilitator screenshares. Have these ready:

- **[Long thread link]** with diverse viewpoints (Pilot-Eval)
- **[Borderline items set]** (3–6 neutralised examples)
- **[Counterfactual pairs]** (2–3 pairs where only a group word differs)
- Links to **DEM-1...DEM-6** mini-surveys.

Segment 1 — Opening & Scenario (0–8')

- **Consent script** (researcher reads).
- Send **DEM-1 Baseline** (1–2').
- **Pick one real scenario** (agenda setting / briefing / communications / moderation).
Success criteria: DEM-1 submitted; scenario stated.
Record: Scenario text in notes.

Segment 2 — Summary in Workflow (AIA: Coverage) (8–25')

- Open **[Long thread link]** → show **AI Summary**.
- **Prompt 1 (Utility):** “For your scenario, does this save time capturing key points?” (*thumbs or 1–5 poll*)
- **Prompt 2 (Coverage/minority views):** “Is any important—especially **minority**—view **missing**? Which?”
- **Prompt 3 (Adoption):** “Where would you paste this **right now** (briefing line, agenda note, public message)? Any **small edit** needed?”
- Send **DEM-2 Summary Check** (2–3').
Success criteria: DEM-2 submitted; at least one **missing view** call-out or confirmation; one **adoption destination** named.
Record: Utility rating, missing view (Y/N + which), **one sentence** they'd add, where they would paste it.

Segment 3 — Moderation Consistency (AIA: Fairness & Stability) (25–45')

For each **[Borderline item]** (3–6 total):

- Show item (context + text).
- Ask **two decisions**:

1. “Under your policy: **Keep / Remove / Unsure** — why?”
 2. “Does the **system’s** decision match what you expect?”
- Send **DEM-3 Moderation block** for the item (repeat per item).

Success criteria: DEM-3 submitted for each item; at least one **policy-based rationale** captured; at least one **counterfactual judgement** (same/different) recorded.

Record: Per item — staff decision, system decision (if shown), rationale, fairness notes.

Segment 4 — Privacy & Security (AIA) (45–55')

- Ask: “Before routine use, what would you **mask** in an export (names, IDs, quotes, other)?”
- Ask: “Any **risks** that would stop use (e.g., tricking the system, spam)? What **safeguard** would reassure you (audit trail, watermark, rate limits, moderation queue rules)?”
- Send **DEM-4 Privacy & Security** (1').
Success criteria: DEM-4 submitted; at least one **masking rule** and one **safeguard** captured.
Record: Masking list; risk list; safeguards requested.

Segment 5 — Uptake & Wrap (55–60')

- Send **DEM-5 Uptake** (where it helps; readiness 1–5; one change for “ready now”).
- Send **DEM-6 Wrap** (Top-3 fixes).
- **Agree Top-3 fixes** (voice or quick poll) and assign **owner / target date** (draft).
Success criteria: DEM-5 + DEM-6 submitted; **Top-3** recorded with an owner/target.
Record: 3+ **uptake examples** (Feature → Step helped → How used next month); Top-3 fixes table.

Optional follow-ups (post-session)

- **Uptake example exports:** Ask which export they’d need (e.g., “Top-5 points” snippet for agendas).
- **Policy mapping:** If disagreements in moderation arose, list the **policy clause** each person cited.

One-page capture grid (researcher uses during session)

Scenario: [task/team/output]

Uptake examples (≥3):

- AI summary → [step] → [how used next month]
- Argument view (if present) → [step] → [use]
- Moderation queue → [step] → [use]

Summary (AIA): Utility [1–5]; Missing view Y/N (which); “Add this sentence: []”; Destination [briefing/agenda/public].

Moderation (AIA):

- Item 1: Staff [Keep/Remove/Unsure] vs System [Allow/Remove/Escalate]; Why: “[...]”; Fairness note: [...]
- Item 2: ...

- Counterfactual Pair A/B: **Same/Different**; Notes: [...]

Privacy/Security (AIA): Mask [names/IDs/quotes/...]; Risks [...]; Safeguards [audit trail / watermark / rate limits / queue rules / ...].

Top-3 fixes:

- — Owner [] — Target []
- — Owner [] — Target []
- — Owner [] — Target []

Annex 4: Performance Testing Protocol

Audience: SIMAVI / KT dev-ops (CERTH coordination)

Scope: Load, spike, stress, (optional) soak, and failure-simulation tests on production-like envs, with **Algorithmic Impact Assessment (AIA)** cross-checks to ensure fairness/consistency isn't degraded under load.

0. Objectives

- Verify priority user journeys meet **latency & error** targets during pilot windows.
- Validate **graceful degradation and recovery**.
- Confirm AI components (Summaries, Moderation) remain **consistent, fair, and usable** under load/failure (AIA cross-check).

1. Test Gates & Schedule

- ❑ **Gate A (Pre-pilot):** baseline load + brief spike; record SLOs; run AIA cross-check (baseline).
- ❑ **Gate B (Mid-pilot regression):** load + spike + failure simulation; confirm no regressions; re-run AIA cross-check.
- ❑ **Gate C (End):** verification load test after fixes; final AIA cross-check.

2. System Under Test (SUT)

- ❑ **Environment:** [Staging / Prod-like / Agreed prod window]
- ❑ **Version/Tag:** [e.g., vX.Y.Z] • **Feature flags:** [on/off list]
- ❑ **Services:** Web app, API, workers, DB, cache, queue, **AI inference endpoints** (summary, moderation).

3. Service Levels & Targets (SLOs)

p95 latency targets (per journey/endpoint):

- Login: **≤ 1500 ms**
- Topic list / Thread view: **≤ 2000 ms**

- Post comment / React: ≤ 2000 ms
- View AI summary (serve or generate+serve): ≤ 2500 ms
- Moderation decision (API): ≤ 500 ms
Availability (pilot windows): $\geq 99.5\%$ • Error rate: $\leq 1\%$ total; 0 criticals.

4. Load Models & Concurrency

24. Estimate active users: $U = \text{participants per site} \times \text{participation rate} \times \text{overlap factor}$.
25. Concurrency target (VUs) = $\max(U, 2 \times U)$ for safety margin.
26. Example bands: **Small 20 VU • Medium 50 VU • Large 100 VU** (pick per site/window).

5. Test Types

Load: ramp-up 5' → hold 15–30' at target VUs.

Spike: jump to 1.5–2× target for 3–5'.

Stress: step beyond capacity to find breakpoints.

Soak (optional): 60–120' steady load (leaks/backlog).

Failure simulation: brief node/network cut; observe **recovery time** and **graceful errors**.

6. Scripts (Journeys)

- **J1:** Login → topic list → open thread → read → react.
- **J2:** Login → open long thread → **request/view AI summary**.
- **J3:** Login → post comment → **report borderline item** → moderation decision.
- **J4:** Static assets fetch (cache efficiency).
Add think-time 1–3 s; randomise threads/posts.

7. Instrumentation & Capture

15. p50/p95/p99 **latency, throughput, error rate** per endpoint.
16. Infra: **CPU/Mem**, DB/queue utilisation, **worker backlog, cache hit rate**.
17. Logs: request IDs, error traces.
18. Take **snapshots** at start / peak / end.

8. AIA Cross-Check (each Gate)

8.1 Moderation consistency under load

- 12) Use a **fixed** labelled set of **40 borderline items** (same text each run).
- 13) Score at **idle** (baseline) and at **peak load**; record decisions/scores.
- 14) Compute **flip-rate** vs baseline; **threshold ≤ 5 percentage points (pp)**.
- 15) Note any differences that align with **identity terms/dialect** (fairness slice).

8.2 Summary coverage under load

- Use **2 fixed long threads** with pre-extracted **key viewpoints** (incl. minority view).
- Generate/serve summaries at **idle** and at **peak**; capture text.
- Check presence of minority-view sentence; flag if missing **under load**.

8.3 Latency budgets & fair fallbacks

- Confirm **Moderation p95 ≤ 500 ms**; **Summary p95 ≤ 2500 ms**.
- If fallbacks used (e.g., cached/stale summary), verify they **do not degrade fairness** (e.g., keep minority view).

9. Pass/Fail Criteria

- All SLOs met; **0 critical** errors; recovery **< 5 min**; graceful error surfacing.
- **AIA cross-check** passes: moderation **flip-rate ≤ 5 pp**; **no coverage regression**; **no identity-slice regressions**.

10. Execution Steps (per Gate)

- Freeze version/config; note flags.
- Warm-up 5'.
- Run: **Load** → **Spike** → **(Soak optional)** → **Failure sim**.
- Run **AIA cross-checks** (8.1–8.3).
- Export raw results/logs; snapshot dashboards.
- Fill **Performance Test Report + AIA–Perf Cross-Check Sheet**.
- File defects; assign owners; schedule re-test if needed.

11. Safety & Ethics

- No personal data in test payloads.
- Do not impact live participant windows without agreement.

12. Artefacts to Produce

8. Gate A/B/C **Performance Test Report**
9. **AIA–Performance Cross-Check Sheet** (CSV/XLSX or doc)
10. Test scripts/configs, monitoring snapshots, incident log

2) Performance Test Report Template (per Gate A/B/C)

Environment: [Staging / Prod-like / Prod] **Version/Tag:** [] **Date:** []
Window: [Start–End, UTC/EET] **Team:** [Names]

1) Scope & Objectives

Brief summary of journeys tested, SLOs, and the gate (A/B/C).

2) Test Plan Summary

- **Load model:** [Small / Medium / Large] ([] VU)
- **Types run:** Load [], Spike [], Stress [], Soak [], Failure sim []
- **Scripts:** J1 / J2 / J3 / J4 (versions [])

3) Results — SLOs

3.1 Latency (ms) & Error Rate

Endpoint / Journey	p50	p95	p99	Error %
Login	[]	[]	[]	[]
Topic list / Thread view	[]	[]	[]	[]
Post / React	[]	[]	[]	[]
AI Summary (serve/gen)	[]	[]	[]	[]
Moderation decision	[]	[]	[]	[]

3.2 Throughput & Capacity

- **Peak RPS:** []
- **Sustained RPS (hold):** []
- **Max stable concurrency before degradation:** []

3.3 Availability & Recovery

- **Availability in window:** []
- **Failure sim:** [what was cut] • **Recovery time:** []
- **Error surfacing:** [graceful / raw stack / blank]

4) System Metrics (peak snapshot)

CPU []% • Mem []% • DB CPU []% • Cache hit []% • Queue backlog max []

5) AIA Cross-Check Results

5.1 Moderation consistency

- **Items checked:** 40 (fixed set)
- **Flip-rate vs baseline:** [] pp (threshold ≤ 5 pp)
- **Identity-slice deltas (Δ FPR/ Δ FNR):** [summary or N/A]

5.2 Summary coverage

- **Threads checked:** 2 (fixed)
- **Minority-view sentence present (Idle \rightarrow Load):** [Y/N] \rightarrow [Y/N]

- **Notes on any differences:** []

5.3 Latency budgets & fallbacks

- **Moderation p95 ≤ 500 ms:** [Pass/Fail] (value: [])
- **Summary p95 ≤ 2500 ms:** [Pass/Fail] (value: [])
- **Fallback used:** [Yes/No] • **Fairness impact:** [None / Risk — explain]

6) Incidents & Defects

ID	Sev	Area	Summary	Owner	Status
[]	[P1/P2]	[API/UI/DB/AI]	[]	[]	[Open/In-prog/Done]

7) Conclusions

Gate [A/B/C] **Pass/Fail**; bottlenecks; immediate mitigations; **retest needed** [Y/N].

8) Attachments

- Test scripts/configs
- Monitoring snapshots
- Raw CSV exports
- **AIA–Perf Cross-Check Sheet**

3) AIA–Performance Cross-Check Sheet (per Gate)

Gate: [A/B/C] Version/Tag: [] Date: []

Section 1 — Moderation Consistency Under Load

- **Fixed borderline set used:** [Yes/No], **ID:** [name/version]
- **Items (IDs):** [list or link]
- **Baseline decisions:** [attach/link]
- **Load decisions:** [attach/link]
- **Flip-rate vs baseline:** [] pp (threshold ≤ 5 pp)
- **Notes (identity-term/dialect differences):** []

Section 2 — Summary Coverage Under Load

- **Threads:** [IDs/links]
- **Idle summary captured:** [Y/N] (path)
- **Load summary captured:** [Y/N] (path)
- **Minority-view sentence present:** Idle [Y/N] → Load [Y/N]
- **Differences / risks:** []

Section 3 — Latency Budgets & Fallbacks

- **Moderation p95 ≤ 500 ms:** [Pass/Fail] (value: [])
- **Summary p95 ≤ 2500 ms:** [Pass/Fail] (value: [])
- **Fallback used (cache/placeholder):** [Yes/No] → **Fairness impact?** [None / Describe]

Section 4 — Sign-off

- **Prepared by:** [] **Reviewed by:** [] **Date:** []
- **Follow-up actions:** []

Annex 5: Logged Data Request Brief

Template A — Data Request Brief (send to devs/admins)

Purpose: Provide Phase-2 evaluation analytics for Braşov & Martin, in a privacy-respecting, pseudonymous form.

Time window: [start date] → [end date]

Sites: Braşov, Martin

Granularity needed:

- **Per session (row = one user session)**
- **Per day per site**
- **Per thread per day**
- **Feature usage snapshots** (summaries, moderation, accessibility)
- **(Optional) Pseudonymous roster** for cohort linking

Privacy: No PII (no names, emails, IPs). Use a stable, salted pseudonymous user ID only available to the study team.

Delivery format: CSV or XLSX, one file per template below.

Template B — Export “Session Summary” (row = one user session)

We need (descriptive)	What it means	Example value
Site (city)	Which pilot site the session belongs to	Braşov / Martin
Pseudonymous user ID	Stable, salted ID for user	u_9b3f...
Session identifier	Unique session code	s_2f1a...
Session start time (ISO)	When the user began acting	2025-09-12T17:03:11Z
Session end time (ISO)	When the user stopped (or auto-ended)	2025-09-12T17:14:59Z
Session duration (sec)	End – Start, excluding long idle	708
Device type (broad)	Phone / Desktop / Tablet	Desktop

We need (descriptive)	What it means	Example value
Study mode tag	Guided / Free (if known)	Guided
Threads viewed (count)	Count of distinct threads opened	4
Summaries viewed (count)	Count of times a summary panel was opened	2
Posts created	New posts (or topics) authored	0
Comments created	Replies authored	1
Reactions made	Likes/emoji reactions	3
Reports submitted	Times “Report” was used	1
Accessibility used (Y/N)	Any accessibility feature used this session	Y

Template C — Export “Daily Site Summary” (row = site × day)

We need	What it means	Example
Date (YYYY-MM-DD)	Calendar day	2025-09-12
Site (city)	Braşov / Martin	Braşov
Active users (count)	Users with ≥1 action or ≥5m active	42
Sessions (count)	Total sessions that day	97
Threads viewed (count)	Total thread views	311
Summaries viewed (count)	Total summary opens	151
Posts made	New posts/topics	7
Comments made	Replies	89
Reactions made	Likes/emoji	265
Reports made	Report actions	18
Avg session duration (sec)	Mean over sessions	690

Template D — Export “Thread/Topic Daily Summary” (row = thread × day)

We need	What it means	Example
Date (YYYY-MM-DD)	Calendar day	2025-09-12
Site	Braşov / Martin	Martin
Topic identifier	Human or internal topic ref	transport_brasov
Thread identifier	Internal thread ref/URL slug	thr_884c...
Unique participants	Distinct users who acted/viewed	23
Comments total	Comments posted that day	37
Reactions total	Reactions that day	94

We need	What it means	Example
Reports submitted	Reports on this thread that day	4
Summary views	Summary panels opened	28
Marked as AIA set? (Y/N)	Thread is in curated AIA list	Y

Template E — Export “AI Summary Usage” (row = one summary view)

We need	What it means	Example
Site	Braşov / Martin	Braşov
Pseudonymous user ID	Stable, salted	u_9b3f...
Session identifier	Link to Session Summary	s_2f1a...
Thread identifier	Where the summary was read	thr_884c...
Summary identifier (if any)	Version/ref of summary	sum_v3_thr884c
View start time (ISO)	When panel opened	2025-09-12T17:06:05Z
View duration (ms)	Time panel was visible/in focus	42000
Was this thread in AIA curated set?	For linkage	Y/N

Why: lets us see if curated AIA threads were actually read, and how intensively summaries are used.

Template F — Export “Moderation Interactions” (row = one report action)

We need	What it means	Example
Site	Braşov / Martin	Martin
Pseudonymous user ID	Stable, salted	u_71a2...
Session identifier	Link to sessions	s_5c11...
Post identifier	The item reported	p_33de...
Thread identifier	Thread context	thr_991a...
Report time (ISO)	When reported	2025-09-12T18:03:10Z
Report category (broad)	Harassment / Hate / Spam / Other / Unknown	Hate
System outcome shown to user?	Whether UI displayed decision	Yes/No
Outcome (if shown)	Allow / Remove / Escalate / Pending	Remove
Was this post in AIA set?	For linkage	Y/N

Why: aligns with AIA (moderation fairness) and tells us whether users see outcomes.

Template G — Export “Accessibility Usage” (row = one toggle/use)

We need	What it means	Example
Site	Braşov / Martin	Braşov
Pseudonymous user ID	Stable, salted	u_9b3f...
Session identifier	Link to sessions	s_2f1a...
Feature	Font increase / High contrast / Reduce motion / Keyboard-only / Screen reader hint	High contrast
State	On / Off / Used	On
Page area	Thread / Topic list / Composer / Settings	Thread
Time (ISO)	When used	2025-09-12T17:07:10Z

Why: quickly surfaces common barriers without collecting medical info.

Template H — (Optional) Pseudonymous Cohort Roster

We need	What it means	Example
Pseudonymous user ID	Stable, salted	u_9b3f...
Site	Braşov / Martin	Martin
Role group (if known)	Citizen / Staff / Moderator	Citizen
Consent date (ISO)	When consent recorded	2025-09-05
Declared accessibility tools	Free text or picklist	"Keyboard-only; Zoom"

Why: lets us split staff/citizen in aggregates; link with surveys via non-PII study code.

Template I — AIA Linkage Sheet (threads & items in curated tests)

We need	What it means	Example
AIA item type	Thread / Post / Pair	Thread
Identifier	How devs/admins refer to it	thr_884c...
Human label	Friendly name	"Bike lanes thread"
Belongs to site	Braşov / Martin	Braşov
In live traffic (Y/N)	Appears to users	Y
Notes	Anything special (language, sensitive)	"RO, long debate"

Why: gives a clean join between AIA test items and usage (E/F) for context.

Minimal metric definitions (to avoid ambiguity)

- **Active user (per day):** a pseudonymous user with ≥ 1 action (post/comment/react/report/summary view) or ≥ 5 minutes active that day.
- **Session:** a continuous period while logged in, ending at logout or after 10 minutes idle.
- **Summary view:** the summary panel was opened and visible for > 2 seconds.

- **Report submitted:** user pressed “Report” and the request reached the backend (regardless of outcome).
- **Accessibility used:** any accessibility feature toggled “On” or “Used” during a session.

Mapping worksheet (for your backend/exports)

We need (from templates above)	Where it lives in your system (table, API, admin export)	How to retrieve (query/report name)	Notes
Session start/end			
Summary views (count/duration)			
Reports & outcomes			
Accessibility toggles			
Thread/day aggregates			

Fill once per site; keep with your data pull scripts.

Privacy & retention (copy into your request)

- No PII (no names/emails/IPs/precise location).
- Use a **stable, salted pseudonymous ID**; keep the salt key outside shared exports.
- Retain raw exports only until the Phase-2 report is delivered [**insert date**], then delete or fully anonymise.
- Access limited to the evaluation team.

Annex 6. Gamification Module evaluation

Annex 6.1: Single Usability & Experience Session protocol

0) What we’re evaluating (plain English)

We want to see if the **missions** → **rewards (XP/badges/levels)** → **feedback** → **leaderboard** loop is:

- **Understandable** without explanations,
- **Motivating** (competence, relatedness, activity),
- **Technically solid** (events fire, UI updates fast, logs recorded).

Suggested topic for UPAT students (pick one and seed content accordingly) OR you may replace it with the one you will decide:

- Campus mobility** (bus frequency, parking, bike lanes),
- Digital campus services** (wifi dead zones, e-secretariat, exam scheduling),
- Food & study spaces** (canteen hours, library seating, quiet zones).

1) People & roles

Participants

- N = 10–12 UPAT students** (18–28, mixed faculties if possible).

- ❑ Mix of heavy social-app users, casual forum users, and **≥3 with little/no gamification experience.**
- ❑ Incentive: **coffee voucher** or canteen card or other, course related.

Staff (3 people)

- **Facilitator (leads session)** — reads the script, keeps time, stays neutral.
- **Note-taker/Observer** — logs issues, quotes, timestamps (uses the checklist below).
- **Tech/Logger** — prepares accounts, seeds content, watches backend logs.

2) What to prepare the day before

Accounts & content

27. Create **12 test accounts**: stu01 ... stu12 (passwords on printed slips).
28. Seed **5–8 proposals** on the chosen topic; add a few realistic comments.
29. Ensure **missions** and **badges** are active; **reset leaderboard.**

Missions (copy as-is or use your own)

ID	Mission	Trigger	X P	Badge	Difficulty	Motivation tag
M1	Make your first post	Create 1 proposal	10	Competence (bronze)	Easy	Competence
M2	Join the dialogue	Write 2 comments	10	Relatedness (bronze)	Easy	Relatedness
M3	Civic vote	Vote on 3 items	8	Activity (bronze)	Easy	Activity
M4	Helpful reply	Receive 1 upvote	12	Relatedness (silver)	Medium	Relatedness
M5	Keep it going	Log in 2 days in a row (active today + tomorrow)	10	Activity (silver)	Medium	Activity
M6	Quality proposal	Receive 2 positive evaluations	15	Competence (silver)	Medium	Competence
M7	Influencer	Enter top-5 leaderboard	20	Competence (gold)	Hard	Competence
M8	Community builder	Interact with 3 different users	15	Relatedness (gold)	Hard	Relatedness

Backend event logging (turn this on)

For each event, log as CSV or JSON (optional)

user_id, timestamp_iso, action_type (post/comment/vote/reply/login),
 mission_id, mission_status (viewed/completed),
 xp_delta, level_before, level_after,
 badge_id, badge_category (competence/relatedness/activity), badge_difficulty,
 leaderboard_rank_after,
 ui_area (profile/active_missions/badges/leaderboard),
 session_id, session_start, session_end,
 feedback_received (upvotes on their content during session).

Room & materials

- Quiet room, projector or large screen, **stable Wi-Fi.**
- Printed: **participant list, account slips, consent, Pre/Post questionnaires** (hardcopies in case online does not work), **Micro-survey cards.**

- Stopwatch or timer. Optional screen recording.

3) Session timetable (2.5–3 hours)

Block	Time	What happens
A. Welcome & consent	10'	Brief purpose, privacy; sign consent; hand out accounts
B. Pre-questionnaire	10'	Baseline motivation & expectations
C. Guided tasks (think-aloud)	50'	Step-by-step missions + micro-surveys after each reward
D. Mid mini-survey	5'	Flow & quick UX clarity check
E. Open exploration	20'	Free play: climb rank, earn a badge
F. Post-questionnaire	12'	GAMEX + gamification-UI usability
G. Debrief (group)	10–15'	What worked, what didn't; suggestions
H. Wrap-up	5'	Incentives, thanks

4) EXACT scripts & instructions

A) Facilitator welcome (read verbatim, 1–2 minutes)

“Hi, and thank you for joining. Today we’re testing how a reward system (missions, points, badges, and a leaderboard) **feels** to use. We’re testing the **system**, not you. Please think aloud: say what you expect, what you’re looking for, and what confuses you. You can skip anything and stop at any time. We’ll collect your clicks and short answers; all data are pseudonymous.”

Consent & accounts: hand out slips; make sure everyone logs in.

B) Pre-questionnaire (collect on paper or Form)

Scale: 1 (Strongly disagree) ... 7 (Strongly agree). Copy these items exactly.

Interest/Enjoyment (IMI-short, 2 items)

19. I think using this platform today will be interesting.

20. I expect to enjoy the tasks I’ll do.

Perceived Competence (2)

3. I feel confident I can complete missions here.

4. I can figure out how to earn points and badges.

Relatedness / Social Motivation (2)

5. Seeing others’ progress would motivate me.

6. I like feeling recognized by a community.

Autonomy (2)

7. I like choosing which tasks to do first.

8. I prefer systems that let me progress in different ways.

Familiarity (mark one)

9. I’ve used apps with points/badges before: No A little A lot

Expectation

10. Gamification will make participation more engaging for me.

C) Guided tasks — Think-aloud protocol

Moderator ground rules to self:

- 16) Don't teach; don't lead. Let participants discover.
- 17) Use **neutral prompts only** (below).
- 18) After **each visible reward** (points pop-up, badge, level, rank change), hand them a **Micro-survey card** (see section D).

Neutral prompts (use when stuck):

- "What are you looking for now?"
- "What do you expect will happen?"
- "Tell me what you're thinking."
- "What would you try next?"

Tasks (*read one at a time; 2–5 min each*)

- **Find missions.**
"Please open your **Profile** and find the list of **Active Missions**. Without clicking, tell me which one you'd choose first and why."
- **Complete M1 – Make a post.**
"Create a new proposal on today's topic (e.g., bus frequency to campus). Add a short title and 2–3 sentences."
- **Complete M2 – Join the dialogue.**
"Comment on two different proposals; say something concrete."
- **Complete M3 – Civic vote.**
"Cast votes on three items you agree/disagree with."
- **Check your rewards.**
"Open your **Profile** → **Points & Badges**. What changed? Explain in your own words what the XP bar and level mean."
- **See the leaderboard.**
"Open **Leaderboard**. Where are you now? What would move you up?"
- **Try to earn a badge (M4).**
"Reply helpfully to someone else and try to get an upvote."
- **Explain mission/badge logic.**
"In your own words: how do missions, points, and badges work here?"

Success criteria (observer checks silently):

- Finds Active Missions without help
- Completes M1–M3
- Spots XP change/feedback promptly
- Finds profile progress & badges
- Finds leaderboard & interprets rank

D) Micro-survey (after each visible reward)

Give this 30-second card or short online form every time a reward appears.

Scale: 1–7 unless noted

- Right now, how **satisfying** was that?
- I **understood why** I got this.
- I felt progress in **Competence** (skill/mastery).
- I felt progress in **Relatedness** (connection/recognition).
- I felt progress in **Activity** (momentum/energy).
- One word for the feeling: _____ (free text)

Observer note: write the **event type** on the card before handing it/ or online form field (mission, badge, level-up, rank change).

E) Mid mini-survey (5 minutes, after Task 8)

Flow Short Scale (1–7):

- I felt fully absorbed in what I was doing.
- I felt in control while doing the tasks.
- I found this enjoyable.

UX clarity (UEQ-S pairs, 1–7):

Confusing–Clear, Complicated–Easy, Inefficient–Efficient

Open (1 line): What helped or hurt your motivation so far?

F) Open exploration (20 minutes)

“Now play freely. Try to **earn any badge you don’t have and move up at least one rank**. Say aloud what you’re attempting.”

Observer watches for:

11. Self-set goals,
12. Reward chasing behaviour,
13. Frustration (no feedback, unclear thresholds),
14. Social interaction attempts.

G) Post-questionnaire (10–12 minutes)

GAMEX short (1–5):

- **Enjoyment** — I felt positive emotions while earning rewards.
- **Absorption** — I was immersed while completing missions.
- **Activation** — The system energized me to act.
- **Creative thinking** — I experimented to reach goals.
- **Absence of negative affect** — The system rarely frustrated me.
- **Dominance/Competence** — I was aware of my progress/improvement.

Gamification-UI usability (1–5):

7. The **Missions** screen was easy to understand.
8. The **Badges** were self-explanatory.
9. My **progress** was always visible.
10. **Feedback timing** felt right (I saw rewards when I expected to).
11. The **Leaderboard** felt useful and fair.

Adoption (1–5):

12. I would like this gamification in real student participation.

Open:

13. What one change would most improve your experience?

H) Group debrief (10–15 minutes)

- “What felt **most rewarding** and why?”

- “What felt **pointless or confusing**?”
- “How should we **pace missions** so it’s lively but not spammy?”
- “What would make badges feel **meaningful** (not ‘empty points’)?”

Facilitator rules: Let everyone speak once before a second round; keep it neutral; capture exact phrases.

5) What the observer writes down (checklist)

- **Findability:** Needed hints to find Missions? Profile? Badges? Leaderboard?
- **Vocabulary confusion:** words like “missions”, “badges”, “level”, “XP”, “rank”.
- **Feedback timing:** reward appears quickly? delayed? missed?
- **Mismatch:** action taken ≠ reward expected (note which).
- **Social cues:** asked peers for upvotes? reacted to leaderboard?
- **Frustration signals:** sighs, back-and-forth, “where is...?”, giving up.
- **Quotes** (verbatim, short) tied to a moment.

Add **timestamps** for major events (to match logs later).

6) Data to collect

A) Backend event logs (CSV)

One row per event:

user_id,timestamp_iso,action_type,mission_id,mission_status,xp_delta,level_before,level_after,badge_id,badge_category,badge_difficulty,leaderboard_rank_after,ui_area,session_id,session_start,session_end,feedback_received

stu03,2025-10-

14T10:23:12Z,comment,M2,completed,10,1,1,relatedness_bronze,relatedness,bronze,8,active_missions,abc-123,10:00,11:55,1

B) Paper/digital instruments

- Pre-questionnaire, Micro-surveys, Mid mini-survey, Post-questionnaire.
- Observer checklist sheets.

C) Optional screen recording

If available, note recording filenames per participant.

7) Quick analysis plan (same day/next morning)

Compute (Excel/Sheets is enough):

- Task completion rates (M1–M3 %) and **time to first reward**.
- Average **micro-survey** scores by event type (mission vs badge vs level vs rank).
- **Flow & clarity** means (mid), **GAMEX & usability** means (post).
- **Top 5 issues** from observer notes & frequency.

Success thresholds:

- $\geq 80\%$ complete M1–M3 without moderator help.
- Micro-survey **satisfaction $\geq 5/7$** and “**understood why**” $\geq 5/7$.
- Post **GAMEX enjoyment & activation $\geq 4/5$** ; usability $\geq 4/5$ (\approx SUS 70+).
- ≤ 3 repeating confusion points.

8) Debrief & fix list (within 48h)**One-pager structure (copy this):**

- **Who/where/when** (N, topic, version tested).
- **What worked** (bullet list of top motivators; include one quote).
- **What to fix** (top 5 friction points; each with: **issue** → **impact** → **fix** → **owner** → **target date**).
- **Key metrics** (task success %, mean micro-survey, GAMEX, usability).
- **Next steps** (quick wins + any deeper changes to consider).

9) Risk & troubleshooting

- **People wait for instructions:** Remind them “There isn’t a right path; say what you think and try what you’d do.”
- **No upvotes appear:** Have one staff account ready to upvote a genuinely helpful reply (don’t overdo it).
- **Reward doesn’t show:** Ask participant to show where they think it should appear; log a **feedback latency** issue.
- **Conflicting jargon:** If multiple labels exist (e.g., “points” vs “XP”), note it; don’t explain—this is a finding.
- **Tech hiccup:** If the platform stalls $>60s$, skip to next task and note **interruption**.

10) Ethical & data handling

- Use **pseudonymous** accounts only; no personal data in logs.
- Store logs/forms on the project drive; restrict to the evaluation team.
- Participants may stop anytime; still receive the incentive.

11) Ready-to-paste recruitment text (students)

“**Help us test a new reward system** for student participation. One **2.5–3h session** on campus. You’ll post, comment, vote, and try to climb a leaderboard about [choose topic]. We give **coffee vouchers** and a **certificate of participation**. No experience needed.”

12) Handouts & templates (copy blocks below)**Micro-survey card (print 12 per participant)**

- Event: Mission Badge Level-up Rank change
- Satisfaction **right now** (1–7): __

- I understood why (1–7): ___
- Competence (1–7): ___
- Relatedness (1–7): ___
- Activity (1–7): ___
One word: _____

Observer sheet (per participant)

- Start time: ___ End time: ___
- Found **Missions** without help? Y N (notes)
- Completed **M1/M2/M3**? Y N
- Saw feedback quickly? Y N (sec est.)
- Found **Badges**? Y N
- Found **Leaderboard**? Y N
- Confusions/quotes (timestamped):
 - [10:32] “Where do I see points?”
 - [10:41] “Why no badge?”
- Handoffs/interventions: ___

Post debrief capture (group)

- Most rewarding (list): ___
- Pointless/confusing (list): ___
- Pacing suggestion: ___
- One change to ship first: ___

Annex 6.2: Instruments

Gamification Module: Single Usability & Experience Session (2.5–3h)

Session Header (staff)

Date: _____ Time: _____ Location: _____

Facilitator: _____ Note-taker: _____ Tech/Logger: _____

Topic seeded: Campus mobility Digital campus services Food & study spaces Other: _____

Version tested: _____ (commit/build/date)

Participant IDs (stu01–stu12): _____

Consent One-Liner (read aloud)

“Participation is voluntary. We are evaluating the system, not you. Your account is pseudonymous and no personal data will be stored in the logs. You may skip any question and stop at any time without penalty.”

Pre-Questionnaire (Baseline Motivation & Expectations)

Participant ID: _____ Time start: _____ Time end: _____

Tick one box per row. Scale: 1 (Strongly Disagree) ... 7 (Strongly Agree).

Item	Statement	1	2	3	4	5	6	7
1	I think using this platform today will be interesting.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	I expect to enjoy the tasks I'll do.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	I feel confident I can complete missions here.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	I can figure out how to earn points and badges.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	Seeing others' progress would motivate me.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	I like feeling recognized by a community.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	I like choosing which tasks to do first.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	I prefer systems that let me progress in different ways.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Item	Statement	1	2	3	4	5	6	7
9	Expectation: Gamification will make participation more engaging for me.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Scale anchors: 1 = Strongly Disagree ... 7 = Strongly Agree

Familiarity with gamification (select one): No A little A lot

Micro-Survey Card (30s; AFTER each visible reward)

Print multiple copies; staff pre-marks the event type.

Event type: Mission completed Badge earned Level-up Rank change Time:

Scale: 1 (Not at all) ... 7 (Very much). Tick one box per row.

Item	Statement	1	2	3	4	5	6	7
1	Right now, how satisfying was that?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	I understood why I got this reward.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	I felt progress in COMPETENCE (skill/mastery).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	I felt progress in RELATEDNESS (connection/recognition).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	I felt progress in ACTIVITY (momentum/energy).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Scale anchors: 1 = Not at all ... 7 = Very much

One word for the feeling: _____

Mid Mini-Survey (Flow & UX clarity) – 5 minutes

Complete once mid-session. Scale: 1 (Strongly Disagree) ... 7 (Strongly Agree).

FLOW (FSS-short):

Item	Statement	1	2	3	4	5	6	7
1	I felt fully absorbed in what I was doing.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	I felt in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Item	Statement	1	2	3	4	5	6	7
	while doing the tasks.							
3	I found this enjoyable.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Scale anchors: 1 = Strongly Disagree ... 7 = Strongly Agree

UX Clarity (UEQ-S pairs; choose one value 1–7 on the continuum):

Confusing 1 2 3 4 5 6 7 Clear

Complicated 1 2 3 4 5 6 7 Easy

Inefficient 1 2 3 4 5 6 7 Efficient

Open: What helped or hurt your motivation so far? _____

Post-Questionnaire (GAMEX + Gamification-UI Usability and Adoption)

Scale: 1 (Strongly Disagree) ... 5 (Strongly Agree).

GAMEX Short:

Item	Statement	1	2	3	4	5
1	Enjoyment — I felt positive emotions while earning rewards.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	Absorption — I was immersed while completing missions.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	Activation — The system energized me to act.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	Creative thinking — I experimented to reach goals.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	Absence of negative affect — The system rarely frustrated me.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	Dominance/Competence — I was aware of my progress/improvement.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Scale anchors: 1 = Strongly Disagree ... 5 = Strongly Agree

Gamification-UI Usability:

Item	Statement	1	2	3	4	5
1	The Missions screen was easy to understand.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	The Badges were self-explanatory.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	My progress was always visible.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Item	Statement	1	2	3	4	5
4	Feedback timing felt right (I saw rewards when I expected to).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	The Leaderboard felt useful and fair.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Scale anchors: 1 = Strongly Disagree ... 5 = Strongly Agree

Adoption intent:

Item	Statement	1	2	3	4	5
1	I would like this gamification in real student participation.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Scale anchors: 1 = Strongly Disagree ... 5 = Strongly Agree

Open: What one change would most improve your experience?

Observer Checklist (per participant)

Participant ID: _____ Start: _____ End: _____

Found Active Missions without help Y N (notes)

Completed M1 (post) Y N Completed M2 (comments) Y N Completed M3 (votes) Y N

Noticed XP change promptly Y N Estimated delay (sec): _____

Opened Profile → Badges Y N

Opened Leaderboard Y N

Vocabulary confusion (missions/badges/level/XP/rank):

Frustration points (timestamp + note):

Representative quotes (timestamped):

Group Debrief Capture (facilitator)

• Most rewarding (list top 3):

- Pointless/confusing (list top 3):
-

- Pacing suggestions (frequency of missions/rewards):
-

- One change to ship first:
-