

From Voices to Shared Understanding: AI-Supported Dynamic Participatory Sensemaking in Citizen Deliberations

Preprint. Version dated September 2026.

JOSHUA C. YANG, ETH Zurich, Switzerland

MAUD REVEILHAC, LUT University, Finland

AURELIA TAMÒ-LARRIEUX, University of St.Gallen, Switzerland

DIRK HELBING, ETH Zurich, Switzerland and Complexity Science Hub, Austria

KENNETH HOLSTEIN, EPFL, Switzerland

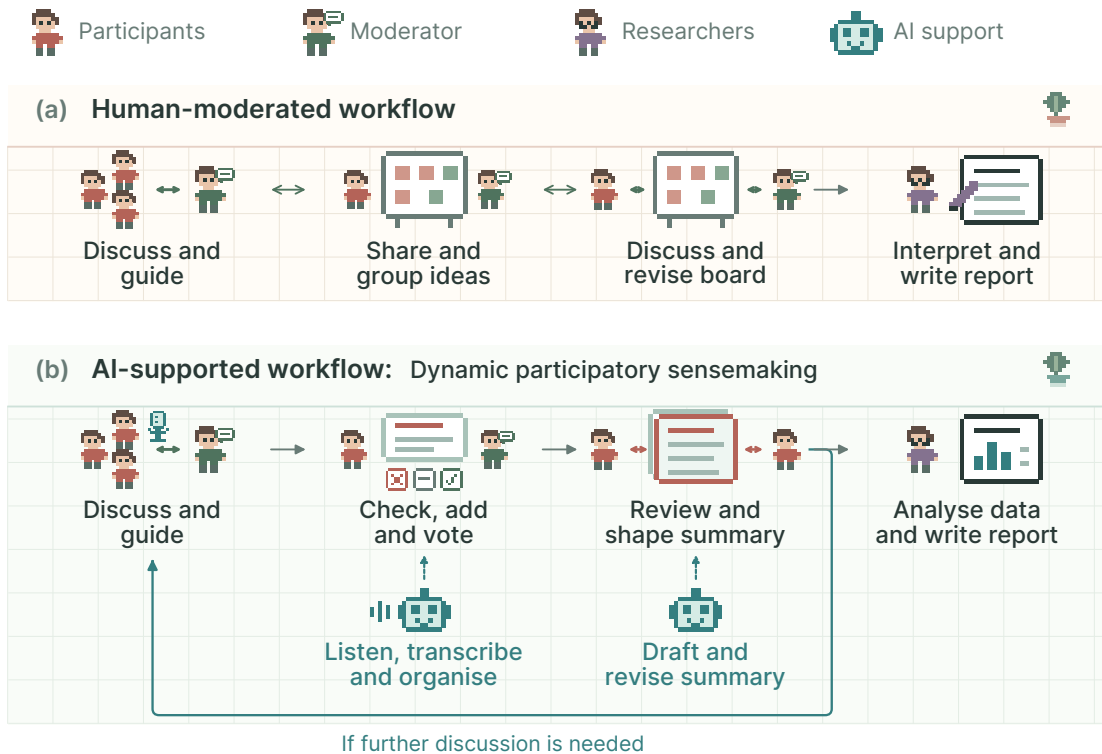


Fig. 1. The two deliberation workflows examined in this study. (a) Participants discuss and revise ideas on a shared board. (b) AI supports transcription, voting cards and summaries, which participants use to guide further discussion and revision through *dynamic participatory sensemaking*. Both workflows involve human facilitation and researcher reporting.

Citizen deliberation offers people with different experiences and priorities a way to shape responses to public problems together. Yet power to shape deliberation outputs often remains in the hands of a select few. Advances in AI offer potential to enable more

Authors' addresses: Joshua C. Yang, Computational Social Science, ETH Zurich, Zurich, Switzerland, joyang@ethz.ch; Maud Reveilhac, Social Sciences, LUT University, Lappeenranta, Finland, maud.reveilhac@lut.fi; Aurelia Tamò-Larrieux, University of St.Gallen, St.Gallen, Switzerland, Aurelia.Tamo@unisg.ch; Dirk Helbing, ETH Zurich, Zurich, Switzerland and Complexity Science Hub, Vienna, Austria, dhelbing@ethz.ch; Kenneth Holstein, EPFL, Lausanne, Switzerland, ken.holstein@epfl.ch.

participatory modes of summarisation, where participants co-construct evolving summaries *during* live deliberations. We developed Murmi to explore the potential for AI-supported *dynamic participatory sensemaking*, by integrating live discussion, voting on summarised position cards, and contestation of summaries. Across four workshops, questionnaire responses from 76 participants more often favoured the AI-supported workflow for perceived output accuracy (51% versus 38% for human moderation), but favoured a human-moderated workflow for feeling heard. Participants valued opportunities to express views and question wording, emphasis and missing concerns. These findings inform the design of AI-supported deliberation in which citizens shape not only the discussion, but how their collective views are interpreted and represented.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: AI-augmented deliberation, participant-facing reporting, micro-voting, citizen workshops, hybrid deliberation and voting

1 INTRODUCTION

Democratic participation holds the promise that people with different experiences and priorities can shape their shared future together. Citizen deliberation creates space to consider what they share, where they disagree and what they might do together [3, 13]. Realising this promise depends partly on how people’s contributions are heard, interpreted and carried forward. AI opens possibilities for involving citizens in this process by connecting ongoing discussion with voting and summary review. We explore what these possibilities mean in practice through dynamic participatory sensemaking in live citizen deliberation.

Voting gives participants another way to respond to one another’s views alongside spoken exchange [29, 34]. It can help focus discussion and preserve dissent amid pressures for consensus [7]. During a live conversation, these responses could identify ideas that need further explanation or discussion. Yet a vote responds to a statement as it has been worded; the reasons, conditions or uncertainty behind a position may be harder to express. The challenge is how voting can enrich the conversation while leaving room for people to clarify what they mean.

HCI research has explored AI-supported discussion and collaborative summarisation [8, 30, 33]. Pol.is lets participants contribute statements and respond to those of others, making patterns of agreement and difference visible [29]. In processes such as vTaiwan, these patterns inform deliberation that connects online participation with discussions among citizens and stakeholders [29, 35]. We examine how spoken contributions, voting and summary review can remain connected during a live citizen workshop. Participants can help shape the voting cards and summaries while using them to identify what needs further discussion.

Work on participatory sense-making examines how meaning develops through interaction between people [11]. In a workshop, a participant might notice that a voting card has left out a condition attached to their proposal and explain why it matters. This can prompt clearer wording and further discussion. We use *dynamic participatory sense-making* to describe this continuing relationship between discussion and how contributions are represented in cards and summaries. Participants interpret these materials together, and their responses can shape both the wording and the conversation that follows. AI helps prepare the materials, while participants and moderators draw on their experience of the exchange to question what they contain. Participation therefore includes having a say in how views are understood and brought together.

To explore this process, we developed *Murmi*. The platform transcribed discussion and extracted propositions for moderators to check before presenting them as voting cards. Participants indicated their positions through brief judgments, or *micro-votes*, added statements of their own and reviewed summaries generated from propositions and responses. The design aimed to help them identify what needed further discussion (Figures 1 and 2).

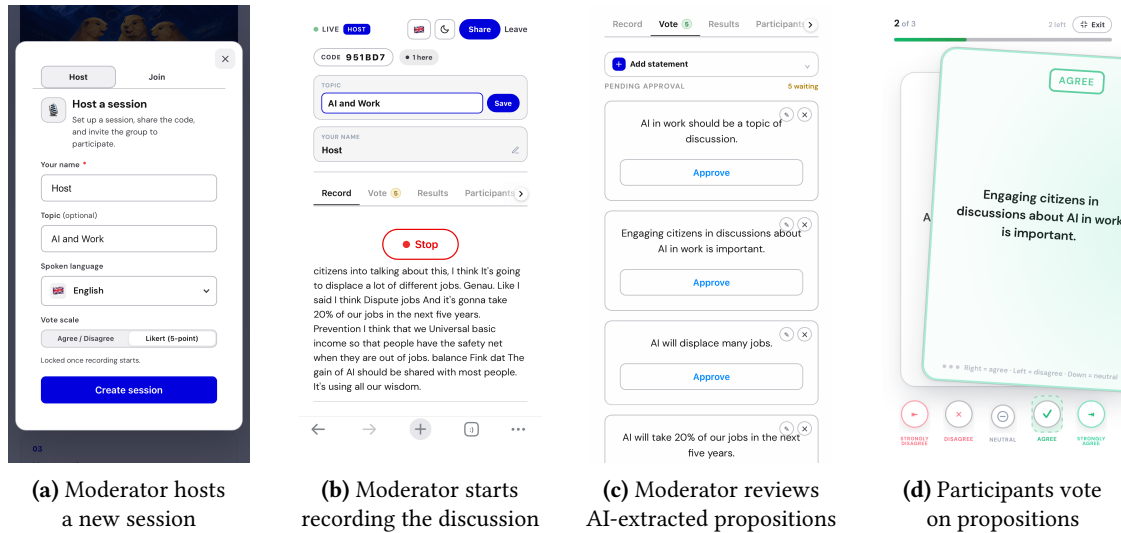


Fig. 2. Platform workflow, panels (a–d). The moderator hosts a session, starts recording (after getting everyone’s informed consent) and reviews AI-extracted propositions. Participants vote on (moderator-approved) propositions. These panels show an illustrative session. The workflow continues in panels (e–h).

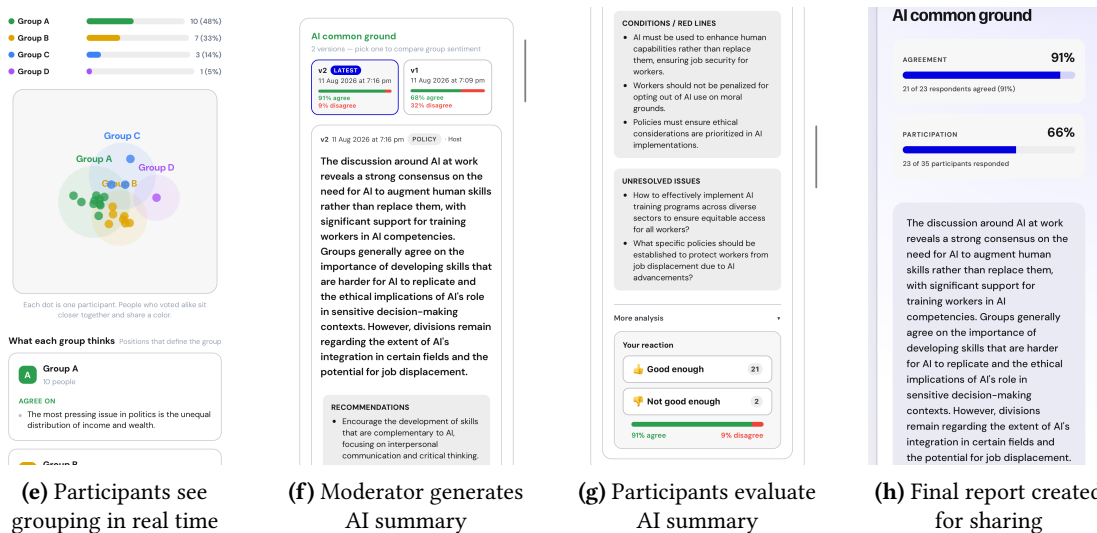


Fig. 2. Platform workflow (continued), panels (e–h): live grouping, summary generation, participant review and reporting. Review includes unresolved issues and whether the summary is “good enough”. Panels (f–h) show aggregate Zurich English workshop data.

Whether these opportunities become part of a shared conversation depends on how people use them. Reading cards and responding to them also take attention during discussion. Moderators decided which extracted propositions reached participants. Their screen was projected so participants could follow and question edits. We examine how overlapping discussion, voting and review worked in practice, including what participants valued and how requests for changes were handled.

We studied four language-group workshops across two Swiss events and three languages. Each room followed the human-moderated workflow and then the AI-supported workflow. Questionnaires from 76 participants and workshop records let us examine their experiences alongside exchanges in which people questioned or tried to change cards and summaries. We ask:

- (1) How did participants assess the discussion outputs and their opportunities to contribute and deliberate in the human-moderated and AI-supported workflows?
- (2) How did participants and moderators work with voting cards and summaries during discussion, and what did they accept, question or seek to change?

We make three contributions to understanding dynamic participatory sensemaking in live citizen workshops. First, we trace how contributions were represented and contested across transcripts, voting cards, responses, summaries and reports. Second, we identify the articulation work of participant review: checking wording, explaining omissions and coordinating changes while discussion continues. Third, we derive design implications for connecting discussion, voting and review: making sources and revisions visible, providing time to consider and correct cards and summaries together, and distinguishing support for propositions from acceptance of a summary and authorisation of a public report.

2 RELATED WORK

We draw on research on how people make sense of discussion with others, how voting relates to deliberation, and how participants review summaries of a group's views.

2.1 Making sense of discussion together

De Jaegher and Di Paolo examine how meaning develops and changes as people respond to one another [11]. We use *dynamic participatory sensemaking* to examine this continuing process when contributions return as AI-generated cards and summaries. Participants can question and revise the cards and summaries, then return to discussing the ideas they contain.

Collaborative systems already support this relationship between discussion and representation. Wikum lets people recursively summarise discussions, while Wikum+ interweaves discussion and summarisation so that summaries can develop as contributions and feedback arrive [33, 36]. MeetMap examines real-time dialogue mapping in online meetings. In a study with ten pairs, participants valued the opportunity to construct maps from AI-generated nodes, as well as the lower effort of editing maps generated by AI [8]. In each system, participants help organise the discussion, though the opportunities to do so and the effort required differ.

AI can also affect how people encounter one another's contributions. Research on AI-mediated communication examines systems that modify or generate communication on a person's behalf [15]; Argyle and colleagues studied suggested rephrasings during political conversations [2]. Research on collaborative programming similarly examines

how AI assistance enters collaboration between peers [10]. Revel and Pénigaud propose AI as a reflector that supports collective sensemaking and agreement-seeking without making binding decisions [27].

Collaborative summarisation also requires people to coordinate their work. Articulation work concerns organising interdependent activities [28], including who prepares a summary, who checks it and who responds when someone asks for a change. Such effort may be less visible than the finished output [31]. In a live workshop, people need time for these tasks alongside speaking and listening.

Our study extends this work on collaborative summarisation and dialogue mapping by examining how participants and moderators work through the meaning of AI-generated cards and summaries during live citizen discussion. We trace exchanges in which people clarify a contribution, question its representation and coordinate changes. The observations show what participants and moderators do when they interpret and revise cards and summaries during discussion.

2.2 Connecting voting with live deliberation

Deliberation and voting offer complementary ways to participate, while their combination raises questions about how positions relate to reasons and discussion [7, 34]. ConsiderIt connects positions with participant-authored pro and con points [20], and PolicyCraft supports discussion and voting on concrete cases [21]. Pol.is gathers participant statements and responses to map patterns of agreement and difference [29]. In vTaiwan, opinion mapping forms part of a process connecting online participation with deliberation among citizens and stakeholders [29, 35]. In these settings, voting accompanies opportunities to consider other people’s contributions.

Prior work has combined opinion mapping with spectrum-based voting to make positions visible before and after discussion, while identifying fixed statement wording as a limitation when ideas evolve during deliberation [35].

Research on AI-augmented deliberation has explored support for several activities, including structuring contributions, assisting facilitators and developing collective statements. BCause presents an architecture for transforming transcripts into structured arguments that moderators can curate and use in reports [1]. In a Finnish Citizens’ Assembly, Jäske and colleagues reported that AI eased the workload of facilitators and deliberators in co-creation and integration, but encountered shortcomings when creating questions for experts [18]. Konya and colleagues combined participant contributions and voting, AI-assisted generation of collective statements, multilingual human review and final votes among Israeli and Palestinian peacebuilders. The resulting statements received at least 84% agreement within each participating group [19].

Building on approaches that connect participant statements, voting and collective outputs, we developed and studied a workflow in which spoken contributions can become voting cards while discussion continues. Our field study examines how participants respond to new voting cards while also speaking and listening. It shows how additional ways to express a position can be valued even when participants want more shared discussion.

2.3 Judging and shaping a discussion summary

Generating statements that participants endorse and summarising a group’s discussion involve related but different judgements. Building on work on generating statements that accommodate diverse preferences [4], Tessler and colleagues’ Habermas Machine generates group statements from written individual opinions and revises them using participants’ critiques. In its human-mediator comparison, participants preferred the AI-generated statements to those

written by other study participants trained for the mediator role [32]. Small and colleagues likewise propose returning AI-generated summaries to participants for checking [30]. These approaches make review part of producing a summary.

Support for separate statements does not by itself settle how they should be brought together. Judgement aggregation examines problems that arise when votes on connected propositions are combined [23], while work on textual aggregation considers criteria for representing diverse views in generated outcomes [14]. Writing a summary also involves deciding which claims to emphasise and how to connect them. Research on under-representation in group summaries examines whose perspectives remain visible after these choices [37]. Participants therefore need to judge both the individual propositions and how the summary represents the group.

Participants may also have information that the system lacks. Holstein and colleagues' work on perceptual complementarity examines how people draw on information unavailable to an AI model. Their experiment found that prompts about such information changed how people combined it with model outputs, without necessarily improving prediction performance [17]. In a workshop, participants can draw on their experience of the discussion when explaining what a card or summary misses. Meeting-extraction research likewise illustrates difficulties in recovering context and intent from spoken interaction [24].

Review also raises questions about who can shape the summary and what their approval means. Boswell describes the difficulty of turning diverse contributions into recommendations under time pressure [6]. McKinney argues for participant verification and contextualisation of AI-generated material in citizens' assemblies [25], while work on participatory algorithmic governance connects involvement with influence over decisions [22]. Moderator approval of a card, participant agreement with it, acceptance of a summary and permission to publish a report on behalf of the group need to be distinguished. Studying review therefore requires attention to what participants were asked to judge and what happened when they objected.

Favourable experiences are also distinct from deliberative outcomes. In a charity allocation task, Parisi and colleagues found that participants preferred LLM facilitation even though measured consensus and participation equality did not significantly improve [26]. We therefore read participants' assessments alongside records of contributions, voting and review. This lets us examine how people used the opportunities available, without treating favourable ratings as evidence of improved deliberation.

Our study adds empirical evidence about the relationship between endorsing individual statements and accepting a summary of the group's discussion. We compare proposition responses with summary-review responses and examine participants' explanations of missing concerns, emphasis and uncertainty. The findings show that support for statements can coexist with rejection of the summary, while acceptance of a group summary can accompany personal disagreement with its recommendations.

3 STUDY DESIGN AND METHODS

We developed the platform for use in citizen deliberation, with support from civic groups. Its design drew on vTaiwan's approach to connecting online opinion gathering with face-to-face discussion [29, 35].

We conducted a participatory field study within a citizen consultation, examining participants' experiences and the handling of contributions in human-moderated and AI-supported workflows. We combined qualitative interpretation of comments and workshop records with questionnaire comparisons and descriptive activity analyses. Research-team members helped organise and facilitate the workshops and develop the platform. Their involvement informed our understanding of its use and meant that our own decisions shaped the setting we studied.

We ran four language-group workshops across two events. Each workshop included two topic discussions, each with a 45-minute time limit. The first followed the human-moderated workflow, using sticky notes and a shared board. The second followed the AI-supported workflow, using live transcription, voting cards and summary review (Figure 3). This sequence was chosen to familiarise participants with the workshop process before they used the platform. The study examined participants’ experiences of the two workflows and how contributions were handled during the AI-supported workflow. Fixed ordering and differences in topic and discussion structure preclude isolating the effect of AI; language, group size and moderator staffing also varied across rooms.

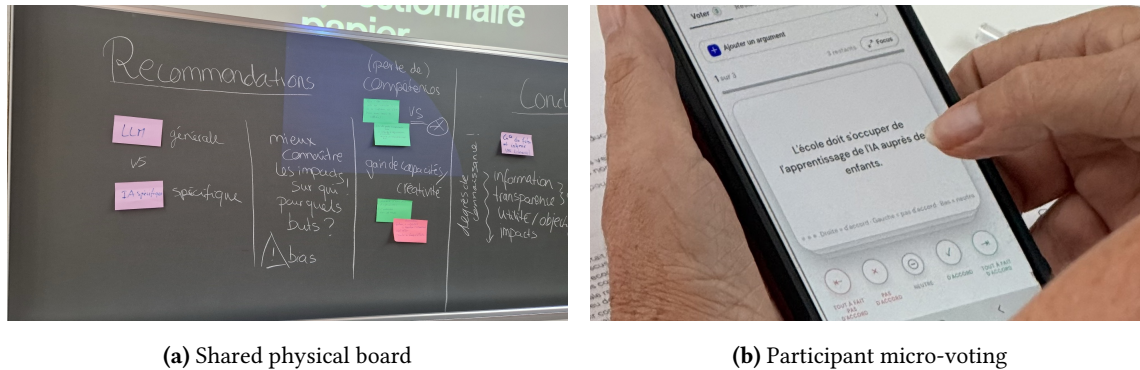


Fig. 3. The two studied workflows in practice. (a) In the human-moderated workflow, moderators organised participant sticky notes and clustered labels on a shared chalkboard. (b) In the AI-supported workflow, participants used their phones to respond to moderator-approved propositions during the room-wide discussion. The phone photograph was cropped to remove the session code and participant initials. Neither panel includes faces.

3.1 Workshop setting and topics

The workshops formed part of a consultation on AI in education and AI at work in August 2026. Each event used two separate rooms: English and German in Zurich, and English and French in Lausanne. We refer to each language-group session as a workshop. Participants stayed in the same room for both discussions. The English rooms discussed education using the human-moderated workflow and work using the AI-supported workflow. The German and French rooms reversed this topic assignment (Table 2). No discussion material was imported from the preceding online study.

A *card* displays a proposition for response. Participants reviewed the physical board or an AI-generated *summary* as a representation of the discussion. The platform’s finalised room *report* combined the AI summary with the propositions and responses available at that point. Researchers subsequently interpreted the workshop material for a separate citizen workshop report.

3.2 Human-moderated and AI-supported workflows

Both workflows began with small-group discussion so participants could formulate their ideas and have an opportunity to speak before sharing with the room. In the human-moderated workflow, groups of five or six organised sticky notes into recommendations, conditions or red lines, and disagreements. They then presented their views to the room. Moderators grouped and labelled contributions on a shared board, invited corrections and asked whether it fairly represented the discussion. Researchers photographed and digitised the board. No per-note agreement counts were recorded, so a note’s presence does not establish group support.

Each Zurich room had two moderators and each Lausanne room one, with unchanged staffing across formats. Moderators guided conversation and managed the board or platform. They did not submit participant responses.

During the room-wide discussion in the AI-supported workflow, participants spoke into a shared Bluetooth microphone, and the platform transcribed speech and extracted candidate propositions. Because audio-extracted cards were created through the moderator’s device, only the moderator could edit them. Moderators could also reject, hold or approve these cards before they appeared in the Murmi app. Review was intended to preserve meaning and avoid repetitive voting. The workshop lead recalls that moderators approved most extracted statements and helped edit wording that did not adequately reflect what a participant had said. The moderator’s interface was projected onto a large screen, showing propositions and response counts (Figure 4). Participants could follow edits on the screen. If an extracted card did not reflect their intended meaning, they could raise their hand and ask the moderator to change its wording or add what was missing.

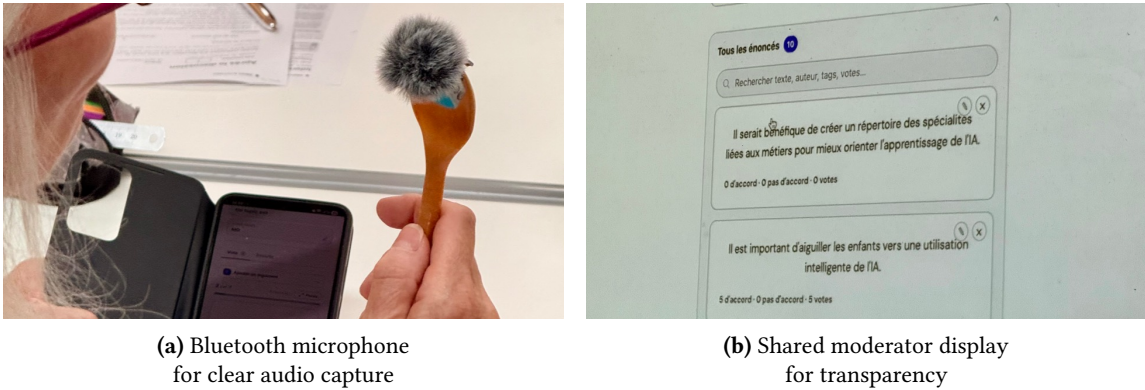


Fig. 4. The physical setup for the AI-supported workflow. (a) Participants took the shared Bluetooth microphone to speak, supporting audio capture for AI transcription and visibly marking whose turn it was. The workshop lead observed that holding it appeared to legitimise taking the floor and make turns more deliberate. (b) The projected moderator interface displayed propositions and response counts and made the review visible to the room. The moderator approved AI-extracted propositions for voting and could edit their wording for the sake of correct representation. Participants could question omissions or misrepresentation.

Participants responded on a five-point scale from strongly disagree to strongly agree. They could also type a proposition or speak a missing view for the platform to process. Participants could directly edit cards they had submitted themselves; changes to audio-extracted cards went through the moderator. The workshop lead recalls that Zurich German participants were initially encouraged to enter statements manually before the spoken round. The platform could also generate *tension cards*, statements intended to bring apparent differences in responses into discussion.

Near the end of the discussion, the platform generated a draft summary from approved propositions and responses. Participants were told to select “good enough” if they were satisfied with the summary and “not good enough” if they found it unacceptable. They could also explain their concerns. The question concerned the acceptability of the summary as a whole. Participants did not vote separately on each recommendation. We used approval by at least 80% of those who reviewed the summary as a practical threshold for ending review without another iteration. Further discussion could precede a moderator-requested revision within the 45-minute limit for the topic. Lausanne French remained below the threshold after its second summary, but the time limit prevented another iteration. Participants saw the finalised room report before completing the post-discussion questionnaire.

Table 1 compares responsibilities across formats. Researchers retained responsibility for interpreting the material and preparing the findings for reporting.

Table 1. Responsibilities, records and reporting artefacts in the two workflows. Both required facilitation, interpretation and participant review. The human-moderated workflow relied on manual note-making, organisation and summarising. The AI-supported workflow combined automated capture and drafting with human checking and correction. These descriptions concern task allocation and retained evidence, not measured workload or effort saved.

Aspect	Human-moderated workflow	AI-supported workflow
Participants	Discuss in small groups, organise sticky notes, share ideas with the room and question the board.	Discuss in small groups, share ideas with the room, vote, submit and edit their own typed cards, request moderator corrections to extracted cards, inspect groups and review summaries.
Moderators	Facilitate discussion, group notes, clarify meanings and revise the board with participants.	Facilitate discussion, review extracted cards and coordinate corrections, voting and summary review.
Researchers	Photograph and digitise boards and interpret the material for reporting.	Analyse transcripts, cards, votes and reviews; prepare retrospective voting maps and interpret findings.
Records and reporting artefacts	Board photographs and notes preserve selected contributions and groupings, without a complete transcript or per-note support.	Transcripts, cards, responses and summaries preserve captured speech and judgements; incomplete links and histories limit reconstruction.

3.3 Value-sensitive design, ethics, and participants

The platform and workshop procedures sought to protect privacy, faithful representation and participants' ability to question cards and summaries. These values could pull in different directions: making discussion available for inspection could help people check a summary while retaining more of what they said.

The ETH Zurich Ethics Commission approved the protocol (project 26 ETHICS-116) and subsequent amendment before fieldwork. Participants gave written informed consent to live transcription and generative processing, with restricted retention of transcripts and no project audio file retained. They could choose an equally compensated human-only route. Audio was processed transiently, and the approved European-processing configuration excluded provider training and retention. Research text was access-restricted. Moderators could see aggregate votes; individual votes were hidden from them. Publications use aggregate results and short anonymised excerpts without personal identifiers. These measures limited exposure without guaranteeing anonymity: people could recognise one another's contributions in the room, and retained text could contain identifying context. Not retaining audio also meant that later checks of transcription and interpretation had to rely on the available text.

Moderator review, projected edits and participant contributions were intended to preserve meaning and make generated wording open to challenge. Participants could question cards and reject summaries, allowing further discussion and revision.

Moderator choices still shaped wording and inclusion, and opportunities to object did not guarantee that a change preserved the intended meaning. We examine what happened when participants used those opportunities. Section 5.3 discusses moderator discretion and safeguards; the study does not establish that review prevented bias or manipulation.

Recruitment combined public outreach and a university participant pool, and was open to adults living in Switzerland who could participate in a workshop language. Participation was voluntary and compensation did not depend on expressed views.

The questionnaire analysis includes 76 respondents. Citizens were encouraged to complete the demographic questionnaire before the workshops, but completion was not required for attendance. Some attended without completing it, limiting the demographic information available for the workshop sample. Using study codes, we could match 51 workshop questionnaire respondents to demographic and AI-use information collected online beforehand (Table 2). This matched subset was relatively young, educated and familiar with AI. Of the 51 participants, 34 were under 35, 39 held a university degree and 42 used AI tools at least weekly. Mean ages were estimated from age bands.

Counting app activity. We count votes on cards and summary-review responses separately. Summary approval is the share of review responses marked “good enough”. App activity was recorded by account or device, so these counts may not correspond to distinct participants. We used anonymous identifiers to link votes and reviews, excluding host activity from these analyses. Appendix A.1 gives the counts used in each analysis.

Table 2. Workshop participants and topic assignment. The human-moderated workflow preceded the AI-supported workflow in every room. Survey counts refer to respondents, not total attendance. Completing the advance demographic questionnaire was encouraged but not required for attendance; some attendees had not completed it. Demographic counts, mean ages and gender describe the 51 survey respondents whose workshop questionnaires could be matched to these advance records using study codes. Gender is women/men/non-binary (W/M/NB). Mean ages use age-band midpoints of 21, 29.5, 39.5, 49.5, 59.5, and 70 years for the open-ended 65+ category.

Room	Survey <i>n</i>	Demographic <i>n</i>	Estimated mean age	W/M/NB	Human-moderated workflow topic	AI-supported workflow topic
Zurich German	20	16	33.2	9/6/1	AI at work	AI in education
Zurich English	35	20	26.8	13/7/0	AI in education	AI at work
Lausanne French	14	10	47.6	5/5/0	AI at work	AI in education
Lausanne English	7	5	41.5	3/2/0	AI in education	AI at work
Total	76	51	34.3	30/20/1		

3.4 Platform workflow and report production

The summary model received approved propositions, response counts, voting-pattern groups, jointly supported proposition pairs and available anonymous feedback on earlier summaries. It did not receive the full discussion transcript. Its instructions asked it to retain unresolved concerns and avoid treating one person’s suggestion as group agreement. Discussion, voting and review sometimes overlapped, allowing source material to change while a summary was being reviewed.

The platform checked joint support for pairs of propositions before suggesting that the summary model combine them: separate majorities can reflect different supporting coalitions [23]. The project-specific thresholds guided generation but had not been validated as criteria for collective agreement or logical consistency. Appendix A.2 gives the rule and its treatment of neutral and missing responses. Participant review assessed acceptance of the generated summary separately [30].

The platform displayed a live map of voting patterns using principal component analysis (PCA) and k-means clustering. Participants could locate their highlighted position and inspect statements supported or opposed within their assigned group. Groups reflected recorded responses and were updated when responses changed (Figure 2(e)).

Researchers also prepared retrospective voting-pattern maps for the separate citizen workshop report. Those maps used different data processing and do not reconstruct the live participant display; Appendix A.3 contains the maps and their stability checks.

We inspected the pre-workshop code, interfaces and exports. The inspected code specified `gpt-4o-transcribe` for transcription and `gpt-4o-mini` for proposition extraction and summary generation (Appendix A.2). Logs did not establish every deployed setting, including whether optional automatic approval was active. We therefore do not assume it was used. Cards lacked source passages and previous wording, edit times and editor identities. A shared recorder label covered different speakers and workshop stages, while extraction treated each passage as one contribution and returned at most one claim. These gaps constrained the tracing analysis below. Appendix A.2 documents the processing rules, grouping calculations and retained data.

We developed the platform for real-world use in citizen workshops and assemblies. Beyond the workshops reported here, a civic group outside Switzerland continues to use it in another language for monthly in-person deliberative workshops involving more than 100 citizens per session.

Code availability. The platform is open source and hosted on GitHub: <https://github.com/Carbon-copy-PS/murmi>.

3.5 Data collection and preparation

The study used shared boards from the human-moderated workflow, platform records from the AI-supported workflow and paper questionnaires about participants' experiences. These sources captured different aspects of each workshop.

Discussion boards. Researchers photographed the shared boards and transcribed notes into a dataset retaining room, topic, contribution type and grouping labels. The records preserve selected written contributions and their organisation, without an equivalent speech transcript or per-note voting dataset for the human-moderated workflow.

Platform records and summary reviews. The platform retained machine-generated speech transcripts, cards, available approval and edit metadata, and proposition responses. These records let us examine how spoken or typed contributions entered the voting material. No project audio was retained for retranscription, and edit and response histories were incomplete.

The platform retained successive summaries, supporting proposition references, participants' "good enough" or "not good enough" judgements and optional reasons, separately from the paper questionnaires. We compared these with finalised room-report snapshots. The retained response log and final-report aggregates differed, so we kept them distinct when reconstructing activity and support (Appendix A.1).

Participant questionnaires. Participants completed paper questionnaire sections after each discussion and a final comparison after both formats. After each discussion, they answered eight questions from 1 (strongly disagree) to 7 (strongly agree). Five assessed the output's coverage of views, recognition of minority views, visibility of disagreement, traceability and acceptability as a fair summary. Three assessed opportunities to correct the output, the feeling of being heard and *group control*. The group-control item asked about participants' experience and did not directly measure their decision-making authority. Questions referred to the physical board after the human-moderated workflow and the digital summary after the AI-supported workflow.

A ninth item asked about learning something new. An optional comment asked about missing or misrepresented views, disagreements or uncertainty. The supplement provides the full multilingual instrument.

After both discussions, participants compared output accuracy, the feeling of being heard, the handling of disagreement and future-workshop preferences. Scales ran from 1 (human much better) through 4 (about the same) to 7 (AI much better), with preference wording for future workshops. Two optional prompts asked about the most important difference and its effect, and what helped, hindered or should be changed in the platform.

Transcription, translation and linkage. Two team members handled both transcription and translation of their assigned workshop materials. The team member responsible for the Zurich workshops was a native German speaker and transcribed the sticky notes and questionnaires, translating the German-language material into English. The second author, a native French speaker, was the team member responsible for the Lausanne workshops. She transcribed and documented the questionnaires, prepared the Lausanne board material, and translated the French questionnaire responses and discussion transcripts into English. The questionnaire transcriptions were organised into a structured response dataset, with scans retained as source material. Responses originally written in English were used directly. The resulting English questionnaire texts formed the corpus for the thematic analysis described below. The first author analysed these English texts and did not undertake translation. Preparing those responses was separate from the platform’s automatic transcription of the spoken discussion.

We linked questionnaire responses to demographic information using study codes where available. Linking questionnaire responses to app activity required a separate match, and we identified any uncertainty. We connected participants’ descriptions to their votes or summary-review responses only where the evidence supported that match.

Context from facilitation. The workshop lead’s later reflections helped us understand the instructions, corrections, reporting work and differences between rooms. We distinguished these from participant comments and compared them with stored records where possible. They did not independently evaluate usability or workload.

3.6 Research analysis

Questionnaire ratings and written answers address participants’ experiences across formats (RQ 1). Transcripts, cards, response logs, summary versions and written review reasons support tracing how contributions and concerns entered the outputs of the AI-supported workflow (RQ 2); board records document the output of the human-moderated workflow. We describe the survey coding, illustrative excerpt interpretation and tracing procedures separately below.

3.6.1 Comparing participants’ ratings across formats. We averaged the eight representation-and-participation items when at least six were answered to form the study-specific *overall representation-and-participation rating*. Because this rating combines judgements about the output and participation, we give priority to the individual items when interpreting it. Complete-case internal consistency was $\alpha = 0.79$ after the human-moderated workflow ($n = 71$) and $\alpha = 0.88$ after the AI-supported workflow ($n = 72$) [9].

We calculated each participant’s AI-supported rating minus their human-moderated rating and averaged these differences across the 75 participants with paired overall ratings. Positive differences favoured the AI-supported workflow; negative differences favoured the human-moderated workflow. We applied the same calculation to each individual item using its available pairs. We report a 95% interval from resampling participants [12], sign-flip and Wilcoxon signed-rank tests, and the standardised paired effect d_z . Benjamini–Hochberg and Holm adjustments addressed multiple comparisons across the eight individual items and, separately, the four direct comparisons [5, 16]. We compared future-workshop preference across rooms using a permutation Kruskal–Wallis test. Because the formats differed in sequence, topic and discussion structure, these comparisons do not isolate the effect of AI.

3.6.2 *Qualitative analysis of questionnaire answers.*

Thematic analysis of survey answers. The first author developed and reviewed the codebook, drawing both on the questionnaire responses and on prior concepts, and conducted the thematic analysis. Using this codebook, the first author grouped the English versions of 148 written answers from 67 respondents into descriptive thematic categories. The unit of coding was one answer to one question, with multiple categories allowed per answer. The corpus comprised 57 method comparisons, 53 platform comments, 20 comments after the human-moderated workflow and 18 after the AI-supported workflow. Blanks and missing-value placeholders were excluded; reports of no concern were retained.

We used six categories for method comparisons, eight for platform comments and six for post-discussion comments. An answer could acknowledge useful synthesis while also describing difficulties in following the discussion. The supplementary codebook provides the categories, coding rules and examples.

All reported survey category counts derive from the first author's thematic coding. We also counted answers assigned to both a synthesis category and a discussion-related category within the same question, to examine whether these assessments occurred together in individual answers. This count used the existing category assignments; no new categories were introduced. Counts describe the submitted answers within each question; categories can overlap, and one respondent can contribute answers to several questions. They therefore cannot estimate the prevalence of an experience among workshop participants. Stable identifiers P1–P76 link questionnaire responses; uncertain links to app activity are identified separately.

Excerpt selection and interpretation. The second author selected and checked the quotations. Excerpts were translated where needed by the team members responsible for transcription and translation and edited for readability, with interpretive paraphrases marked. Selected comments were checked against respondents' recorded preferences.

The selected questionnaire excerpts cover all four rooms, different future preferences and positive, critical and mixed experiences. Post-discussion examples include a report of no missing content alongside concerns about representation. Appendix B.1 documents this illustrative selection.

3.6.3 *Following contributions from discussion to report.* We transcribed 140 board rows, retaining room, topic, contribution types and grouping labels. For the AI-supported workflow, the first author read all 143 transcript entries and 157 retained candidate cards chronologically, recording source, approval, duplication, meaning changes and workshop stage. A serious meaning change altered the actor, object or position enough to express a materially different claim. A stage error turned procedural speech into a policy proposition. Transcript-to-card matching relied on wording and timing because source links were absent from the stored records and no audio was retained under the consent protocol. The counts cover only clearly identifiable examples; they do not estimate error rates.

To follow correction episodes from discussion to cards and summaries, the first author compared passages across all four transcripts with card wording, creation times, edit flags, responses and summary references. Unreliable speaker identification precluded assigning survey codes to transcript speakers. Missing histories prevented complete reconstruction or linking a review judgement to a particular change (Appendix D). We examined exchanges about displayed groups alongside questionnaire comments, without reconstructing individual screen views or assignments.

To assess the evidence attached to the reported summary examples, we checked whether summary references resolved to published propositions, inspected the code handling them, and read contrasting linked-text examples alongside responses and comments. This distinguished a working reference from a proposition that substantively supported a summary claim (Appendix C.3).

3.6.4 Comparing votes and summary-review responses. Further checks covered 136 approved propositions, 45 summary items, 58 supporting links and four finalised room reports. The final aggregates contained 2,579 votes. The analysed event log contained 2,415 votes after excluding host activity. The incomplete timestamped log therefore supported different analyses from the final aggregates. Appendix B.3 gives their reconciliation and the coverage and timing calculations. We also examined whether later cards received fewer votes (Figure 9) and whether first-summary reviewers had been more active voters. The appendix also defines how response coverage was calculated.

The archive contained nine summary versions and 107 review responses. Seven versions were generated before the final room reports, with 105 review responses attached to them. The two later Zurich German versions accounted for the remaining two responses. Comparisons used anonymous identifiers. We paired judgements where identifiers matched and examined all ten written reasons. Appendix C.1 reports these records and the comparison between summary approval and support for cited propositions. Review responses lacked timestamps and reviewer sets changed across versions. We also compared each room’s last pre-report summary with the propositions and responses present when that report was finalised. Updated proposition responses retained only their latest value and timestamp.

4 FINDINGS

We begin with questionnaire ratings and written responses about the physical boards and digital summaries and participants’ experience of contributing across the two workshop formats (RQ 1). We then use activity records and recorded exchanges to examine how participants and moderators worked with cards and summaries while discussion continued. Finally, we compare votes with summary-review responses and examine participants’ reasons for accepting or questioning the group summary (RQ 2).

4.1 Participants distinguished expressing views from discussing them together

Questionnaire responses distinguished registering a position from taking part in a shared conversation. Some participants valued one while criticising the other in the same response.

4.1.1 Ratings of the discussion outputs and the experience of participating. After each discussion, participants rated the resulting summary and their experience of taking part. The combined representation-and-participation rating showed no clear overall difference: means were 5.40 after the human-moderated workflow and 5.17 after the AI-supported workflow, with a paired difference of -0.23 (95% CI -0.51 to $+0.04$, sign-flip $p = 0.104$). These contrasts describe the workshop formats as experienced, rather than an isolated effect of AI.

Group-control ratings averaged 0.76 points lower after the AI-supported workflow (95% CI -1.20 to -0.32 , Holm $p = 0.018$). Ratings concerning accuracy, minority views, disagreement, traceability and acceptance of the summary showed no clear overall difference, while correction opportunities and the feeling of being heard also leaned towards the human-moderated workflow (Figure 5). The ratings describe perceived influence and experience; decision authority was not measured directly.

The separate final questionnaire asked participants to compare the formats directly. For output accuracy, 51.4% favoured the AI-supported workflow and 37.5% favoured the human-moderated workflow; for handling disagreement and less common views, the shares were 50.0% and 31.4%, respectively. Feeling heard favoured the human-moderated workflow: 49.3% preferred it on this criterion, compared with 27.4% preferring the AI-supported workflow. Only feeling heard showed a clear preference after Holm adjustment across the four direct comparisons (Holm $p = 0.015$). The other

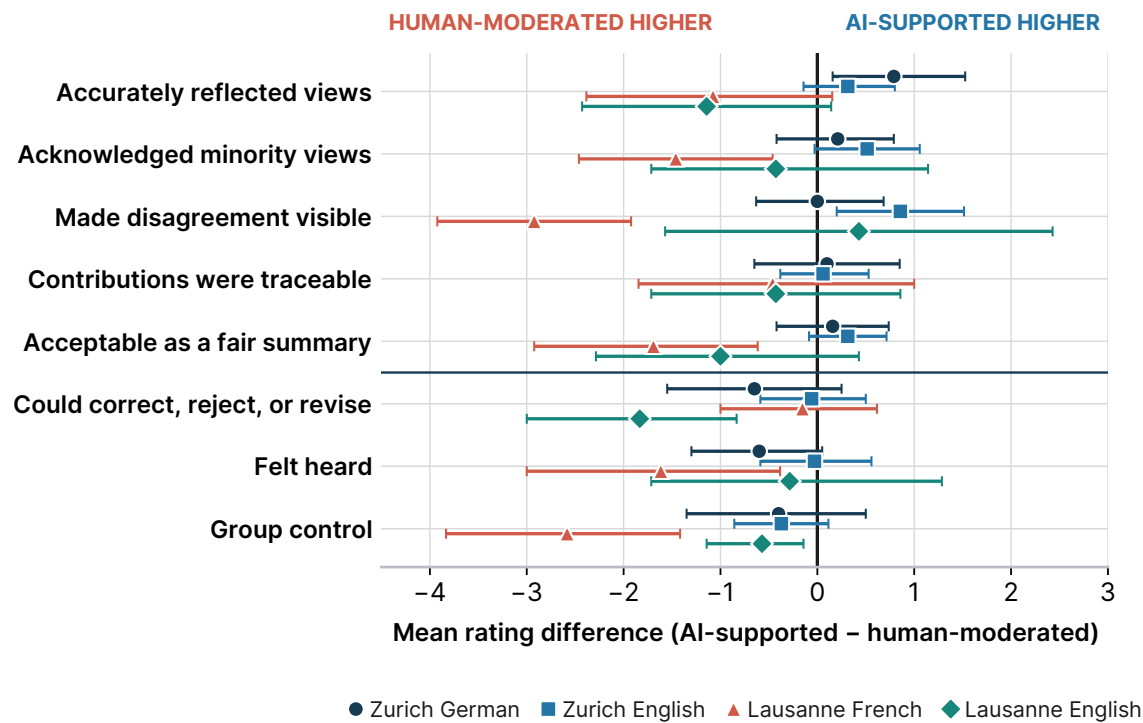


Fig. 5. Room-level paired differences (ratings for the AI-supported workflow minus those for the human-moderated workflow, in scale points). Original ratings range from 1 to 7; paired differences can range from -6 to +6. Points are mean paired differences and lines are 95% participant-bootstrap intervals. Colour and marker shape identify rooms. Positive values favour the AI-supported workflow. Negative values favour the human-moderated workflow. The first five items concern the summary and the last three concern participation.

comparisons describe tendencies in these workshops. Table 3 reports the means, response counts and shares rating the formats about the same, alongside contrasting participant comments.

The written responses help explain what participants valued about each format. Participants who favoured the AI-supported workflow described broader capture, easier expression and efficiency. Those who favoured the human-moderated workflow valued nuance, clarification and discussing together. P6 gave a mixed assessment, criticising accuracy while valuing summarisation and tension extraction (Lausanne English, paraphrased). Table 3 retains both sides of that response.

Fifty-five comments could be linked to a future-workshop preference. Appendices B.1, B.2 and B.4 present comments by theme, post-discussion diagnostics and further rating comparisons.

Table 3. Final comparisons and illustrative participant responses. Each mean summarises a direct comparison after both discussions. For future workshops, the scale expresses method preference. Only ratings for feeling heard differed clearly from the scale midpoint after adjustment across the four comparisons (Holm $p = 0.015$). Other apparent advantages are descriptive.

Criterion and mean 1 = human much better 4 = about the same 7 = AI much better	Case for the AI-supported workflow	Case for the human-moderated workflow
Output accuracy Mean: 4.31 Towards AI $n = 72$ 11.1% about the same	51.4% favoured the AI-supported workflow “I preferred the AI-moderated output which captured more statements.” (P24) “Keeping track of all views. Visibility. Summary.” (P36)	37.5% favoured the human-moderated workflow “AI was inaccurate and dropped important points.” (P6) “AI support allowed us to discuss more, but I do not feel it grasped all the nuances in the discussion.” (P1)
Feeling heard Mean: 3.34 Towards human $n = 73$ 23.3% about the same	27.4% favoured the AI-supported workflow “Everyone could take a position on every statement. Opportunities to state my own position were more limited in the human-moderated method.” (P69) “With AI, people can also give their point of view anonymously.” (P22)	49.3% favoured the human-moderated workflow “Without AI, I was clearly more able to express myself and felt heard.” (P10) “AI-supported mode seems efficient, but face-to-face communication with human moderation lets people give more authentic responses.” (P43)
Handling disagreement and less common views Mean: 4.24 Towards AI $n = 70$ 18.6% about the same	50.0% favoured the AI-supported workflow “People tried to explain their point, and we could have voted on it. That was easier than confronting them when I disagreed.” (P2) “AI is better at summarisation and automated tension extraction.” (P6)	31.4% favoured the human-moderated workflow “With human moderation, we address the group; the moderator understands subtleties and asks for clarification.” (P15) “With AI there was little disagreement and no nuance; the result did not take these nuances into account.” (P10)
Preference for a future workshop Mean: 3.71 Towards human $n = 72$ 26.4% about the same	33.3% favoured the AI-supported workflow “The app was much more efficient, also in terms of its analysis.” (P72) “AI support records the voices, so there is no need for handwritten notes, and the model is good enough to summarise and classify subjects.” (P23)	40.3% favoured the human-moderated workflow “With human moderation, there were real exchanges and real listening between participants.” (P16) “Murmi allowed less explanation and discussion than the blackboard method. The blackboard felt more like joint work.” (P74)

Shares include all valid responses in each row. Ratings 1–3 favour the human-moderated workflow, 4 means about the same and 5–7 favour the AI-supported workflow. Each excerpt comes from a respondent who favoured that method on that criterion. These 16 excerpts from 14 respondents were translated, where necessary, and edited for readability. The general open question asked about method differences without asking for an explanation of each rating. P74’s excerpt retains the original platform name, Murmi. Anonymity is P22’s perception.

4.1.2 Voting and typing offered additional ways to contribute. P2 described voting as easier than confronting a speaker when disagreeing (Lausanne English). P69 valued being able to respond to every statement and described more limited opportunities to express a position in the human-moderated workflow (Zurich German, translated; see Table 3). These responses describe ways to register a view beyond taking a speaking turn.

Typing also helped some participants contribute. P29 described what helped as *“to be able to write, to be heard even if shy”* (Zurich English), while also reporting distraction and less listening. P70 valued posting an opinion immediately (Zurich German, paraphrased). Participants also valued recording: P45 described live transcription and the creation of arguments from participants’ ideas as helpful, and P46 said real-time capture kept thoughts from being forgotten (Zurich English, paraphrased). These are perceived benefits of contributing and recording. They do not establish complete capture or equal influence.

4.1.3 Participants valued the AI-generated outputs but wanted more discussion. Among the 57 written comparisons, 32 were coded for a human discussion or contextual-detail advantage, 23 for accuracy or detail concerns, and 20 for AI synthesis, structure or visibility. Five of those 20 answers also carried the human discussion or contextual-detail code. Some respondents therefore described both useful synthesis and an advantage of human discussion in the same answer. P24’s comparison illustrates how broader capture could coexist with less discussion:

“Less discussion + critical reflections came in the bigger AI moderated discussion. I preferred the AI-moderated output which captured more statements.” (P24, Zurich English, verbatim excerpt)

P24 favoured the human-moderated workflow for a future workshop (3/7), despite preferring the AI-supported output. Their comment describes less discussion in the AI-supported round, although both workflows began in small groups before sharing with the room. We interpret the comment as an account of how much discussion P24 experienced.

Several participants described the difficulty of contributing through the app while following the discussion:

“difficult to assess or vote and contribute to the debate at the same time.” (P1, Lausanne English)

P1 suggested a quiet review period; P65 reported distraction while writing and found Post-its more collaborative (Zurich German, paraphrased). Of 53 platform answers, 25 concerned synthesis or visualisation, 14 concerned divided attention or reduced exchange, and eight concerned insufficient time or pacing. Nine of the 25 synthesis answers also concerned divided attention or reduced exchange. These codes show which concerns appeared in the same answers. Cognitive overload was not measured.

Later cards generally received fewer votes and also left less time to respond before reporting (Appendix B.3, Figure 9). These voting patterns do not establish whether participants were listening at the same time.

The Zurich German comments and the workshop lead’s recollection raised a possible connection between the opening activity and participants’ attention. The workshop lead felt that asking people to enter statements first drew attention towards completing entries in the app and away from talking together. An early transcript passage notes that participants had already entered many points and invited a short spoken round. P58 similarly described less of a group result when everyone wrote their own opinion (paraphrased from German). These descriptions suggest competition between activities, while the transcript shows that spoken discussion continued. However, they do not isolate the effect of the opening instructions.

Participants also described what they wanted from the exchange. P48 wanted the tool embedded in social interaction and wrote, *“I would have liked to hear from the others more, maybe switch tables etc.”* (Zurich English, verbatim excerpt). Read alongside P24’s concern about critical reflection and P58’s description of individual writing, this suggests that opportunities to contribute did not always provide the shared discussion these participants wanted.

The workshop lead and one participant described different experiences of documentation and discussion. The workshop lead described less manual summary writing and more attention available for guiding the discussion, with moderator numbers unchanged between formats. P26 valued report generation while noting missed speech:

“the arguments are more individual, less collaborative. But very efficient for workshop host!” (P26, Zurich English)

P26 selected the midpoint for future-method preference. They saw the workflow as efficient for the host while expressing concerns about how it affected shared discussion. Neither their comment nor the workshop lead’s reflection measured effort saved.

P20’s and P67’s comments expressed contrasting expectations of deliberation and consensus (Table 6).

Participants’ assessments distinguished the physical board or digital summary from their experience of contributing to it (RQ 1). Some valued broader capture and the additional ways to express a position through voting or typing, while wanting more exchange with others. The ratings also showed lower perceived group control after the AI-supported workflow, and correction opportunities leaned towards the human-moderated workflow. Valuing the resulting board or summary did not necessarily mean feeling heard or able to influence it. These assessments concern participants’ experiences of the workshop arrangements; they do not establish whether their views changed.

4.2 Participants challenged meaning as discussion became cards and summaries

Voting continued after the first summary was generated in all four rooms, and three rooms produced a second version before reporting (Figure 6). In Zurich English and Lausanne French, additional propositions also entered between versions. Participants were still contributing and responding while the summary developed.

During this overlap, participants and moderators questioned wording, added statements and challenged what summaries included, emphasised or left out. They discussed both the meaning of individual cards and what the combined summary conveyed about the group. The following episodes trace what participants asked to preserve, what they and moderators did, and what appeared in the later record. Appendix B.2 presents participants’ concerns about selection in the human-moderated workflow.

4.2.1 Participants and moderators worked through corrections together. When participants questioned how a contribution had been captured, correcting the card became another task to work through during the discussion. In Lausanne English, an early exchange concerned whether the state had sufficient AI expertise to govern it. The machine transcript records a request to make a statement specifically about the state’s fluency, discussion of deleting the wrong statement and an objection that the text was almost the reverse of what had been said. It also records an instruction to let the app’s transcription run in the background and continue the discussion. In one example recalled by the workshop lead, the moderator revised a card about the government’s role after a participant explained that its wording did not capture their intended meaning. The lead also recalled a participant entering another card. The correction illustrates how a participant could shape the wording presented for voting through an exchange with the moderator.

An approved card from this interval described the state’s need for capability to govern AI. It was marked edited, received seven responses and was cited by two recommendations in the final summary. Its earlier wording and edit time were not retained, so the stored record cannot establish which action produced this version or whether the participant accepted the revised wording. Appendix B.3 reports the wider edit counts and capture problems.

Direct entry offered another way to contribute, but one participant in Zurich German described uncertainty about how to use it:

“I find the report good and interesting, but I do not understand how it came about, so I am a little mistrustful. No one explained to me how to create an ‘opinion card’ that the whole group could then ‘vote’ on.” (P67, translated)

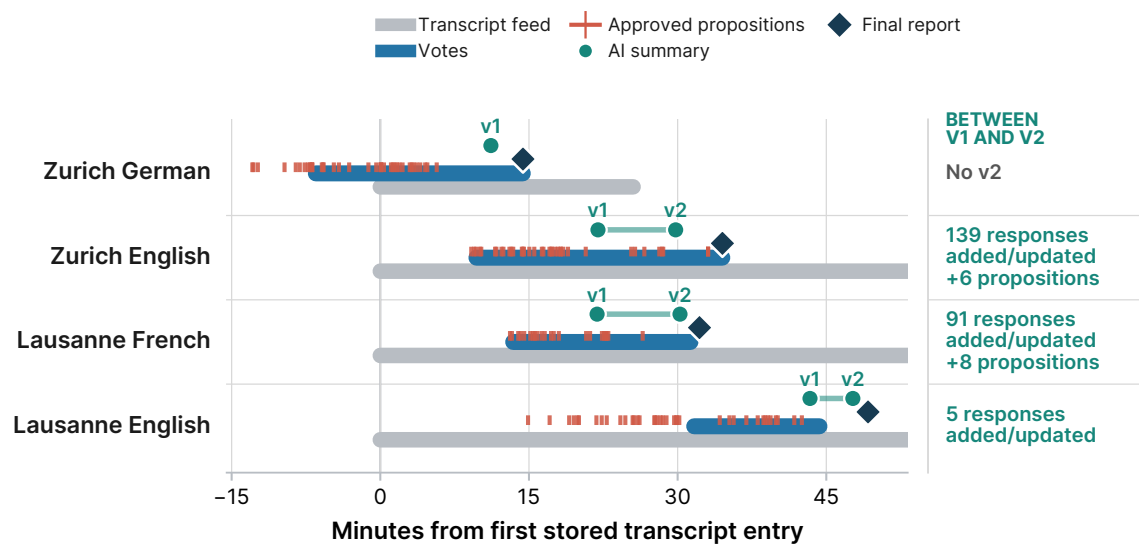


Fig. 6. Contributions and responses continued as summaries developed. Each row shows one room: grey and blue bars span recorded transcription and voting activity; coral ticks mark creation of propositions eventually approved for voting; teal circles mark summary generation; navy diamonds mark final reports. The right-hand column counts show responses added or updated, and additional propositions, between summary versions 1 and 2. These are changes visible in incomplete records, not counts of new voters. Zurich German completed the review with version 1.

P67 also selected 6 on the final future-preference scale, favouring the AI-supported workflow. This respondent's favourable preference accompanied uncertainty about the process; it does not establish how other participants understood the platform.

4.2.2 Tracing concerns through cards and successive summaries. The following cases show what was preserved and what changed as discussion became voting cards and summaries. We matched spoken comments to cards using wording and timing because the records did not directly link them.

Lausanne English: a workload concern retained across summaries. In Lausanne English, the discussion moved from preparing workers to use AI to protecting them from growing work pressures. Someone described “a real risk of mental overload and burnout” and added, “I start to feel it myself.” Reducing working hours was then suggested as a possible response to more intense work. These ideas became separate voting cards about overload, work intensity and shorter hours. The votes supported the two descriptions of the problem more widely than the proposal to reduce hours, which also received neutral responses.

Both summary versions carried the issue forward as an open question: “How can we assess the impact of AI integration on worker mental health and workload?” Neither presented shorter hours as an agreed recommendation. Here, the shared summary retained the concern while preserving a distinction between recognising a problem and agreeing on what to do about it. Table 13 provides the detailed trace.

Lausanne French: what the shorter statement left out. When the first Lausanne French summary described broad agreement on integrating AI learning, someone objected: “We have resolved absolutely nothing.” Later, when the group was asked what was missing, someone said school must reinvent itself but questioned “where” and “how”, and

said the summary left out the emotion in the discussion. A voting card created in the same minute read “School must reinvent itself”. It kept the call for change but left out the uncertainty about what that change should mean. These excerpts are translated from French.

The ten votes recorded for this statement were all in favour, yet written feedback continued to question the summary: “Same comment as before. Too much flattening of arguments and ideas.” Most of the summary’s wording stayed the same. People had another chance to explain what was missing and vote, but agreement with a short statement did not show that the summary captured what mattered to them. We cannot tell whether the repeated criticism referred to this particular card. Appendix D.2 gives the full sequence, including the timing and limits of the records.

The records do not establish whether these exchanges explain changes in approval. Appendix C.3 provides a further example from Zurich English.

Participants and moderators worked with the cards and summaries by clarifying wording, adding contributions and discussing what a card or summary missed (RQ 2). Their questions concerned the meaning of individual contributions and whether the summary preserved uncertainty and unresolved differences. The English summary kept a workload concern open without settling on a solution; the French card kept a call for change while leaving out uncertainty about its direction. The recalled moderator correction shows one way a participant influenced a card, while the continuing French objections show that further discussion and voting did not necessarily settle concerns about the summary.

We also checked the propositions cited by the summaries. In Zurich English, eight second-version summary items cited the same proposition about workforce replacement, including items about training and health-care ethics. Even when a link opened an existing proposition, that proposition could be unrelated to the summary claim. Appendix C.3 gives the source-link audit and a contrasting example with relevant references.

4.3 Supporting statements did not necessarily mean accepting the summary

Participants could mark a summary “good enough” or “not good enough” and explain their response. Accepting a group summary could differ from endorsing its recommendations. Two reviews of the second Lausanne English summary illustrate the distinction:

“I disagree with the recommendations, but it’s OK with what the group thinks” (Accepted)

“Missing texture, details and color. Some issues dropped. Some misunderstood.” (Rejected)

The first response distinguishes personal agreement from accepting a description of the group’s views. The second questions what that description preserves. Neither response can be linked to a particular edit. Appendix C.1 retains all written review reasons and Appendix C.2 examines agreement with cited propositions alongside summary-review responses.

Linking votes with summary-review responses also showed support for propositions alongside rejection of a summary. Of the 17 responses rejecting a first summary, eleven were linked to votes supporting at least three quarters of the cited propositions answered. The calculation uses the latest retained responses; votes at the time of review could not be reconstructed. It also depends on which propositions were cited, and those links could be incomplete or unrelated to the summary claims. It does not explain the reasons for each rejection (Appendix C.2).

The five written reasons accompanying first-version rejections concerned missing content, lost nuance, and summaries that made the discussion seem more settled than it was. A separate survey response from P5 described a perceived mismatch in emphasis:

“We had a long discussion about if the government was able to take on that role and it’s appearing as a detail in the report (hierarchy is missing).” (P5, Lausanne English)

In Lausanne French’s discussion of AI in education, P16 and P17 likewise reported that the role or purpose of schooling remained unclear or underrepresented in the final output. These questionnaire responses concern the same topic as the review episodes, but the records do not identify these respondents as the speakers or reviewers in those sequences or connect their answers to a particular edit.

Concerns about emphasis also extended across formats. P70 felt that human presentations conveyed what particularly mattered to each presenter, while the AI result appeared smoothed and optimised (Zurich German, paraphrased). P17 reported that ideas had not been prioritised in the human-moderated workflow either (Lausanne French, paraphrased). These comments distinguish including an issue from preserving its perceived importance; they do not establish an agreed weighting of the discussion’s concerns.

Review responses across summary versions. Three rooms reviewed a second summary before reporting (Figure 7). Lausanne French’s approval fell from 75% (9 of 12 review responses) to 69% (9 of 13); nine responses approved each version, but the responses did not all come from the same accounts or devices. Of ten reviews paired across versions using anonymous identifiers, one changed from approval to rejection and none changed towards approval. Review ended below the 80% target because the 45-minute time limit left no time for another iteration. Lausanne English rose from 40% (2/5) to 86% (6/7). Of its five paired reviews, two changed to approval and none changed away. Approval also rose in Zurich English, while Zurich German completed review at 80% with the first version. These snapshots combine changes in responses with differences in who submitted a review. They do not isolate effects of particular revisions. Appendix C.1 reports all linked judgements.

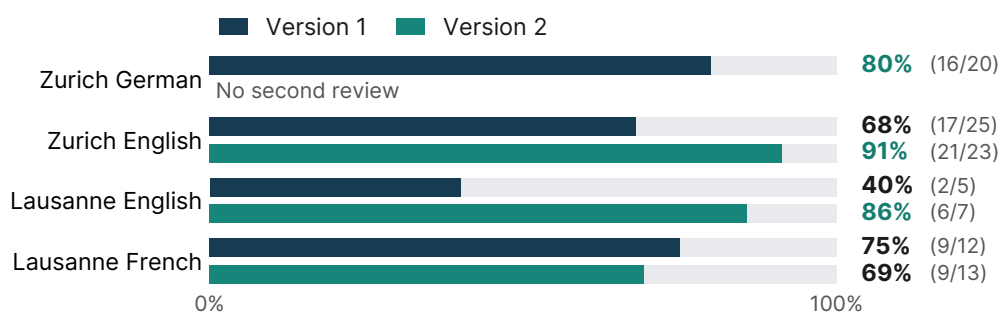


Fig. 7. Summary of approval by version. Coloured segments show approval. Light grey completes each bar to 100%. Teal percentage labels mark approval at or above 80%, the practical target for summary acceptance. Labels give percentages followed by approvals/summary-review responses. Responses were not necessarily submitted by the same accounts or devices across versions. These snapshots differ from the final-report snapshots in Appendix A.1: Lausanne English’s second version has 6/7 approvals here, compared to 5/6 in its final report. Zurich German completed review with version 1. Later stored versions were generated after its report.

The two complete Zurich English summaries are compared in Appendix C.3.

In reviewing summaries, participants judged how the group was represented as well as whether they agreed with particular statements. Written reasons questioned missing concerns, their relative emphasis and the impression that unresolved issues had been settled. Acceptance could accompany disagreement with recommendations, while support

for cited propositions could accompany rejection of the summary. Overall approval and changes between versions do not establish that a particular concern was resolved.

5 DISCUSSION

Participants valued AI support for some aspects of the workshops and wanted more shared discussion in others. Around half favoured the AI-supported workflow for output accuracy and handling disagreement or less common views. Participants more often favoured the human-moderated workflow for feeling heard. Future-workshop preferences were mixed: 33% favoured the AI-supported workflow, 40% the human-moderated workflow and 26% rated them alike (Table 3). These assessments concern the workshop arrangements as a whole and cannot be attributed to AI alone.

Cards and summaries sometimes became part of the discussion itself. A card could offer another way to express a position while losing a qualification that mattered to its speaker. A summary could be accepted as a fair representation of the group despite personal disagreement with its recommendations. Participants drew on their experience of the conversation to question these interpretations.

We examine this relationship through *dynamic participatory sensemaking*: discussion shapes cards and summaries, and participants' responses to them can prompt clarification, further contributions and revision. Participation depends on whether people can influence those interpretations, including how their views appear alongside others'. In the correction recalled from Lausanne English, a participant helped shape a card's wording. In Lausanne French, concerns persisted through further discussion and voting. These episodes suggest that participants need to recognise their meaning in the cards and summaries, have time to consider them together and be able to follow what happens to requests for change. Whether this process improves mutual understanding or agreement remains to be studied.

5.1 Preserving people's meaning before drawing views together

For someone who has just explained why an issue matters to them, seeing their words return on a voting card can invite a difficult question: is this still what I meant? It also gives others another way to engage with the idea. Some participants valued being able to register disagreement without confronting the speaker, and direct entry offered a way to express a view when it was difficult to get a speaking turn (Section 4.1). A card could give an idea more attention than it received in the spoken exchange, provided people could still recognise the meaning behind its wording.

The cases suggest that participants need different things from voting cards and discussion summaries. A voting card needs to preserve an individual's intended meaning; a discussion summary needs to represent the group's views. The Lausanne French card about schooling omitted the speaker's uncertainty about what change should mean (Section 4.2.2), while a reviewer in Lausanne English accepted a summary despite disagreeing with its recommendations (Section 4.3). Review needs to allow participants to question both a card's wording and the account of the group's discussion.

We therefore argue for keeping argument extraction close to what people say. A voting card should preserve the speaker's reasons, conditions and uncertainty, so that others respond to the view they intended to express. Where a short card cannot carry this context, an accompanying explanation could make it available before voting. The discussion summary can then draw views together, showing shared concerns alongside differences and unresolved questions. Giving each step a clearer purpose could help people first feel understood in their own terms, then consider how their views belong in a summary they share with others.

For that possibility to become part of everyday workshop practice, people need ways to explain what a card misses, attach a reason or bring a disputed statement back to the room. Voting can open these conversations, but a response

count cannot tell us whether someone has considered an idea or understood what its author meant [7, 34]. Our records likewise do not establish whether people became more open-minded, changed their views or understood one another better.

5.2 Dynamic participatory sensemaking redistributes work and capability

Dynamic participatory sensemaking brings some of the work of interpreting a discussion into the exchange between participants and moderators. Choices about selection and emphasis that researchers would make during later reporting became open to participants while they were still in the room. Participants could explain what a card missed and question how a summary brought their views together. Researchers retained responsibility for interpreting the material and preparing the findings. Supporting this process therefore requires attention to who can question an interpretation, who helps work through it and who carries a requested change forward.

This work extended across discussion, voting cards and summaries. In Lausanne English, participants and moderators clarified an intended contribution and coordinated a card correction. Its wording mattered to the speaker, to people voting on it and to those preparing the summary. The coordination involved in opening reporting to participant review is a form of articulation work [28, 31]. Missing edit history limits how fully we can reconstruct its outcome.

Review also competed for attention: 14 of 53 platform comments mentioned divided attention, and eight mentioned insufficient time. The workshop lead described less manual writing and more attention for facilitation, with unchanged staffing and no measured time saving. Participants checked meaning, moderators coordinated review and researchers interpreted incomplete records. This effort deserves recognition alongside the finished report [28, 31].

These experiences suggest testing a different workshop rhythm: discussion, time to read and correct cards, voting, and a return to conversation before summary review. Participants could focus on each activity and return to earlier contributions when needed. Future work should compare this phased format with overlapping discussion and voting under comparable conditions, examining whether participants feel heard, can follow through on corrections and prefer the experience.

These opportunities need to be accessible beyond those already active in the app. Identifiers with a first-summary review response had voted on a larger share of the available cards before that summary was generated. P67 favoured the AI-supported workflow but said no one had explained how to create a card. A short practice activity, spoken correction requests and moderator invitations to less-active participants could help people take part. Future studies should examine who needs help, who raises corrections and whose concerns reach later versions. Our study did not establish whether particular demographic groups were disadvantaged. These proposals need evaluation alongside the phased format.

Listening, clarifying misunderstandings and working through disagreements still call for human skill. Supporting dynamic participatory sensemaking therefore means making room for this work alongside voting and summary production.

5.3 From richer records to accountable public reporting

Reading transcripts, cards, votes and summaries together helped us see how contributions were represented and where concerns remained unresolved. Readers encountering only the final report would have less access to these exchanges. We did not study how policymakers or other readers used the reports. Future work should examine whether showing the sources behind recommendations, alongside participant feedback and revision histories, helps readers understand

how a report was produced and what remained unresolved. This would also require deciding which material participants are comfortable sharing beyond the workshop.

Missing source passages and revision histories made it difficult to establish what supported a summary claim. Final reports could also combine a reviewed summary with later changes to propositions or responses. A polished summary backed by counts may appear more authoritative than these records warrant. Readers need access to the relevant sources and the exact version participants reviewed. Accepting a proposition, accepting a summary and authorising a public report remain different judgements [25].

Participants should also be able to follow what happened to a concern they raised. Summary claims could link to source cards and responses, with objections linked to revisions or marked unresolved. After an edit, the person who raised the concern should be asked whether the new wording does it justice. Reports should distinguish how many people reviewed a summary from how many approved it, and show whether later changes were reviewed. Such a record could also give citizens a basis for asking how public institutions responded to their recommendations.

These opportunities to question the record still relied on how the moderator responded. Selection and wording could steer what became available for a vote, even when the voting itself was transparent. Participants could follow edits on screen, but did not have a complete record of what had been changed or left out. In wider use, participants should be able to review these decisions and raise concerns with someone other than the moderator. The software would also need to prevent duplicate accounts and moderator voting, and limit spam, without making participation unnecessarily difficult. Our study did not establish that moderator review prevented bias or deliberate manipulation.

Making the record easier to inspect also needs to respect the privacy commitments described in Section 3.3. A source passage that explains a summary may contain context that identifies its speaker. Participant access to evidence within a workshop and public access to a report therefore need separate decisions about what can be shown and to whom. Future designs should evaluate whether restricted source access and approved excerpts provide enough context to question a claim while limiting exposure. Permission to publish should concern the particular report and supporting material to be released; accepting a summary during a workshop does not settle that decision.

5.4 Extending deliberation across places, times and languages

A shared record could help people continue a discussion remotely, return later with further thoughts or encounter contributions from other groups in their preferred language. They could revisit one another's reasons, add what was missing and carry a face-to-face conversation beyond the meeting. Our workshops used German, English and French in separate groups and did not test these forms of exchange. Participants would need ways to check translated contributions and correct wording that no longer conveyed their meaning. Remote, asynchronous and cross-language participation remain directions for future study.

A parliamentary debate offers one example of what this could look like. In a multilingual setting such as Switzerland, citizens could follow contributions in their preferred language, respond to statements drawn from the debate and add reasons or questions from elsewhere. Some could participate as the debate unfolds; others could return to the material later. Representatives could consider these contributions during or after the debate. Such a process would need ways to check meaning across translations and make clear how citizens' contributions enter the exchange.

Following discussions over time could show whether participants can get cards and summaries revised, respond to one another's reasons and keep unresolved concerns in view. The aim is for deliberation to continue beyond one meeting, while keeping later recommendations connected to the concerns, disagreements and hopes people expressed.

6 LIMITATIONS

The study had to fit the practical needs of four workshops held as part of a public consultation. Citizens first took part in a human-moderated workflow to become familiar with the workshop before taking part in the AI-supported workflow. Participants' experience in the first round may have influenced how they approached the second, making the workflows difficult to compare independently. Topics, languages, group sizes and facilitation also varied across rooms, so the findings reflect how these elements worked together in this setting. Demographic information was available for only some participants, which further limits our understanding of how their backgrounds may have shaped their experiences.

To understand how participants' concerns reached the summaries, we read the transcripts alongside cards, responses and successive versions of the text. This made some changes visible, but we had no retained audio to return to, and source links and parts of the edit history were missing. Connections based on wording and timing therefore leave some uncertainty about how a change came about. The records also do not establish whether participants felt that the revision addressed their concern.

Participants' ratings and comments help explain what they valued and what made the workshops difficult, though they offer less insight into how their views developed or how well they came to understand one another. Since comments were optional and sometimes translated, the responses presented here also depend on who chose to respond and how we interpreted their words. The workshop lead's later reflections helped us understand the demands of facilitation, but further work would be needed to establish how the platform changed the effort required of participants, moderators and researchers.

Working closely with the platform as its designers and facilitators gave us insight into its use, while also making our own role part of the setting we studied. Citizens and facilitators may find different ways of working with it as they gain experience or use it for other purposes. Following those adaptations over time, along with what happens to reports after a workshop, would help establish how AI support fits into a deliberative process that citizens themselves can shape.

7 CONCLUSION

We developed Murmi to explore dynamic participatory sensemaking through live discussion, voting and summary review. Across four language-group workshops at two Swiss events, questionnaire responses from 76 participants more often favoured the AI-supported workflow for output accuracy, while the human-moderated workflow was more often favoured for feeling heard. These assessments reflect the workshop arrangements as a whole and cannot be attributed to AI alone.

Participants valued additional ways to express their views, but contributing through voting and typing did not always provide the exchange they wanted with others. Their questions about cards and summaries also showed why representation belongs within that exchange: wording could retain a topic while losing its uncertainty, and support for individual statements did not necessarily mean accepting how the discussion was summarised. Reviewing these materials gave participants opportunities to explain what they meant and question how contributions were brought together, although the records did not consistently establish whether their concerns were resolved.

In these workshops, dynamic participatory sensemaking involved participants returning to cards and summaries to consider how their views were represented. Supporting it requires time to consider contributions together, ways to challenge wording and emphasis, and follow-through on requests for revision. Participants and moderators help

determine what the discussion means and what remains unsettled. Designs should give participants a continuing say in how their contributions become shared summaries and public recommendations.

ETHICAL CONSIDERATIONS AND ADVERSE IMPACTS

The ETH Zurich Ethics Commission approved the protocol (project 26 ETHICS-116) and amendment before fieldwork. Section 3.3 describes consent, data protection and the opportunities participants had to challenge generated cards and summaries. These safeguards leave questions that matter beyond the workshops: who can influence what is included, whether an objection receives a meaningful response, and what participants have actually authorised when they accept a summary. We do not treat support for propositions or acceptance of a summary as evidence of consensus or permission to publish a report on the group's behalf. Wider use would also require access controls that prevent duplicate or unauthorised participation without unnecessarily collecting personal information or linking names to votes. For example, organisers could issue each attendee a unique random access code, with controlled replacement, to limit duplicate entries without storing names alongside votes. These controls and their effectiveness against deliberate manipulation remain to be tested.

GENERATIVE AI DISCLOSURE

We used OpenAI Codex to assist with programming the deliberation platform, writing SVG code for the workflow illustration in Figure 1, and formatting LaTeX tables. Generative AI was not used to interpret the study data. The authors take responsibility for the final manuscript and research artefacts.

AUTHOR CONTRIBUTIONS

Joshua C. Yang, Maud Reveilhac and Aurelia Tamò-Larrieux jointly conceived and designed the study, contributed to data collection and analysis, and drafted the manuscript. Joshua led the experiment's design and implementation, the platform's design and development, and the writing and final preparation of the manuscript, and serves as the corresponding author. Maud initiated and led the broader TA-Swiss project, developed the overall research plan, contributed to securing funding, coordinated the citizen workshops from planning through delivery, contributed substantially to the experimental design and implementation, and revised the manuscript. Aurelia co-led the TA-Swiss project and contributed intellectual and practical insights that shaped the study's direction, interpretation and writing. Both Maud and Aurelia played central roles in conducting the workshops in Zurich and Lausanne. Dirk Helbing provided intellectual guidance on digital democracy, contributed ideas that shaped the broader research direction and study design, and reviewed and edited the manuscript. Kenneth Holstein contributed to the paper's conceptual framing and positioning within HCI, helped develop the analytical framing and interpretation of the findings, and refined the articulation of the contributions and implications. He provided substantive feedback on the argument, structure and presentation, and contributed to reviewing and revising the manuscript.

ACKNOWLEDGMENTS

This study was conducted as part of a TA-SWISS-funded project. Joshua C. Yang, Maud Reveilhac and Aurelia Tamò-Larrieux were members of the project team.

We thank Carbon Copy, a Zurich-based nonprofit organisation, especially Philip Widmer and Christian Horvat, for substantial technical and logistical support, including contributions to the development of the Murmi platform and

assistance with collecting the pre-workshop questionnaires. We also thank Maurice Flechtner for developing DRI questionnaire items used in the experiment, and Fynn Bachmann for early discussions that helped shape the idea behind the platform and its subsequent development. Gerold Schneider and Simon Mayer provided intellectual and project support. Vlada Druta, Nisha Yadav and Paula Deutschmann helped collect and organise the study data. Finally, we are grateful to all the citizens who took part in the workshops and contributed their time, experiences and perspectives.

REFERENCES

- [1] Lucas Anastasiou and Anna De Liddo. 2026. From Shallow Discourse to Structured Deliberation: The BCause Approach. In *Human-Computer Interaction – INTERACT 2025 (Lecture Notes in Computer Science, Vol. 16111)*, Carmelo Ardito, Simone Diniz Junqueira Barbosa, Tayana Conte, André Freire, Isabela Gasparini, Philippe Palanque, and Raquel Prates (Eds.). Springer Nature Switzerland, Cham, 94–99. https://doi.org/10.1007/978-3-032-05008-3_19
- [2] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale. *Proceedings of the National Academy of Sciences* 120, 41, Article e2311627120 (2023). <https://doi.org/10.1073/pnas.2311627120>
- [3] André Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark E. Warren. 2018. Deliberative Democracy: An Introduction. In *The Oxford Handbook of Deliberative Democracy*, André Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark E. Warren (Eds.). Oxford University Press, 1–32. <https://doi.org/10.1093/oxfordhb/9780198747369.013.50>
- [4] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. 2022. Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 38176–38189. <https://doi.org/10.52202/068431-2766>
- [5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [6] John Boswell. 2021. Seeing Like a Citizen: How Being a Participant in a Citizens’ Assembly Changed Everything I Thought I Knew about Deliberative Minipublics. *Journal of Deliberative Democracy* 17, 2 (2021), 1–12. <https://doi.org/10.16997/jdd.975>
- [7] Simone Chambers and Mark E. Warren. 2025. Why Deliberation and Voting Belong Together. *Res Publica* 31, 2 (2025), 279–297. <https://doi.org/10.1007/s11158-023-09635-x>
- [8] Xinyue Chen, Nathan Yap, Xinyi Lu, Aylin Gunal, and Xu Wang. 2025. MeetMap: Real-Time Collaborative Dialogue Mapping with LLMs in Online Meetings. *Proceedings of the ACM on Human-Computer Interaction* 9, 2, Article CSCW132 (2025), 35 pages. <https://doi.org/10.1145/3711030>
- [9] Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3 (1951), 297–334. <https://doi.org/10.1007/BF02310555>
- [10] Taufiq Daryanto, Xiaohan Ding, Kaike Ping, Lance T. Wilhelm, Yan Chen, Chris Brown, and Eugenia H. Rho. 2026. Human-Human-AI Triadic Programming: Uncovering the Role of AI Agent and the Value of Human Partner in Collaborative Learning. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 541, 21 pages. <https://doi.org/10.1145/3772318.3791773>
- [11] Hanne De Jaegher and Ezequiel Di Paolo. 2007. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences* 6, 4 (2007), 485–507. <https://doi.org/10.1007/s11097-007-9076-9>
- [12] Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap* (1st ed.). Chapman & Hall/CRC. <https://doi.org/10.1201/9780429246593>
- [13] James S. Fishkin. 2018. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press. <https://doi.org/10.1093/oso/9780198820291.001.0001>
- [14] Umberto Grandi, Robert Bredereck, Théo Delemazure, Ulle Endriss, Jan Maly, Nick Mattei, Nicolas Maudet, Oliviero Nardi, and Stanisław Szufa. 2026. Social Choice with Text: Collective Decision Making in the LLM Era. Position paper presented at the ACM AI Leadership Summit. <https://hal.science/hal-05548117v3> HAL preprint hal-05548117, version 3.
- [15] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100. <https://doi.org/10.1093/jcmc/zmz022>
- [16] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. <https://www.jstor.org/stable/4615733>
- [17] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghui Cheng. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1, Article 152 (2023), 20 pages. <https://doi.org/10.1145/3579628>
- [18] Maija Jäske, Katariina Kulha, Mikko Leino, Maija Setälä, and Toni Wessman. 2026. Can AI Support Inquiry, Co-Creation, and Integration Functions in Citizens’ Assemblies? *Politics* (2026). <https://doi.org/10.1177/02633957261466242> Advance online publication.

- [19] Andrew Konya, Luke Thorburn, Wasim Almasri, Oded Adomi Leshem, Ariel Procaccia, Lisa Schirch, and Michiel Bakker. 2025. Using Collective Dialogues and AI to Find Common Ground between Israeli and Palestinian Peacebuilders. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 312–333. <https://doi.org/10.1145/3715275.3732022>
- [20] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting Reflective Public Thought with ConsiderIt. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, New York, NY, USA, 265–274. <https://doi.org/10.1145/2145204.2145249>
- [21] Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 805, 24 pages. <https://doi.org/10.1145/3706598.3713865>
- [22] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 181 (2019), 35 pages. <https://doi.org/10.1145/3359283>
- [23] Christian List and Philip Pettit. 2002. Aggregating Sets of Judgments: An Impossibility Result. *Economics and Philosophy* 18, 1 (2002), 89–110. <https://doi.org/10.1017/S0266267102001098>
- [24] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 2208–2220. <https://doi.org/10.1145/2998181.2998335>
- [25] Sammy McKinney. 2024. Integrating Artificial Intelligence into Citizens’ Assemblies: Benefits, Concerns and Future Pathways. *Journal of Deliberative Democracy* 20, 1 (2024), 1–12. <https://doi.org/10.16997/jdd.1556>
- [26] Aaron T. Parisi, Nithum Thain, Alden Hallak, Vivian Tsai, and Crystal Qian. 2026. Real-Time Group Dynamics with LLM Facilitation: Evidence from a Charity Allocation Task. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 25–59. <https://doi.org/10.1145/3805689.3812253>
- [27] Manon Revel and Théophile Pénigaud. 2026. AI-Enhanced Deliberative Democracy and the Future of the Collective Will. *Philosophy & Technology* 39, 2, Article 81 (2026). <https://doi.org/10.1007/s13347-026-01044-1>
- [28] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW Seriously: Supporting Articulation Work. *Computer Supported Cooperative Work (CSCW)* 1, 1–2 (1992), 7–40. <https://doi.org/10.1007/BF00752449>
- [29] Christopher T. Small, Michael Björkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *Recerca. Revista de Pensament i Anàlisi* 26, 2 (2021), 1–26. <https://doi.org/10.6035/recerca.5516>
- [30] Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. 2023. Opportunities and Risks of LLMs for Scalable Deliberation with Polis. <https://doi.org/10.48550/arXiv.2306.11932> arXiv:2306.11932 [cs.SI]
- [31] Susan Leigh Star and Anselm Strauss. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work (CSCW)* 8, 1–2 (1999), 9–30. <https://doi.org/10.1023/A:1008651105359>
- [32] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science* 386, 6719, Article eadq2852 (2024). <https://doi.org/10.1126/science.adq2852>
- [33] Sunny Tian, Amy X. Zhang, and David Karger. 2021. A System for Interleaving Discussion and Summarization in Online Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3, Article 241 (2021), 27 pages. <https://doi.org/10.1145/3432940>
- [34] Charlotte C. L. Wagenaar and Ank Michels. 2026. Hybrid Combinations of Deliberation and Voting: An Illustrated Typology and Initial Expectations. In *Connecting Deliberation and Voting in Democratic Innovations: Variations and Perspectives on Hybrid Designs*, Charlotte C. L. Wagenaar, Take Sipma, and Frank Hendriks (Eds.). Edward Elgar Publishing, 18–36. <https://doi.org/10.4337/9781035353323.00010>
- [35] Joshua C. Yang and Fynn Bachmann. 2025. Bridging Voting and Deliberation with Algorithms: Field Insights from vTaiwan and Kultur Komitee. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 3207–3223. <https://doi.org/10.1145/3715275.3732205>
- [36] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 2082–2096. <https://doi.org/10.1145/2998181.2998235>
- [37] Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024. Fair Abstractive Summarization of Diverse Perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 3404–3426. <https://doi.org/10.18653/v1/2024.naacl-long.187>

A DATA AND IMPLEMENTATION

A.1 App activity counts and analysis datasets

Table 4 gives the account or device counts used to analyse app activity. The same account or device could submit both votes and summary-review responses. Summary approval uses only the review responses, so missing responses are not treated as disagreement.

The table uses each room’s final-report snapshot. Version-specific reviews use different snapshots. Zurich German’s first summary had 16 approvals among 20 summary-review responses, compared with 14 among 17 in its final report. Lausanne English’s second summary had 6/7 approvals, compared with 5/6 in its final report. The Zurich English report illustrated in Figure 2(h) received 23 summary-review responses, of which 21 approved the summary (91%). Reviews were submitted from 23 of the 35 identifiers (66%).

Table 4. App activity in the final-report snapshots. The first two counts refer to accounts or devices, rather than individual votes. Each account or device could contribute one review response per summary. Approval is the percentage of review responses marked “good enough”.

Room	Accounts/ devices	With votes	Review responses	Approvals	Approval rate
Zurich German	36	26	17	14	82%
Zurich English	35	32	23	21	91%
Lausanne French	21	18	13	9	69%
Lausanne English	9	7	6	5	83%

The final aggregates contained 2,579 votes. The analysed event log contained 2,415 votes from 82 non-host account or device identifiers, compared with 83 identifiers with votes in the final aggregates. The report-map analysis required at least seven responses and retained 78 records. These datasets serve different analyses and should not be combined into a single participant count.

A.2 Platform configuration and retained data

Table 5 documents the processing and intervention points described in Section 3.4. Settings come from the code available before the workshops. Complete AI requests were not saved, so individual generations cannot be reconstructed exactly.

Table 5. Platform processing, human intervention and retained evidence.

Stage	Processing	Human intervention and retained evidence
Audio to text	gpt-4o-transcribe produced live text and a second pass replaced the completed turn, preserving English, German or Swiss German, and French without translation.	Participants saw live and final captions. Audio remained in memory for the final pass, but no project audio file was written. All speakers shared one recorder label.

Continued on next page

Table 5 continued

Stage	Processing	Human intervention and retained evidence
Text to turn	Captions under the same label were merged across gaps of at most 12 seconds. Two seconds without new text completed the unanalysed suffix. Fillers and common hallucination markers were filtered.	No speaker or phase label separated argument, instruction, review or report narration. Retained text had chronology but no source-audio link.
Turn to card	gpt-4o-mini, temperature 0.2, returned at most one standalone claim. Instructions requested the overall argument and room language, and excluded procedural, incomplete or duplicate material. Exact duplicates were blocked.	Moderators could edit, approve, hold, reject or add cards. Participants could enter up to 240 characters and directly edit their own submitted cards. Only moderators could edit audio-extracted cards; participants could request these changes aloud while viewing the projected interface. Rejection deleted the candidate. Source spans, confidence, phase, model version and edit history were not retained.
Card to responses	Approved cards offered five choices, from strongly disagree to strongly agree. The export distinguished missing and neutral responses.	Participants used swipes or buttons. Moderators did not vote. The incomplete response log is reconciled in Appendix A.1.
Responses to draft	gpt-4o-mini, temperature 0.5, received approved propositions, counts and coverage, group contrasts, jointly supported pairs, earlier versions and anonymous review reasons. It requested recommendations, conditions, trade-offs and unresolved questions with proposition references.	Moderators requested generation. Invalid references were removed. If none remained, storage attached the first valid proposition, which could be unrelated to the claim (Appendix C.3).
Draft to review	Participants were instructed to select “good enough” when satisfied with the summary and “not good enough” when they found it unacceptable, and could give a reason. Later generation received prior anonymous feedback.	Summary versions and reactions were retained, but reaction timestamps and a complete transformation history were unavailable. Review applied only to the selected version.

Audio configuration. One recorder browser captured mono audio with echo cancellation, noise suppression and automatic gain control, resampled it to 24-kHz PCM16 and streamed frames by WebSocket. Server voice-activity detection used a 0.5 threshold, 300-ms prefix padding and 500-ms silence. The second transcription pass used temperature 0.

Live grouping and summary input. A Pol.is-inspired pipeline [29] projected voting patterns into two dimensions with PCA. K-means compared two to four groups where possible and selected the highest silhouette score. Support rates excluded neutral responses, whereas coverage included them and used the app-record roster. Equation 1 gives the project-specific rule for suggesting combinations of propositions. The implementation called split support “conflict”. The thresholds supplied guidance to the model. They did not validate collective agreement or the generated summary.

Joint-support rule. Two propositions can each attract majority support from different coalitions, so approval of each separately does not establish collective backing of their combination [23]. Checking joint and split support gives the summary model evidence about where these coalitions overlap and which disagreements a combined recommendation

could obscure. Before suggesting that the model combine them, the platform required both substantial *joint support* (supporting both) and limited *split support* (supporting one while opposing the other). For propositions j and k , let B_{jk} and F_{jk} count these two patterns, respectively, and S_{jk} count all identifiers with non-neutral responses to both. A pair was flagged only if at least two records gave non-neutral responses to both, at least 50% supported both and at most 30% split their support:

$$\text{flag for combination}_{jk} = \begin{cases} 1, & \text{if } S_{jk} \geq 2, B_{jk} \geq 0.5S_{jk}, F_{jk} \leq 0.3S_{jk}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

For example, a pair with 60% joint support and 20% split support would qualify, whereas one with the same joint support but 40% split support would not. Records opposing both counted in S_{jk} , but neither in B_{jk} nor F_{jk} . Neutral or missing responses to either proposition were excluded. We chose these thresholds as project-specific heuristics. They had not been validated as criteria for collective agreement or tests of logical consistency. This rule guided the prompt without guaranteeing that the generated summary followed it or was acceptable to participants. Participant review assessed acceptance separately, consistent with proposals to return generated summaries to participants for checking [30].

A.3 Voting-pattern maps for the citizen workshop report

The maps in Figure 8 use the retained non-host responses to 136 approved propositions, keeping the latest response per record and proposition. The seven-response minimum excluded two records each in Zurich English and Lausanne French, leaving 78 records (Appendix A.1).

Agreement strengths were collapsed to +1, disagreement strengths to −1, and neutrality retained as 0. PCA used observed-mean imputation by proposition. Projection scores were scaled by the square root of proposition count divided by answered count. K-means selected among two to five groups where feasible by silhouette score. Each room was fitted separately. Topic labels are author-coded descriptions of distinguishing propositions.

Across 100 proposition-bootstrap replicates per room, median adjusted Rand agreement with the original partition was 0.14–0.44. The same group count recurred in 21–49% of replicates. These are exploratory partitions. The live platform used evolving responses and encoded missing and neutral responses alike, whereas this reanalysis used the retained matrix and observed-mean imputation. Its positions and letters cannot verify live assignments.

Figure 8 shows the retrospective grouping of retained responses. The first two principal components accounted for 25.0–55.6% of response variance. Coordinates and cluster labels are specific to each workshop. They cannot identify the groups or speakers in the live exchanges retained in Appendix D.

B PARTICIPATION AND ADDITIONAL SURVEY RESULTS

B.1 Selected participant responses

The selection below retains 37 responses from 31 respondents: 24 method comparisons, seven platform responses and six post-discussion diagnostics. Tables 6 and 7 contain 35 responses, and two human-output diagnostics follow the second table. They support the experiences described in Section 4.1 and the concerns about representation in Section 4.2. The selection covers all four rooms, different future preferences and positive, critical and mixed responses. It includes a report of no missing content. These are illustrative responses from 148 written answers, not estimates of prevalence. Some main-text excerpts reappear here with their wider context and attribution.

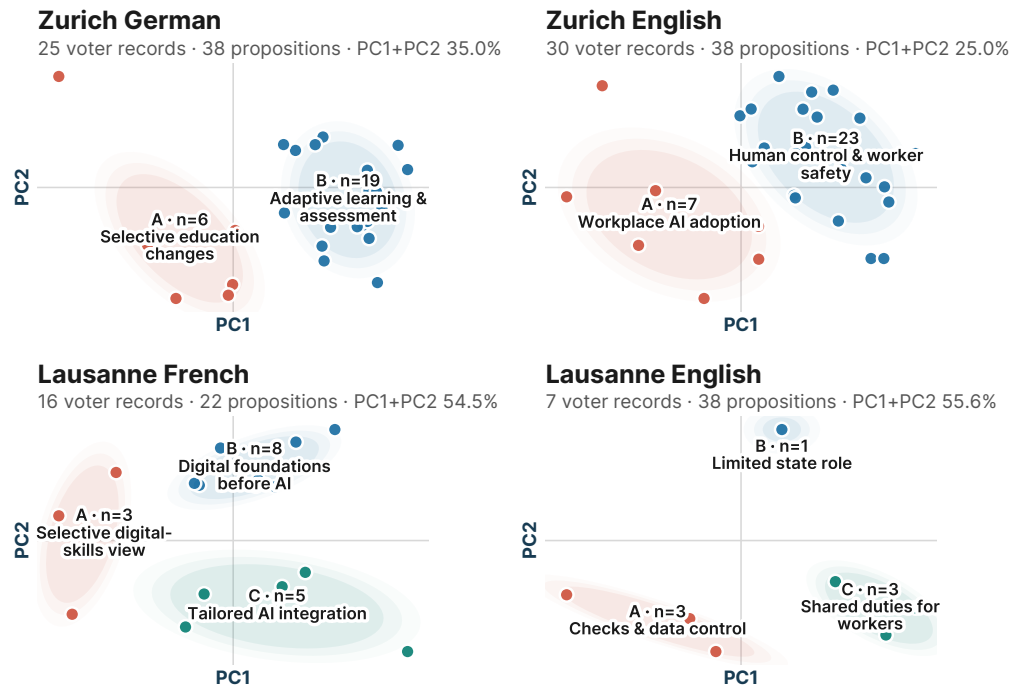


Fig. 8. Exploratory voting-pattern maps prepared for the citizen workshop report using PCA and k-means clustering. Each panel shows one workshop. Each dot is an anonymous identifier with at least seven votes (78 identifiers in total), not a verified individual. Colours and letters identify clusters within a workshop, with author-coded labels describing distinguishing propositions. Coordinates and cluster labels are not directly comparable across panels or with the live display. Shading is a visual grouping aid, not an uncertainty interval. The maps illustrate how the collected voting data were visualised for reporting; they do not identify stable participant groups.

Reading the responses. P1–P76 are survey respondent codes; workshop locations and languages are written out in the tables. Each response identifies its source question: method comparison (57 answers in the full corpus), platform experience (53), or missing or misrepresented content after an AI-supported workflow (18). The corresponding question after the human-moderated workflow received 20 answers. The supplement retains the full multilingual prompts. The future-workshop preference was recorded separately from each written response: scores of 1–3 favour the human-moderated workflow, 4 means about the same and 5–7 favour the AI-supported workflow. The tables show this preference in words alongside the original score out of 7; it is not a rating of the quoted response. Missing preferences are labelled not recorded. Language and editorial status follow each excerpt. Workshop language does not establish the language of the answer. Paraphrases are unquoted. The responses describe participants’ perceptions. Capture, anonymity and influence were not independently verified through these responses.

Table 6. What participants valued and questioned about the two workshop formats. Shaded topic headings group the responses by what they discuss. Each row gives the respondent's workshop and separately recorded future-workshop preference. Scores of 1–3 favour the human-moderated workflow, 4 means about the same and 5–7 favour the AI-supported workflow.

Participant / workshop	Future-workshop preference	Selected response / source question
Topic: What the discussion outputs captured or missed		
P23 Zurich, English	AI-supported workflow (5/7)	"AI support records the voices, so there is no need for handwritten notes, and the model is good enough to summarise and classify subjects." (<i>Method comparison; English, edited excerpt</i>)
P36 Zurich, English	AI-supported workflow (7/7)	"Keeping track of all views. Visibility. Summary." (<i>Method comparison; English, edited excerpt</i>)
P51 Zurich, English	AI-supported workflow (6/7)	"In the AI method, everything is recorded and transparent. Everyone can also express themselves freely." (<i>Method comparison; English, edited excerpt</i>)
P1 Lausanne, English	Human-moderated workflow (1/7)	"AI support allowed us to discuss more, but I do not feel it grasped all the nuances in the discussion." (<i>Method comparison; English, edited excerpt</i>)
P3 Lausanne, English	Human-moderated workflow (1/7)	<i>Paraphrase:</i> P3 felt that the AI-supported workflow did not adequately reflect the sensitivity and subthemes of the arguments. (<i>Method comparison; English, paraphrase</i>)
P6 Lausanne, English	Human-moderated workflow (2/7)	"AI was inaccurate and dropped important points. AI is better at summarisation and automated tension extraction." (<i>Method comparison; English, edited excerpt</i>)
P29 Zurich, English	About the same (4/7)	"AI gathered all the thoughts, but not completely correctly. The human summary covered more topics and was accurate." (<i>Method comparison; English, edited excerpt</i>)
P75 Zurich, German	AI-supported workflow (7/7)	"The pattern of opinion could be shown better visually with AI." (<i>Method comparison; German, translated excerpt</i>)
Topic: Expressing views and discussing with others		
P2 Lausanne, English	AI-supported workflow (7/7)	"People tried to explain their point, and we could have voted on it. That was easier than confronting them when I disagreed." (<i>Method comparison; English, edited excerpt</i>)
P22 Zurich, English	AI-supported workflow (6/7)	"With AI, people can also give their point of view anonymously." (<i>Method comparison; English, edited excerpt</i>)

Continued on next page

Table 6 continued

Participant / workshop	Future-workshop preference	Selected response / source question
P69 Zurich, German	AI-supported workflow (5/7)	"Everyone could take a position on every statement. Opportunities to state my own position were more limited in the human-moderated method." (<i>Method comparison; German, translated excerpt</i>)
P10 Lausanne, French	Human-moderated workflow (1/7)	"With AI there was little disagreement and no nuance; the result did not take these nuances into account. Without AI, I was clearly more able to express myself and felt heard." (<i>Method comparison; French, translated excerpt</i>)
P15 Lausanne, French	Human-moderated workflow (1/7)	"With human moderation, we address the group; the moderator understands subtleties and asks for clarification." (<i>Method comparison; French, translated excerpt</i>)
P16 Lausanne, French	Human-moderated workflow (1/7)	"With human moderation, there were real exchanges and real listening between participants." (<i>Method comparison; French, translated excerpt</i>)
P62 Zurich, German	Human-moderated workflow (1/7)	"The second method did not feel like a discussion." (<i>Method comparison; German, translated excerpt</i>)
P74 Zurich, German	Human-moderated workflow (1/7)	"Murmi allowed less explanation and discussion than the blackboard method. The blackboard felt more like joint work." (<i>Method comparison; German, translated excerpt</i>)
P24 Zurich, English	Human-moderated workflow (3/7)	"Less discussion + critical reflections came in the bigger AI moderated discussion. I preferred the AI-moderated output which captured more statements." (<i>Method comparison; English, verbatim excerpt</i>)
Topic: Keeping up with the discussion and the platform		
P72 Zurich, German	AI-supported workflow (5/7)	"The app was much more efficient, including its analysis. It enabled everyone to express an opinion quickly in their own words." (<i>Method comparison; English, edited excerpt</i>)
P43 Zurich, English	About the same (4/7)	"AI-supported mode seems efficient, but face-to-face communication with human moderation lets people give more authentic responses." (<i>Method comparison; English, edited excerpt</i>)
P65 Zurich, German	About the same (4/7)	"I was more distracted when writing in the app. The Post-it method felt more collaborative." (<i>Method comparison; German, translated excerpt</i>)
P1 Lausanne, English	Human-moderated workflow (1/7)	"It was difficult to assess or vote and contribute to the debate at the same time. A quiet time at the end of the roundtable would have been more appropriate." (<i>Platform experience; English, edited excerpt</i>)

Continued on next page

Table 6 continued

Participant / workshop	Future-workshop preference	Selected response / source question
P26 Zurich, English	About the same (4/7)	"I love how it generates a report, but some of the verbal statements were not picked up by AI. The arguments are more individual, less collaborative. But very efficient for the workshop host!" (<i>Platform experience; English, edited excerpt</i>)
P29 Zurich, English	About the same (4/7)	<i>Paraphrase:</i> Writing helped P29 be heard despite shyness, but the app also distracted from listening. P29 suggested discussing the topics before voting. (<i>Platform experience; English, paraphrase</i>)
P52 Zurich, English	Human-moderated workflow (3/7)	"The dynamic sorting could be renewed less frequently, instead of changing all the time." (<i>Platform experience; English, edited excerpt</i>)
P59 Zurich, German	AI-supported workflow (5/7)	"Voting within groups would be good before publishing the final points to the whole group, so the cards can be revised." (<i>Platform experience; German, translated excerpt</i>)
P67 Zurich, German	AI-supported workflow (6/7)	"It all went too quickly for me. I find the report good and interesting, but I do not understand how it came about, so I am a little mistrustful. No one explained to me how to create an 'opinion card' that the whole group could then 'vote' on." (<i>Platform experience; German, translated excerpt</i>)
Topic: What participants wanted from deliberation		
P48 Zurich, English	Human-moderated workflow (3/7)	"I would have liked to hear from the others more, maybe switch tables etc." (<i>Platform experience; English, verbatim excerpt</i>)
P14 Lausanne, French	About the same (4/7)	"The human element is the basis of everything, even if AI adds something." (<i>Method comparison; French, translated excerpt</i>)
P20 Lausanne, French	Human-moderated workflow (2/7)	"With AI moderation, deliberation does not do its job: we stop at the first polarisation." (<i>Method comparison; French, translated excerpt</i>)
P67 Zurich, German	AI-supported workflow (6/7)	"Very different opinions in the small group; a lot of time lost trying to find 'consensus'." (<i>Method comparison; German, translated excerpt</i>)
P76 Zurich, German	About the same (4/7)	"These 38 cards describe the discussion a little better. But I would use both methods together so there is also a human discussion. Pure AI moderation is not perfect." (<i>Method comparison; German, translated excerpt</i>)

The final theme contrasts expectations of deliberation. P20 wanted movement beyond initial positions, while P67 regarded time spent finding consensus as time lost. These responses do not establish whether participants' opinions changed. P6 and P29 each valued aspects of capture while raising concerns about accuracy or discussion.

B.2 Missing or misrepresented content after each discussion

Participants gave these responses about the output of one discussion. They answered the overall method-comparison question separately. They contextualise Sections 4.2 and 4.3 without identifying transcript speakers or linking a concern to a particular edit. Table 7 retains four AI-output diagnostics. Two further human-output diagnostics are presented immediately below it.

Table 7. Participants’ descriptions of missing or misrepresented content after the AI-supported workflow. Future-workshop preferences were recorded separately: scores of 1–3 favour the human-moderated workflow, 4 means about the same and scores of 5–7 favour the AI-supported workflow.

Participant / workshop	Future-workshop preference	Selected response / source question
P5 Lausanne, English	Not recorded	“We had a long discussion about if the government was able to take on that role and it’s appearing as a detail in the report (hierarchy is missing).” (<i>AI-output diagnostic; English, verbatim, terminal punctuation added</i>)
P16 Lausanne, French	Human-moderated workflow (1/7)	“The nuances, for example the question of the role of school, did not appear clearly.” (<i>AI-output diagnostic; French, translated excerpt</i>)
P17 Lausanne, French	About the same (4/7)	“The question of the purpose of school was underrepresented.” (<i>AI-output diagnostic; French, translated excerpt</i>)
P21 Lausanne, French	Not recorded	“No.” (<i>AI-output diagnostic; French, translated</i>)

For the output of the human-moderated workflow, P24 wrote, “*More unpopular views were tried to be amended before the write-up*” (Zurich English). P75 reported that some small groups dominated the result through their many cards (Zurich German, paraphrased). These responses describe concerns about selection without establishing systematic exclusion. P21’s report of no missing content after the AI-supported workflow does not establish that a correction resolved a concern.

B.3 Speech capture, card sources and response coverage

Across the four rooms, the archive contained 143 transcript entries and 157 candidate cards: 117 from speech, 37 entered directly, and three generated to highlight tensions. Moderators approved 136 cards, including 96 created from speech. The conservative audit identified two clear reversals, ten duplicate or near-duplicate cards and 19 candidates drawn from instructions, review, report narration or talk outside the policy discussion. Categories overlap. In Lausanne English, a claim about the state’s expertise was represented as a limitation in AI’s ability to educate workers. Later, a suggestion to begin with an imperfect law became a rejection of using an imperfect AI system. These examples show changes to what people were asked to vote on. They do not estimate the rate of transcription errors.

We ordered propositions by creation time within each room and divided them into four nearly equal groups. Coverage was the share of non-host identifiers in the timestamped log responding to a proposition. Support, neutrality and opposition used all recorded responses, including neutral responses, as their denominator. We also measured time from card creation to the final report. Card order and available response time were therefore intertwined. A further test accounting for room compared voting activity between first-summary reviewers and other identifiers. Updated responses retained only their latest value and timestamp.

The datasets are reconciled in Appendix A.1. Figure 9 shows response coverage and recorded choices across the card sequence. Card-order correlations with response coverage were $\rho = -0.84$ in Zurich English, -0.92 in Lausanne French and -0.56 in Zurich German (adjusted $p \leq 0.00046$), and -0.26 in Lausanne English (adjusted $p = 0.119$). Later cards also had less available time, so these associations cannot identify why coverage declined. We found no corresponding change in support or opposition over the sequence.

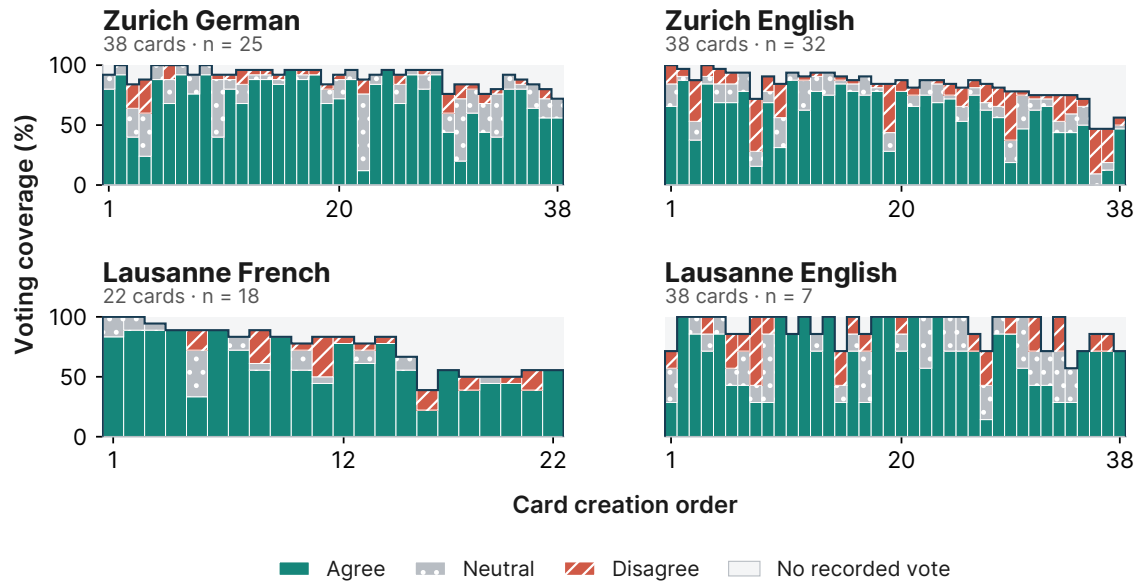


Fig. 9. Voting coverage and response direction across cards in creation order. Each column is one approved card. Agree, neutral and disagree segments use the same denominator: the room's n non-host account or device identifiers with at least one retained vote. The total coloured height shows coverage; the pale space above it indicates no recorded vote. Strong agreement and disagreement are combined with agreement and disagreement respectively. Dots mark neutral votes and diagonal hatching marks disagreement. Later cards generally had lower coverage and less time for responses before reporting. These patterns do not measure attention or attitude change.

Across the 157 retained candidate cards, four were marked edited: two in Zurich German and one in each English-language room. All four were approved and received responses. The flag identifies cards whose wording changed, but does not establish the number of edits, whether voting preceded an edit or whether a correction resolved a concern.

An instruction in Zurich German to analyse the data first became an approved proposition and received 23 responses. Moderator approval therefore did not establish that every card represented a participant's policy view.

Direct entry provided another input route. In Zurich German, 26 of 38 approved cards were entered directly and twelve came from audio. Their topics included age limits, basic skills, privacy, teacher training, assessment, equal access and human skills. People used direct entry for these topics, although the records leave open whether an entry repaired an omission or changed the later summary. The survey comments separately report a cut sentence that nearly reversed meaning (P29, Zurich English) and a request for two microphones (P2, Lausanne English). The decision to discard audio reduced privacy risks and limited later checks to the retained text.

Recorded voting activity extended across the AI-supported workflow. Among non-host identifiers with card responses, the median number of responses per record ranged from 18.5 to 36 across rooms, and the median interval

between first and last recorded responses ranged from 5.5 to 23.0 minutes. Between 46% and 75% of consecutive responses were submitted within ten seconds of each other. These measures describe recorded voting activity. They cannot show whether participants were listening or inspecting a summary at the same time.

Among the 82 identifiers in the voting-log analysis, 61 had a recorded response to the first summary and 21 did not. Using retained votes timestamped before the first summary was generated in each room, median coverage of the propositions available by that point was 89.5% among those with a review response and 76.3% among those without one (Holm $p = 0.047$). This difference does not establish exclusion or explain why a summary judgement was submitted. The late Lausanne French card in Section 4.2.2 was absent from summary citations, but the records do not establish that response coverage explains its absence.

B.4 Additional comparisons of participant ratings

These checks supplement the estimates in Section 4.1. Of 75 paired overall ratings, 41 favoured the human-moderated workflow, 30 favoured the AI-supported workflow and four were equal. The Wilcoxon signed-rank comparison gave $p = 0.207$, alongside the sign-flip result reported in the main text. Future preference did not clearly differ from the scale midpoint (Wilcoxon $p = 0.158$, Benjamini–Hochberg $q = 0.257$ across four comparisons), but differed across rooms ($\epsilon^2 = 0.18$, permutation $p = 0.001$). Table 8 retains the room-level context. These contrasts do not isolate format or room-size effects.

Table 8. Future-method preference by room. Higher values favour the AI-supported workflow on the seven-point scale.

Room	Valid responses	Mean preference
Lausanne English	6	3.00
Lausanne French	13	2.00
Zurich German	19	4.16
Zurich English	34	4.24

C SUMMARY REVIEW AND REPORTING

C.1 Linked judgements and all written review reasons

This evidence supports Section 4.3. Among 36 identifiers linked across both reviews, 30 kept their judgement, five moved to “good enough” (three in Zurich English and two in Lausanne English), and one moved away (Lausanne French). Changing reviewer sets, summary content and missing review timestamps prevent attributing these transitions to particular corrections. We examined Lausanne French because it had the lowest final-report approval (69%, versus 82–91% elsewhere). Its difficulties should be understood in that context and cannot be assumed to represent all four workshops.

Table 9 retains all ten written reasons among 107 review responses: five accompanied first-version rejections and five accompanied second-version responses, including one approval. Eight identifiers supplied them. SR1–SR8 identify records only within this table. Room codes are LE (Lausanne English), LF (Lausanne French), ZE (Zurich English) and ZG (Zurich German). “Accepted” and “Rejected” abbreviate “good enough” and “not good enough”. A judgement without a written reason is distinguished from an absent judgement. Pairing requires the same room and anonymous identifier. SR1 is probably linked to P2, while SR2 has a verified link to P6, but neither response can be linked to an edit. The survey diagnostics in Appendix B.2 remain a separate source.

Table 9. All retained written review reasons, paired across summary versions where the same account or device identifier is available.

Record / room	Version 1	Version 2
SR1, LE	Rejected No written reason recorded.	Accepted “I disagree with the recommendations, but it’s OK with what the group thinks.” <i>English, edited</i>
SR2, LE	Rejected No written reason recorded.	Rejected “Missing texture, details and color. Some issues dropped. Some misunderstood.” <i>English, verbatim</i>
SR3, LF	Rejected “Some statements were oversimplified. Yet in a debate, everything hinges precisely on nuance.” <i>French, translated</i>	Rejected “Same comment as before. Too much flattening of arguments and ideas.” <i>French, translated</i>
SR4, LF	Rejected “Overall, it is fairly good. Some dimensions that were presented are missing and were overly smoothed to fit a general mould.” <i>French, translated</i>	Rejected “As before, it is too smooth, too fixed; it does not show the diversity of the discussions.” <i>French, translated</i>
SR5, ZE	Rejected “How AI could be used to create a new job market.” <i>English, terminal punctuation added</i>	No review recorded for this identifier.
SR6, ZE	Rejected “Safety & Security measures.” <i>English, terminal punctuation added</i>	No review recorded for this identifier.
SR7, ZE	Rejected No written reason recorded.	Rejected “I think the group discussion helps but does not stimulate people that talk less or have more unpopular opinions to speak more.” <i>English, verbatim</i>
SR8, ZG	Rejected “Too general; no reflection on the overall opinion.” <i>German, translated</i>	No second review before reporting.

SR3 and SR4 repeat their concerns across versions. SR1 accepts the summary while disagreeing with its recommendations. SR5 and SR6 have no linked second review, which does not establish resolution. Zurich German completed review with version 1. As a contrasting setting for revision, Lausanne English retained 38 propositions, card responses from seven identifiers and eleven summary items across both versions. The overview and six citation sets changed,

while all eleven substantive item texts remained unchanged. Of the five identifiers with judgements on both versions, two changed to acceptance. The available histories cannot connect these changes to a specific comment or revision.

C.2 Votes and summary-review responses

To compare agreement with source propositions and acceptance of their summary, we calculated support for each identifier with a summary judgement. Let J_{iv} contain the unique propositions cited by summary version v that non-host account or device identifier i answered, and let $a_{ij} = 1$ for agreement or strong agreement and 0 for any other recorded response. Its support share was

$$s_{iv} = \frac{1}{|J_{iv}|} \sum_{j \in J_{iv}} a_{ij}, \quad |J_{iv}| > 0. \quad (2)$$

Each cited proposition counted once, even if several summary items cited it. Unlike the responses used in the pairwise rule in Equation 1, this measure included explicit neutral responses in the denominator. Missing responses were excluded, as were records with no answered cited propositions. We used the latest retained responses and pooled the first summary from each room, giving each included identifier equal weight within the approving and rejecting sets.

The 44 approving records supported an average of 77.9% of their answered cited propositions. The 17 rejecting records supported 76.1%. The difference was 1.8 percentage points (room-stratified permutation $p = 0.683$). Eleven of the 17 rejecting records supported at least three quarters of their answered cited propositions. Reviewers could support many cited propositions and still reject the summary. These results do not establish statistical equivalence between the groups or identify why each record rejected the summary. The archive does not reconstruct votes at the moment of review, and a working citation need not substantively support a summary claim.

C.3 Zurich English: summary wording and supporting references

Tables 10 and 11 let readers inspect the summary changes discussed in Section 4.3. Version 1 was generated on 11 August 2026 at 19:09:04 and version 2 at 19:16:55 (Europe/Zurich). The generation snapshots held 29 and 35 approved propositions respectively, with card responses from 32 identifiers in each. Approval rose from 17 of 25 summary-review responses to 21 of 23, but these came from different sets of identifiers and need not match the feedback available when generation was requested.

All eleven original policy items retained their wording. Version 2 added a recommendation about access to training and a condition about ethical implementation. Equitable access had already appeared as an unresolved question. The overview, group descriptions and all retained items' reference sets changed. Table 11 therefore prints each policy item once, alongside both versions' references. Group descriptions are presented by version because matching letters do not establish continuity of membership. The system's own descriptions of its revisions are reproduced below with attribution. These descriptions have not been independently verified.

The source-link audit clarifies what these references establish. Across rooms, 40 of 45 summary items had links that resolved, but the storage fallback in Appendix A.2 could attach an unrelated proposition. In Zurich English, eight version-2 items cited S01, "AI tends to replace people in the workforce", including items about training, public-sector use and health-care ethics. The other five had no stored links. A contrasting Lausanne English item retained relevant references and divided responses (Table 13). A resolving reference alone does not show that the proposition supports the claim.

SR5 asked about a new job market and SR6 about safety and security. Neither has a linked second review. The job-displacement question persisted while a general ethical condition was added. During the intervening discussion, participants raised financial issues and who benefits. The 19:13:45 card S32 stated, “Companies should not prioritize profit and efficiency at the expense of ethical considerations.” It received 24 recorded responses, entered the final proposition set and was not cited by the final summary. Its timing and topic suggest a connection to the discussion, but its ethical framing was not identical to the spoken concern, and no source link establishes the connection. Neither the card nor higher approval establishes that a particular objection caused or was resolved by the revision. SR7’s second-version concern instead addressed opportunities for quieter people and unpopular views.

Table 10. Complete stored overview and group descriptions for Zurich English. Version 1: 19:09:04. Version 2: 19:16:55. Row headings are ours.

Version	Stored text
Overview	
v1	“The discussion around AI at work reveals a strong consensus on the need for AI to augment human skills rather than replace them, with significant support for training workers in AI competencies. However, there are notable divisions regarding the role of AI in sensitive areas such as decision-making and its implications for job security. While some groups emphasize the need for skills that are less technical and harder for AI to replicate, others express concerns about the ethical implications of AI’s role in the workforce.”
v2	“The discussion around AI at work reveals a strong consensus on the need for AI to augment human skills rather than replace them, with significant support for training workers in AI competencies. Groups generally agree on the importance of developing skills that are harder for AI to replicate and the ethical implications of AI’s role in sensitive decision-making contexts. However, divisions remain regarding the extent of AI’s integration in certain fields and the potential for job displacement.”
Group descriptions (labels are local to each version)	
v1	A: “Support for AI Augmentation” “Group A emphasizes the importance of using AI to augment human skills and supports developing skills that are harder for AI to replicate.” B: “Strong AI Advocacy” “Group B strongly advocates for the augmentation of human skills by AI and supports the need for workers to be trained in AI skills.” C: “Skill Development Focus” “Group C is focused on the importance of skill development, particularly in augmenting human capabilities with AI.” D: “Comprehensive AI Support” “Group D supports a comprehensive approach to AI, advocating for its use while also emphasizing skills development.”

Continued on next page

Table 10 continued

Version	Stored text
v2	A: “Augmentation Focus” “Group A emphasizes the need for AI to augment human skills rather than replace them, advocating for training in interpersonal skills and AI competencies.” B: “Strong Support for AI Training” “Group B strongly supports the integration of AI in the workforce, particularly through training programs that enhance workers’ AI skills and ensure ethical considerations in its application.”

System-generated version descriptions. Version 1 states: “This is the first version of the mediation summary, capturing the diverse opinions surrounding the use of AI in the workplace and identifying common ground and areas of disagreement.” Version 2 states: “This version refines the mediation summary and recommendations based on new insights from the group discussions, emphasizing the importance of ethical AI use and the need for comprehensive training programs while addressing unresolved questions about job displacement and equitable access.” These descriptions are reproduced as stored and are not independent assessments of the revisions.

R, C, T and Q denote recommendations, essential conditions, trade-offs and unresolved questions. S01–S35 abbreviate ZH-EN-ST001–ZH-EN-ST035. “Added” means absent as a policy item in v1, not necessarily absent as an idea from the discussion. Wording is reproduced from the application-normalised outputs, not reconstructed screens or raw model responses. References reproduce stored links without endorsing their relevance.

Table 11. All Zurich English policy items, with unchanged wording printed once. The last two columns show stored proposition references in each version.

Item	Policy wording	v1 links	v2 links
R1	“Encourage the development of skills that are complementary to AI, focusing on interpersonal communication and critical thinking.”	S04, S05	S01
R2	“Implement training programs for workers to enhance their AI skills and ensure they can work effectively alongside AI technologies.”	S11, S16	S01
R3	“Establish policies to support workers transitioning from jobs that may be replaced by AI, ensuring they have access to retraining opportunities.”	S18	S01
R4	“Define clear guidelines on the ethical use of AI in decision-making processes, especially in sensitive areas like healthcare.”	S08, S21	S01
R5	“Promote the integration of AI in public sector decision-making to enhance efficiency while maintaining human oversight.”	S19, S20	S01
R6	“Ensure that training programs for AI skills are accessible to all workers to prevent inequities in the workforce.”	Not in v1	None stored
Added			
C1	“AI must be used to enhance human capabilities rather than replace them, ensuring job security for workers.”	S02	S01
C2	“Workers should not be penalized for opting out of AI use on moral grounds.”	S12	S01

Continued on next page

Table 11 continued

Item	Policy wording	v1 links	v2 links
C3 Added	“Policies must ensure ethical considerations are prioritized in AI implementations.”	Not in v1	S01
T1	“Balancing the integration of AI in the workforce with the ethical implications of its use in sensitive decision-making.”	S08, S21	None stored
T2	“The need for AI skills training versus the risk of creating a divide between those who can adapt to AI and those who cannot.”	S11, S03	None stored
Q1	“How to effectively implement AI training programs across diverse sectors to ensure equitable access for all workers?”	S24, S26	None stored
Q2	“What specific policies should be established to protect workers from job displacement due to AI advancements?”	S18, S23	None stored

C.4 Changes after the reviewed summary

This comparison supports the distinction between reviewing a summary and approving the later report in Section 5.3. Three reports paired the reviewed summary with responses added or updated after generation. Zurich English also added three propositions. Only Lausanne English retained unchanged propositions and responses. Participants saw the final report before the questionnaire, but their recorded approval applied to a summary version.

Figure 6 shows the overall sequence and changes between summary versions. Table 12 instead reports changes from the last pre-report summary to the report. These are different intervals. Added or updated responses are confirmed changes in incomplete retained records, not necessarily newly submitted votes.

Table 12. Recorded changes from the last pre-report summary to the report. Minutes measure the interval from summary generation to the report snapshot.

Room	Version	Added cards	Response changes	Minutes
Zurich German	v1	0	112	3.24
Zurich English	v2	3	165	4.72
Lausanne French	v2	0	2	1.97
Lausanne English	v2	0	0	1.53

Zurich German retained 38 propositions while the number of identifiers with card responses rose from 23 to 26. Zurich English’s summary used 35 propositions and card responses from 32 identifiers. The report held 38 propositions, and 48 response changes concerned the new cards.

Late tension prompts in Zurich English. The transcript announced three AI-generated tensions at 19:19:57. All three cards were created at 19:20:11, after the last reviewed summary at 19:16:55 and 87 seconds before the report snapshot at 19:21:38. Their published response counts were 22, 20 and 19, compared with 15, 15 and 18 in the incomplete non-host event log. No further summary was generated. This records a late trial and responses, but not its effect on deliberation or the reviewed summary.

D WORKED EXAMPLES OF CORRECTION AND DISCUSSION

These examples support Section 4.2. They connect machine-transcript passages with cards and summaries through wording and timing. Stored timestamps use Europe/Zurich and need not mark the beginning or end of a contribution. Recorder labels do not identify speakers, and no survey codes are assigned to transcript passages. Source links, earlier wording and parts of the edit and review history were unavailable. These limits apply throughout the examples.

D.1 Lausanne English: a workload concern across summaries

Table 13 follows the workload concern discussed in Section 4.2.2. Minutes are rounded from the first stored transcript entry at 18:44:10, including setup. C1–C3 abbreviate LS-EN-ST021–LS-EN-ST023. All were approved, had no stored edit flag and received responses from the same seven non-host identifiers. Support combines agreement and strong agreement, and opposition combines disagreement and strong disagreement. Response windows describe the latest retained responses, all preceding v1. Speech-to-card connections are inferred, whereas summary-to-card citations are stored. The question also appeared unchanged in the final report. Neither version recommends reduced hours, and no item-specific response establishes satisfaction with the representation.

Table 13. Lausanne English workload sequence. Transcript excerpts, card wording and summary questions are reproduced as stored. Minutes indicate transcript entries, card creation or summary generation. Response windows appear separately.

Stage / minute	Recorded content	Responses / citations
Speech / 28	“there is a real risk of mental overload and burnout”, “I start to feel it myself.”	Within discussion of protecting workers.
C1 / 29	“The risk of mental overload and burnout among workers is heightened due to parallel project work.”	7 support, 0 neutral, 0 oppose. Votes: minutes 34–41.
Speech / 30	“Maybe reducing the work time hours. The work is more intense.”	A possible response to increased intensity.
C2 / 30	“Reducing work time hours could be beneficial in the context of AI integration in the workplace.”	4 support, 3 neutral, 0 oppose. Votes: minutes 35–41.
C3 / 30	“The intensity of work has increased due to the integration of AI.”	7 support, 0 neutral, 0 oppose. Votes: minutes 35–41.
Summary v1 / 43 and v2 / 48	“How can we assess the impact of AI integration on worker mental health and workload?”	Same wording. v1 cites C1 and C2. v2 cites C1 and C3.

D.2 Lausanne French: records behind continuing objections

This chronology supports the shorter description in Section 4.2.2. Minutes are rounded from the room’s first stored transcript entry, including setup. French excerpts are translated. Lausanne French was the only revising room where approval decreased, although most summary-review responses still approved version 2.

- (1) **Minutes 16–18: uncertainty enters discussion and cards.** A transcript passage raises teachers’ fear of losing their previous way of teaching and asks what school should be for. Two related cards describe fear of losing skills and the need to reconsider school’s purpose. These concerns were present before summary review.
- (2) **Minutes 22–24: the first summary’s framing is challenged.** The summary describes broad consensus on integrating AI learning. Following an invitation to explain objections, the discussion challenges its settled tone:

“We have resolved absolutely nothing.” The objection concerns how the discussion is represented, as well as which topics appear.

- (3) **Minutes 26–27: renewed discussion produces a shorter claim for voting.** Asked for missing arguments, a contribution says school must reinvent itself, expresses uncertainty about “where” and “how”, and describes emotion absent from the summary. A card created in the same minute states “School must reinvent itself”, retaining the topical claim without that uncertainty or criticism of the summary. The next minute, the transcript records an announcement of further voting on newly extracted arguments before another report is generated. It does not identify this card individually.
- (4) **Minutes 29–31: support for the card coexists with continuing criticism.** The card’s ten retained responses were supportive, across an interval in which version 2 was generated at minute 30. Two identifiers contained rejections of versions 1 and 2 with repeated objections. One comment read: “Same comment as before. Too much flattening of arguments and ideas.” Review comments lack timestamps, so their order relative to individual card responses cannot be established.

The card “L’école doit se réinventer” received ten supportive responses with retained timestamps spanning minutes 29–31. Nine precede version 2’s generation at minute 30. These are the latest retained responses, not every ballot at generation. The card remained in the final proposition set but was absent from final summary citations. Nine of the ten substantive summary items kept their wording between versions, one recommendation was reworded and both open questions remained unchanged. Matching wording and timing support the card’s connection to the spoken concern, although no source identifier establishes that connection directly. No unambiguous transcript passage identifies a subsequent complaint as a reaction to this particular card. The repeated rejections concern different records from the one that moved from approval to rejection.

The same anonymous reviewer identifiers link the two first-version objections about oversimplification and missing dimensions to second-version rejections repeating those concerns (SR3 and SR4 in Table 9). This establishes persistence across versions without locating the comments relative to edits or votes. P16’s and P17’s separate AI-output diagnostics concern school’s role or purpose (Table 7). They contextualise the sequence without identifying transcript speakers or the authors of the linked review comments. The later card and repeated objections document continued discussion of this concern. Its resolution cannot be established from support for the card or changes to the summary.

D.3 Related correction and grouping exchanges

Table 14 brings together the shorter exchanges used to contextualise correction and participation. The further Zurich English summary-review case is presented with its generated outputs in Appendix C.3. The late tension prompts belong to the report comparison in Appendix C.4.

Table 14. Additional recorded exchanges about correction, input and live grouping. The evidential limits stated at the start of Appendix D apply throughout.

Room / focus	Recorded exchange and its limits
Lausanne English Card correction	Entries at 18:58:44–19:03:00 concern the state’s AI-governance capability, which statement to delete and whether the interface interrupted discussion. A card created at 19:01:16 was marked edited, had seven responses and was cited by two final recommendations. Its creation time is not its edit time. Previous wording and deletion history were not retained, so the correction action and its adequacy cannot be reconstructed.
Zurich German Direct entry	At 18:58:22, the transcript notes many existing entries and invites a short spoken round, consistent with the workshop lead’s recollection that typing came first. Later prompts invite voting and summary review. No equally clear correction-to-summary connection was identified. The 26 directly entered cards establish use of another input route, not that they repaired omissions (Appendix B.3).
Lausanne English Group identity	At 19:28:30 the discussion turns to groups A and B during review. At 19:30:52 a contribution questions whether the dominant group matches the discussion’s balance of government and company responsibility. At 19:31:21: “I feel that I’m group B, I feel that I’m not group A.” These are voiced interpretations, not verified assignments, counts or evidence of a subsequent change.
Zurich English Changing display	At 19:14:49 participants try to inspect group C, note that it disappeared while the display updated and return to A and B. The cause and effect on views are not established. P52’s separate survey request for less frequent “dynamic sorting” appears in Table 6, but does not identify a particular display.
Lausanne French Explaining groups	At 19:01:28 and 19:02:06, explanations introduce the map’s points and changing groups through education-related concerns. At 19:14:48, the explanation stresses shared ground and differences of emphasis, suggesting people might recognise themselves in several clusters. This documents explanation, not participants’ comprehension.

Figure 2(e) illustrates the live grouping interface. Neither these exchanges nor the retrospective maps in Figure 8 reconstruct individual screen views, assignments or changes in opinion.