# Bringing Everyone to the Table: An Experimental Study of LLM-Facilitated Group Decision Making

MOHAMMED ALSOBAY, Microsoft Research, USA
DAVID M. ROTHSCHILD, Microsoft Research, USA
JAKE M. HOFMAN, Microsoft Research, USA
DANIEL G. GOLDSTEIN, Microsoft Research, USA

Group decision-making often suffers from uneven information sharing, hindering decision quality. While large language models (LLMs) have been widely studied as aids for individuals, their potential to support groups of users, potentially as facilitators, is relatively underexplored. We present a pre-registered randomized experiment with 1,475 participants assigned to 281 live groups completing a hidden profile task—selecting an optimal city for a hypothetical sporting event—under one of four facilitation conditions: no facilitation, a one-time message prompting information sharing, a human facilitator, or an LLM (GPT-4o) facilitator. We find that LLM facilitation increased information shared within a discussion by raising the minimum level of engagement with the task among group members, and that these gains came at limited cost in terms of participants' attitudes towards the task, their group, or their facilitator. Whether by human or AI, there was no significant effect of facilitation on the final decision outcome, suggesting that even substantial but partial increases in information sharing were insufficient to overcome the hidden profile effect studied. To support the design and evaluation of LLM-mediated group decision-making systems, we release our data and our experimental platform, the Group-AI Interaction Laboratory (GRAIL), as an open-source tool.

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

Additional Key Words and Phrases: decision making, human-AI interaction, collective intelligence, experiments

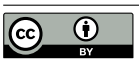
**ACM Reference Format:**
Mohammed Alsobay, David M. Rothschild, Jake M. Hofman, and Daniel G. Goldstein. 2026. Bringing Everyone to the Table: An Experimental Study of LLM-Facilitated Group Decision Making. *Proc. ACM Hum.-Comput. Interact.* 10, 6, Article CSCW059 (October 2026), 36 pages. https://doi.org/10.1145/3816907

## 1 Introduction

From tribal councils to committee meetings, group decision-making is a cornerstone of human societal behavior across eras and cultures. Groups are assembled to make decisions in critical functions such as hiring [72], patient care [18], and crisis response [40], and factors moderating the performance of groups have been studied extensively [1, 23, 32, 36, 37, 46]. Although group-based tasks are ubiquitous in work environments, groups face challenges such as social loafing [38], groupthink [6], and polarization [26], as well as various biases in information sharing, discussion, and evaluation [50, 60, 67].

Many efforts have been made to measure group aptitude and to design interventions that improve group performance. One such class of interventions involves the use of algorithms to influence

Authors' Contact Information: Mohammed Alsobay, Microsoft Research, New York, NY, USA, malsobay@microsoft.com; David M. Rothschild, Microsoft Research, New York, NY, USA, davidmr@microsoft.com; Jake M. Hofman, Microsoft Research, New York, NY, USA, jmh@microsoft.com; Daniel G. Goldstein, Microsoft Research, New York, NY, USA, dgg@microsoft.com.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2573-0142/2026/10-ARTCSCW059
https://doi.org/10.1145/3816907

collective behavior, ranging in sophistication from simple, noisy bots that improve coordination in networked groups [12], to graph neural networks that rewire human networks to promote cooperation [47]. With the launch of ChatGPT in November 2022, large language models (LLMs) have entered the public lexicon and become the focus of research exploring how these models can augment human abilities. While a broad body of literature has explored how LLMs can assist *individual* users in a wide range of domains, including writing, programming, art, and education [33, 51, 53, 79], our current understanding of how these models can assist *groups* of people using them together is relatively nascent [11, 39, 44, 75].

Recent work has demonstrated the ability of LLMs to help groups achieve consensus [64, 73] and discuss divisive topics [5] through static interventions (i.e., summaries or suggestions delivered to group members). In contrast to post hoc tools such as meeting summaries, integrating LLMs as active participants in discussions offers the opportunity to support groups and increase discussion quality in real-time, by increasing desirable outcomes such as intentionality [10] and inclusion [25]. More specifically, as meeting facilitators, LLMs may excel in aspects of facilitation that would be taxing for their human counterparts, particularly in managing discussions that proceed at a high volume and rapid pace and keeping track of who has and has not participated. In this direction, recent work has addressed the design principles underlying such active LLM-based interventions [44] and empirically studied the effect of LLM-based interventions such as dissent on group decision-making and design processes [11, 39]. Together, these studies point to a growing design space of LLM facilitation strategies that vary in their timing, target, and interaction style, and in doing so, shape how groups coordinate and deliberate.

In this work, we study the role of LLM-based facilitators in decision-making discussions in which critical information is asymmetrically distributed among group members, requiring effective information sharing to identify the optimal solution. In such settings, groups often fail to surface and effectively integrate uniquely held information, due to a combination of factors including uneven participation, conversational dynamics, and biases in how information is attended to and interpreted. These dynamics can lead to suboptimal decisions, particularly when key information is known to only a subset of participants, or even a single individual.

We introduce LLMs as facilitators to aid groups completing "hidden profile" tasks (HPTs) [67], a well-studied experimental paradigm of decision-making under asymmetric information. By enacting structural interventions that support participation and information synthesis, LLMs may be effective aids to groups facing such decisions. To further our understanding of LLM facilitation in this setting, we ask the following research questions (RQs):

- **RQ1:** Relative to no facilitation, written guidance, and human facilitation, does LLM facilitation increase information sharing in group decision-making under asymmetric information?
- **RQ2:** Beyond the volume of information shared, how does LLM facilitation affect the process of information aggregation, and the quality of the decision made?
- **RQ3:** How do participants subjectively evaluate LLM-facilitated experiences of group decision-making?

To investigate these questions, we conducted a pre-registered randomized experiment (AsPredicted #192061) in which 281 groups of five decision makers completed a hidden profile task and were randomly assigned to complete the task either unassisted or with the help of a human facilitator or an LLM facilitator. We find that LLM facilitation increased the volume and breadth of information shared, without negatively impacting self-reported attitudes of group cohesion and productivity. However, despite the increased information sharing, groups still made suboptimal decisions, emphasizing the strength of the biases at play and motivating research into algorithmic facilitation tools that are targeted at overcoming well-documented biases in group discussion and

decision making. In support of future research, we also release the full experimental dataset and the Group-AI Interaction Laboratory (GRAIL), an open-source interface for studying group-AI interaction in conversational tasks.

## 2 Related Work

In this work, we study the impact of LLM facilitation on group decision making under asymmetric information. Our research is motivated by the well-documented biases and communication failures exhibited by groups in settings of asymmetric information, and builds upon previous work in the design of group-AI interaction and group decision support technologies. We contribute to these lines of research by complementing existing studies with large-scale, pre-registered experimental evidence describing the impact of LLM facilitation on group behavior while completing a challenging hidden profile task.

### 2.1 The Hidden Profile Task Paradigm of Group Decision-making Research

Since the seminal work of Stasser and Titus [67], hidden profile tasks (HPTs) have been used extensively to study the dynamics of decision making in groups whose members have asymmetric information [9, 41, 58, 65, 69]. In an HPT, a group of decision makers is tasked with selecting one of several predefined options, with information about the options distributed among different group members. Each group member receives a distinct information set (or "profile") about the various options, typically delivered as a list of facts about each option. The visibility (public/shared/private) and polarity (positive/negative/neutral) of pieces of information and their distribution among group members is deliberately structured such that individually accessible information suggests one alternative as optimal, while the full information set would identify a different option as superior. This information asymmetry creates a decision scenario where optimal group performance requires effective information exchange and integration rather than simple majority voting or averaging of individual preferences.

One familiar with the literature on the wisdom of crowds [71] and collective intelligence [78] may intuitively expect groups to outperform individuals in decision making tasks, including HPTs. However, prior research finds that groups typically do not exhaustively share information due to effort and time constraints [67]. The distribution of information under the HPT induces two key information sharing biases that hinder group progress. In groups completing an HPT, shared information tends to be surfaced more frequently than private information (shared information bias [35, 52]), and participants are more likely to share information that supports the option that is appealing under their personal information set (individual preference bias [28, 60]).

More broadly, these dynamics are amplified by well-documented group processes. For example, social loafing [38] can reduce participation, increasing the chance that privately held information is never shared; communication apprehension [45, 52] can discourage individuals from contributing unique or dissenting information; and groupthink [6] can lead groups to converge prematurely on an initially favored option without fully considering alternatives. Collectively, these factors contribute to a central challenge of the HPT paradigm: groups often fail not because the relevant information is unavailable, but because it is not effectively surfaced and integrated during discussion.

### 2.2 Computational Approaches to Group Facilitation

Broadly put, the role of a facilitator is to enact process and content interventions that help a group achieve its objectives [49, 76]. *Process interventions* introduce and reinforce structure in a group's discussion, such as imposing an agenda or by managing specific discussion formats (e.g., the Delphi method and the nominal group technique [21]). In conversation, facilitators enacting process interventions typically perform actions such as timekeeping, managing participation, and

supporting synthesis, without introducing their own substantive judgments or domain expertise. In contrast, *content interventions* involve the facilitator contributing their own judgment, interpretations, or recommendations to the discussion at hand, such as advocating for a particular option or introducing new information and perspectives.

Computational approaches to facilitation build on a rich body of work in the CSCW and broader HCI communities on group decision support systems (GDSSs) [16] and conversational agents (CAs) [17]. In 1993, Phillips & Phillips argued that, while computers could support facilitation as a medium for interaction, "the key functions of the facilitator, observing, attending, maintaining awareness of feelings, and intervening, cannot, of course, be carried out by computers…" [56]. Since then, computational approaches to these functions have arguably progressed to the level of practical utility, and subsequent work has explored technological solutions to enact the structural and conversational roles of the facilitator in group settings. For example, Kim et al. [31] demonstrate the utility of CAs that interactively add structure (e.g., turn-taking) to group debates, while Do et al. [19] show how agents can sensitively provide feedback to underperforming group members, and Shamekhi et al. [61, 62] examine how embodiment influences perceptions of rapport, trust, intelligence, power, and efficacy of facilitative agents in decision-making groups.

In the context of decision making under asymmetric information, process interventions may, in theory, help groups avoid suboptimal decisions. In hidden profile tasks, group members collectively possess the information required to identify the optimal choice, yet this information is often unevenly shared and insufficiently integrated into the final decision. By eliciting contributions from all members and structuring how information is shared and synthesized, process interventions are well positioned to mitigate these failures of information exchange. This makes hidden profile tasks a natural setting in which to study the effects of process-oriented facilitation. At the same time, while such interventions may be beneficial in theory, it remains an empirical question whether LLM-based systems can implement such facilitation effectively in practice.

Effective LLM facilitators may allow us to easily conduct these facilitatory functions with greater fidelity and scale, and across a wide range of domains, potentially broadening access to facilitative support. Professional facilitators, such as those found on Find a Facilitator, command fees in the range of $1,500-$3,000/day[1], which may be out of reach for groups in a wide range of decision making settings. The comparatively low (and decreasing) marginal cost of LLM-based systems may offer a more scalable and cost-effective alternative.

### 2.3 Group-AI Interaction

A broad range of algorithms have been used to influence group behavior with the goal of overcoming detrimental group processes, ranging in sophistication from simple, noisy bots that improve coordination in networked groups [12], to graph neural networks that rewire human networks to promote cooperation [47] and variational autoencoders used to influence group dynamics in a hidden profile task [55].

More recently, large language models (LLMs) have been explored as a tool to influence group dynamics and outcomes in a range of settings. One form of such interventions is characterized by static or post-hoc LLM output provided to groups; in these applications, the LLM serves as a tool that provides input to the group in a highly structured format, rather than as an active participant in a group's interactions. Examples of such interventions include Tessler et al.'s [73] LLM-based "Habermas Machine" which iteratively generates consensus statements for social and political

[1]Interquartile range of 316 facilitators in the United States found by searching for the "meeting facilitation" service on Find a Facilitator (https://findafacilitator.com/), as of February 27th, 2025.

issues, as well as approaches that use LLMs to help users phrase their messages to increase the sense of feeling understood in divisive conversations [5] and improve out-group cooperation [13].

As Mao et al. [44] and Do et al. [19] emphasize, transitioning from structured forms of support to an interactive, multi-user chat setting expands the design space from *what* to say, to include the *when* and *who*. To address the added complexity of multi-party interactivity in chat-based settings, frameworks such as those proposed by Mao et al. [44] and Papachristou et al. [54] introduce multi-component LLM systems in which each component tackles a single facet of interaction (e.g. timing, content, recipient selection, intent extraction, and coordination), while the "Observe, Ask, Intervene" framework proposed by Houtti et al. [25] focuses on the process by which such interactive interventions should be delivered.

Within the context of meetings and group discussions, LLMs have been explored as "devil's advocates" that introduce disconfirmatory opinions [11], as coordinators of meeting scheduling processes [54], and as tools that support temporal work across recurrent meetings [75]. We build upon and extend this literature by conducting a pre-registered, large-scale (281 groups), highly interactive (5-6 members per group) experiment studying the role of single LLMs as decision-making facilitators in a setting that is particularly challenging due to its asymmetric and adversarial information landscape.

## 3 Methods

### 3.1 Participants

In this pre-registered experiment (AsPredicted #192061), 1,475 participants were recruited from Prolific and randomly assigned into groups of five decision makers (with a sixth participant added as facilitator in one treatment condition). All participants resided in the United States and the mean age was 39 years (SD = 12.5), with 54% identifying as male and 45% as female (the remainder declined to answer).

Participants were compensated a flat amount of $3.75 for full completion of the study, and the median hourly compensation rate was $14.40/hr. The study was approved by the Microsoft Research Institutional Review Board (MSR IRB; Record #10883), and all participants provided explicit consent to participate in this study. For recruitment materials, please see Appendix A.

### 3.2 Experiment Design & Implementation

*3.2.1 Facilitation.* The focal manipulation of this study was the availability of facilitative support to assist groups during deliberation. Each group was randomly assigned to receive one of four facilitation treatments:

(1) **No facilitation** ("None", N=70 groups), in which a group completed the task without any facilitative support.
(2) **A one-time message** ("Message", N=71 groups), in which the group was shown a message saying "People may have different information about what is being discussed in this meeting, so encourage everyone to share all of the relevant information they have." This notice was shown at the beginning of the group's deliberation period, and gradually disappeared once the discussion began. As a minimal pre-deliberation prompt, this intervention mirrors active control conditions previously studied by Stasser et al. [66, 70].
(3) **Human facilitation** ("Human", N=70 groups), in which a sixth human participant was added to the group to act as a facilitator. This participant did not receive any private task information, and therefore operated as a neutral, external lay facilitator focused on structuring the group's discussion rather than contributing substantively to the decision.
(4) **LLM facilitation** ("LLM", N=70 groups), in which an LLM (GPT-4o) acted as a facilitator.

Neither the human nor LLM facilitator was provided with information relevant to any of the options in the hidden profile task, and therefore both functioned as external participants whose role was limited to structuring the group's deliberation rather than contributing task-relevant content. Both facilitators were presented with the same core prompt to "first, make sure that everyone is heard from and shares what they know and, second, act as a scoreboard and keep track of pros and cons", and were also given the reminder presented in the "Message" facilitation condition (in the "Human" and "LLM" conditions, the message was only provided to the facilitator, not the participants). If the group was assigned to have an LLM facilitator, the LLM sent a message to the group every 90 seconds, in addition to responding every time it was directly addressed by the group; each request sent to the LLM includes the full timestamped transcript preceding the request, as well as general information about meeting objectives, attendees, and time elapsed/remaining. For implementation details of each facilitative condition, including screenshots, please see Appendix C. The implementation of the LLM facilitator and its prompt structure are further detailed in Appendix D.

*3.2.2 The Hidden Profile Task.* In this experiment's hidden profile task, groups were tasked with selecting one of three fictional cities—Eldoron, Myloria, and Cragnio—to host an international sporting event akin to the Olympics. Each city was described by 10 different facts, and facts had five categories: weather, tourism, transportation, community, and a miscellaneous category for neutral information that did not plausibly influence the decision to host a sporting event. Each fact about a city had a positive (e.g., "friendly and welcoming residents"), negative (e.g., "higher than average humidity"), or neutral valence (e.g., "considered a hub for biotech in the region"), validated through a pre-experiment survey. The structure of the task, including group size, number of options, and information architecture, was modeled after the "Grogan Air" hidden profile task [4], with the content modified for the international sporting event context. For details of the hidden profile task design and valence measurement survey, please see Appendix E.

Under full information and a unit-weight linear model (UWLM) of utility [22] which assigns a score of 1 to positive facts, 0 to neutral facts, and -1 to negative facts, Eldoron is the most favorable option with a total score of 6, followed by Myloria with a score of 2, and Cragnio with a score of 0. However, each group member only had access to a subset of facts; facts could either be public (available to all group members), shared (available to multiple members), or private (available to only one member). To obscure the utility of Eldoron and create the "hidden" profile, the polarity (positive/negative/neutral) and visibility (public/shared/private) of the facts were manipulated such that Myloria was the most appealing option in each individual information set, as detailed in Table 1. For the full list of facts and the distribution of information among the participants, please see Figure 12 in the Appendix.

*3.2.3 Experimental flow.* During onboarding, participants are informed of the purpose and duration of the meeting, the size of their group, and the availability and type of facilitator supporting their group, if they have been assigned one. After reviewing the instructions and completing a brief comprehension check, each group proceeds through three stages of the experiment, illustrated in Figure 1, using the chat interface illustrated in Figure 2:

(1) **Icebreaker:** Groups begin by spending 1 minute introducing themselves to each other, using the same chat interface they will use for the main decision making task. During this stage, participants gain familiarity with the interface, allowing them to immediately begin the decision-making discussion in the next stage. Participants are not shown any task-relevant information during this stage. While human facilitators may participate in the icebreaker stage, the LLM facilitator does not.

| Information Set | Eldoron | Myloria | Cragnio |
|---|---|---|---|
| Full | **6** | 2 | 0 |
| Blue | 2 | **3** | 1 |
| Green | 1 | **3** | 2 |
| Orange | 2 | **3** | 2 |
| Pink | 1 | **3** | 2 |
| Red | 1 | **3** | 2 |

Table 1. **Utility scores for each city (Eldoron, Myloria, Cragnio) by information set.** Under full information, Eldoron has the highest utility, whereas individual participant information sets (Blue, Green, Orange, Pink, Red) were manipulated to create a hidden profile favoring Myloria as the most attractive city. Utility scores reflect the sum of valence values assigned to positive (+1), neutral (0), and negative (-1) facts based on a unit-weight linear model (UWLM).

(2) **Decision Making Discussion:** After the icebreaker stage, each participant is shown their personal information set, and the group is given 10 minutes to discuss the task. Groups must stay for the entirety of the 10 minutes. During this stage, facilitators in the "Human" and "LLM" conditions communicate with their group through the same chat window the group is using for their discussion. Instead of receiving task-specific information, as is shown to the five "decision making" participants, the human facilitator's interface shows the facilitator's instructions.
(3) **Post-task Survey:** After the decision making discussion, all participants (including human facilitators) complete an exit survey individually. They are asked to report the decision made by the group, as well as their subjective evaluations of the task, their group, and their facilitator (if they were assigned one). For screenshots of the introduction and exit survey, please refer to Appendix G.

Within each group, participants are randomly assigned colors ("Blue", "Pink", "Green", "Orange", "Red") as nicknames[2]; across all groups, each color is associated with an information set (which includes all public information, and any shared or private information assigned to them), presented in "personal briefs". Through the interface shown in Figure 2, participants can view their personal brief on the left side of the screen, and communicate through a chat window on the right side of the screen. The interface is configured to disable direct copying and pasting of the information presented in the personal brief, preserving the limitations on information sharing that make it necessary to choose which information to share, and preventing participants from easily sharing all available facts or copy-pasting the content into an external LLM to be summarized.

### 3.3 Outcome and Process Measures

To answer RQ1, our pre-registered primary outcome is the volume of information covered within each group, operationalized as the total number of distinct facts mentioned during the group's discussion, ranging from 0 to a maximum of 30 total facts. To quantify this outcome, each message sent by participants is annotated by GPT-4o to identify all directly or indirectly mentioned facts (for further details, please refer to Appendix F). Our focus on information coverage is motivated by the fact that information aggregation is a necessary (albeit insufficient, as we discuss in Section 4.3)

[2]We selected colors as nicknames for ease of use by participants and to avoid stronger implicit associations (e.g., by roles or human names), while acknowledging the possibility that color labels themselves could still influence the behavior of both participants and the LLM.

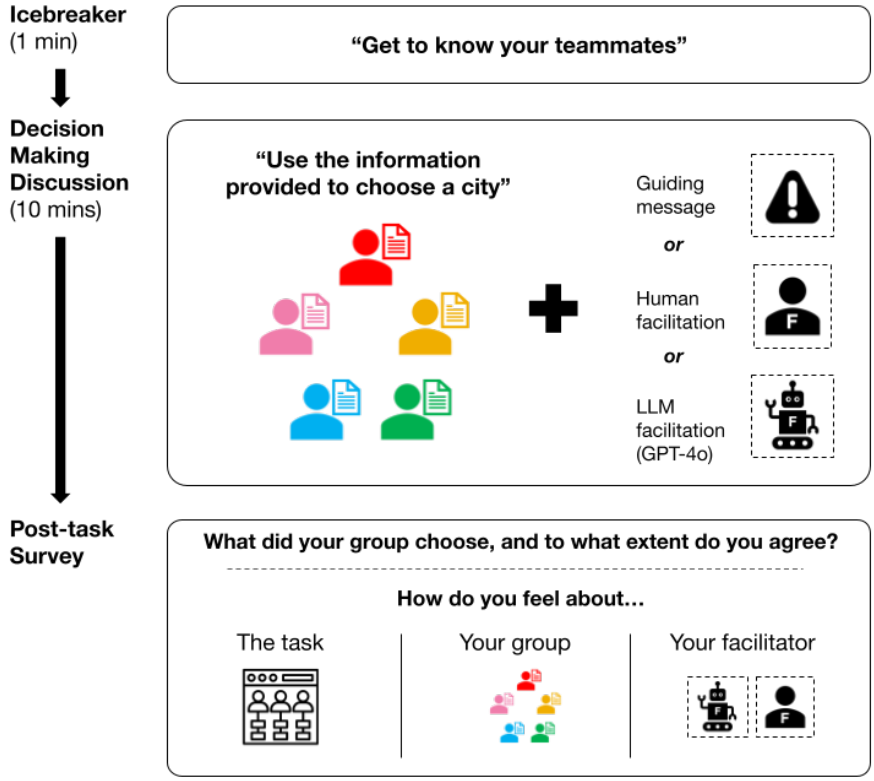



Fig. 1. **Experimental design.** After reading an overview of the task and completing a brief comprehension check, participants were divided into groups of 5 decision makers (and a 6th participant member acting as facilitator, if assigned one) to proceed through the 3 stages of the experiment. (A) First, participants completed a 1-minute icebreaker exercise in which they met their group using the same discussion interface they would use for the main task. (B) After the icebreaker, groups had 10 minutes to complete the hidden profile task, and were randomized to either a control condition with no facilitative support, or one of the three facilitative treatments: "Message", "Human", or "LLM". (C) After the main discussion, each participant completed an exit survey individually, in which they reported the group's decision and their subjective attitudes towards the task, their group, and the facilitator, if they were assigned one.

condition for effective decision-making not only in our instantiation of the HPT, but across a wide range of HPT designs and group performance evaluation schemes [48].

To investigate RQ2, which studies the impact of our facilitative treatments on the discussion dynamics, we measure several process outcomes. Discussion density, defined as the ratio of facts surfaced to the number of messages sent in the discussion (excluding facilitator messages), allows us to differentiate between an increase in fact sharing driven by longer conversations, as opposed to an increase driven by a different discussion focus. To understand the impact of facilitation on option exploration, we measure the Gini coefficient of the facts surfaced for each of the three options within a group, where a lower Gini coefficient implies more even fact coverage of the options.

In settings of asymmetric information, such as the hidden profile task, individual engagement is critical; if a group member does not contribute to the discussion, key information may be locked out of the decision making process. Within each group, we measure the upper and lower bounds of

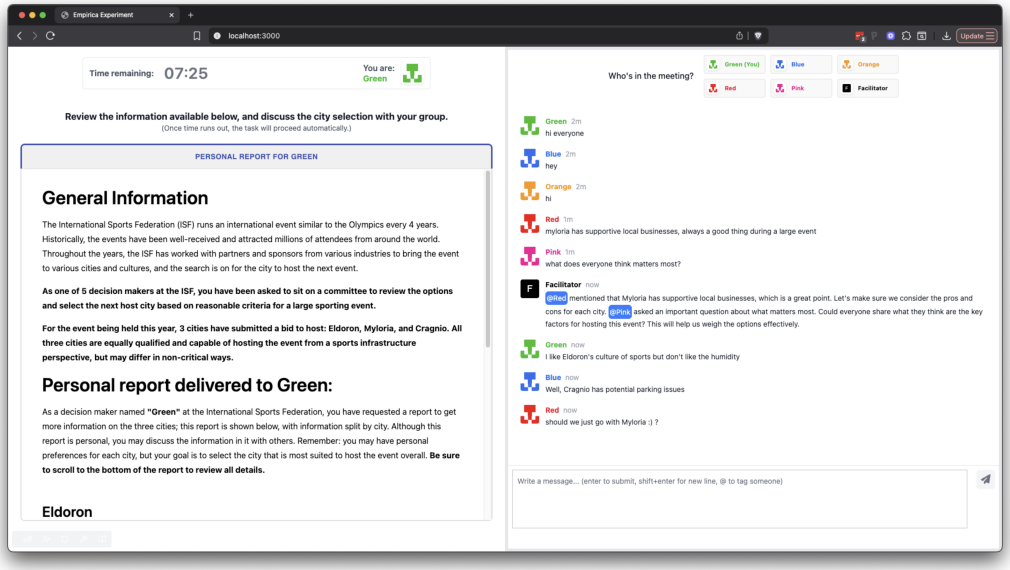



Fig. 2. **Experiment interface.** On the left half of the screen, participants could review the general instructions and the unique set of information available to them about each of the 3 cities. On the right half, the group communicated through a shared chat window; participants could use "@" tags to directly address others, including the facilitator.

individual engagement by measuring the minimum and maximum number of new facts added to the discussion by an individual group member. A distinct but related measure is whether a group achieves full participation, which we define as whether every member has surfaced at least one new fact not previously mentioned in the discussion.

Finally, we define measures describing the information surfaced, and the decision made by the group. Within each group, we define an option's "normative score" as the score that would be given to an option on the basis of the facts surfaced by the group, assuming fact polarity were perfectly perceived by the groups as intended, and use these scores to find the decision the group would make under a unit-weight linear model of decision making. We contrast this normative decision to the actual decision made by the group, defined as the option reported to be the group's decision by a majority of its members, excluding human facilitators. All 281 groups yielded a clear majority, with 19 groups having a 4-1 split of votes, 8 groups having a 3-2 split, and the remaining groups making unanimous decisions.

RQ3 is motivated by the fact that, while a given treatment may lead to outcome or process improvements, such improvements could possibly come at a cost to the subjective experience of decision makers. To study this, we measure individual participants' subjective evaluation of three dimensions–the task, their group, and their facilitator, if they had one–in a post-task survey. Subjective task experience is measured through the NASA Task Load Index [24], a 6-item instrument capturing perceived workload across mental, physical, and temporal demand, as well as performance, effort, and frustration.

Following the TLX, the subjective evaluation of the group begins with a general, free-text question ("How did you feel about your group in the task?"), followed by four 5-point Likert scale questions intended to capture commonly-studied facets of group decision-making and discussion

quality, namely participation and inclusion ("I feel that I was able to contribute to the group discussion"), agency ("I feel that I was able to influence the group's decision"), productivity ("The group's discussion felt productive"), and structure ("The group had a structured way of collecting and summarizing information to reach a decision"). The final item of the group evaluation intended to measure group cohesion, by asking a binary question of whether the participant would hypothetically prefer to repeat the task with the same group or a new, random group.

Lastly, following the structure of the group evaluation, the facilitator evaluation begins with a similarly phrased free-text question about the facilitator, followed by four 5-point Likert scale questions to evaluate the performance of the facilitator, along dimensions of information elicitation ("The facilitator encouraged the group to share information"), information synthesis ("The facilitator helped the group summarize the information shared to reach a decision"), task focus ("The facilitator was able to keep the group focused on the task"), and distraction and process interruption ("The facilitator was distracting"). The facilitator evaluation concludes by asking "You completed the task with [no, a human, an AI] facilitator. If you were to repeat the task, which type of facilitation would you prefer?", with the available options being "None", "Human", and "LLM".

For analysis, TLX and Likert-based responses are aggregated to the group level by taking the mean of individual non-facilitator responses within each group, aligning the unit of analysis with that of treatment assignment. Subjective survey completion rates exceeded 98% for all TLX items, 95% for all group and facilitator Likert and multiple-choice items, and 68% for optional free-text responses.

### 3.4 Statistical Analyses and Data Processing

As pre-registered (AsPredicted #192061), we investigate RQ1 through confirmatory, pairwise, one-sided, independent-samples t-tests of the difference in mean fact coverage between experimental treatments, with sample size chosen based on pilot study data to achieve 70% power in detecting a difference of 2 facts shared at a significance level of 5%. In addition to the p-values for these pre-registered analyses, we also report adjusted p-values using the Benjamini-Hochberg procedure [7] as a conservative robustness check. All other analyses in this work are exploratory and reported with unadjusted p-values, and all subsequent t-tests are two-sided. In our exploratory analyses, we employ one-way analysis of variance (ANOVA) and Pearson's chi-squared test to test for differences between treatments in continuous and binary outcomes, respectively.

To probe the relative weight of facts in a group's decision, we estimated a multinomial logit model of option selection as a function of normative UWLM scores. Let $Y_g$ denote the option selected by group $g$, and let $S_{g,E}$, $S_{g,M}$, and $S_{g,C}$ denote the surfaced normative scores for Eldoron, Myloria, and Cragnio, respectively. For each non-baseline option $j \in \{\text{Eldoron}, \text{Cragnio}\}$, we specify:

$$\log \frac{\Pr(Y_g = j)}{\Pr(Y_g = \text{Myloria})} = \beta_{jE} S_{g,E} + \beta_{jM} S_{g,M} + \beta_{jC} S_{g,C}$$

We set Myloria as the reference category, so each coefficient $\beta_{jk}$ captures the association between the surfaced normative score of option $k$ and the log-odds of choosing option $j$ (relative to Myloria), holding other scores constant. Importantly, because normative scores are themselves shaped by the experimental treatment, these estimates should be interpreted descriptively: they capture how surfaced information and decisions co-vary across groups, rather than isolating the causal effect of additional information holding the decision process fixed. As such, the coefficients reflect a combination of how groups weight information and how treatments jointly influence both information sharing and decision-making.

*3.4.1 Data processing.* Given the sensitivity of the decision-making task to private information that may be lost to a group if a participant leaves the study, our pre-registered inclusion criteria are that each participant sends at least one message, the group collectively sends at least 10 messages, and that all participants submit the task completion code (displayed at the end of the study) on Prolific. Groups were sampled from Prolific until the pre-registered count of 70 valid groups per experimental condition was met. During data collection, 7 groups in each of the "None", "Message", and "LLM" conditions and 15 groups in the "Human" condition failed to satisfy the inclusion criteria; one additional group was sampled in the Message condition as a consequence of ensuring equal-probability assignment across treatments. While the difference in exclusion rates is not statistically significant ($\chi^2(3, N = 317) = 4.56$, p = 0.21), the higher exclusion rate in the "Human" condition may reflect its larger group size (6 vs. 5), which mechanically increases the probability that at least one participant causes a group's exclusion. For more details on pre-registration, please refer to Appendix B.

## 4 Results

### 4.1 LLM facilitation increased information sharing in group discussions under asymmetric information

As illustrated in Figure 3A, we find that neither the one-time message treatment ($\Delta$ = 1.32, t(139) = 1.52, p = 0.066, Benjamini-Hochberg adjusted p = 0.082) nor human facilitation ($\Delta$ = 0.43, t(138) = 0.51, unadjusted and B-H adjusted p = 0.31) significantly increased the average number of facts shared relative to no facilitation. However, we find that LLM facilitation increased information sharing relative to all other conditions: $\Delta$ = 2.9, t(138) = 3.58, $p < 0.001$ , B-H adjusted p = 0.001 for "None"; $\Delta$ = 1.58, t(139) = 2.1, p = 0.021, B-H adjusted p = 0.035 for "Message"; and $\Delta$ = 2.47, t(138) = 3.38, $p < 0.001$ , B-H adjusted p = 0.001 for "Human". To contextualize the observed increase in information coverage with LLM facilitation relative to no facilitation ("LLM" vs. "None"), the effect size (Cohen's *d*) is 0.61, indicating a moderate-to-large practical effect of LLM facilitation on information coverage.

### 4.2 Process-level effects of LLM facilitation on group discussion and information sharing

Despite LLM facilitators eliciting more facts from groups, we detect no significant difference in the average number of overall messages sent by the group members (excluding the facilitators) between treatments (F(3,277) = 0.46, p = 0.71), and find that LLM-facilitated discussions were more information dense than unfacilitated discussions on average, where density is defined as the ratio of covered facts to the number of non-facilitator messages in the discussion ($\Delta$ = 0.067, 95% CI [0.025, 0.109], t(138) = 3.17, p = 0.002; see Figure 3B); together, these findings indicate that the increase in fact coverage is not solely a mechanical consequence of longer discussions. Moreover, the information covered in LLM-facilitated groups may be more evenly distributed among the three candidate cities than in groups with no facilitation, as suggested by the decrease in the average Gini coefficient of covered facts for each city ($\Delta$ = -0.05, 95% CI [-0.091, -0.009], t(138) = -2.42, p = 0.017; Figure 3C), hinting that LLM facilitation may encourage groups to more holistically evaluate the decision at hand.

By the volume and length of their messages, human facilitators appeared to adopt a more conversational style than their LLM counterparts. Human facilitators communicated more frequently, sending 9 messages to their group on average, while the LLM facilitator sent an average of 7 messages (by design, as detailed in Appendix D; t(138) = 2.74, p = 0.007), yet the average human facilitator message was much shorter with a length of 9 words, compared to the average LLM

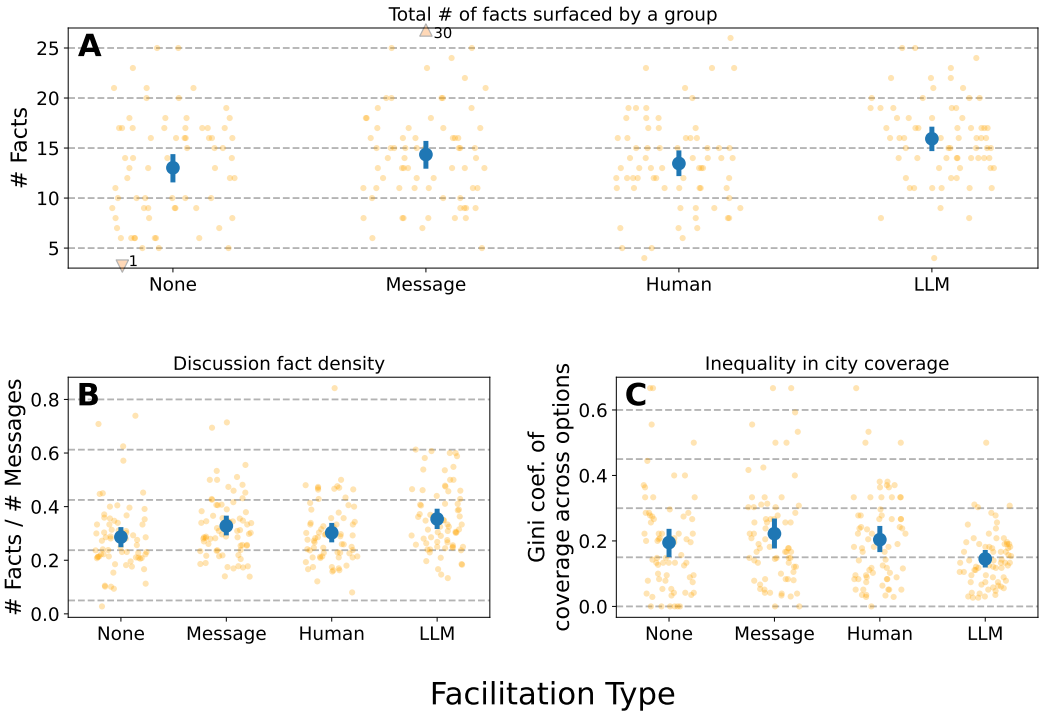



Fig. 3. **LLM facilitation increased fact sharing volume, density, and breadth during discussion.** (A) shows the total number of facts surfaced by a group. (B) shows the density of facts in a conversation (# facts surfaced / # messages, excluding facilitator messages). (C) shows the Gini coefficient of each group's covered information across each of the three options. Error bars indicate 95% confidence intervals, and may be smaller than the marker. Scatter markers each represent a single group. For readability, border markers in panel (A) indicate points out of the y-axis bounds.

facilitator message containing 79 words. However, by evaluating the *content* of the facilitator messages, we find that one potential mechanism driving the increase in information sharing is that of direct engagement with group members to participate in the discussion (e.g., "Red, we haven't heard from you…"). On average, 25% of a human facilitator's messages in a discussion directly addressed a group member, while LLM facilitators did so in 68% of the messages they shared within a single discussion; the LLM facilitator directly addressed a group member at least once in every group it interacted with, yet only 77% of human facilitators ever did so.

Consistent with these differences in facilitator behavior, LLM facilitation increased the minimum number of new facts (i.e., not previously covered in the group's discussion) shared by any member of the group in comparison to no facilitation ($\Delta = 0.41$, 95% CI [0.145, 0.684], $t(138) = 3.04$, $p = 0.003$), but not the maximum number of new facts shared by an individual (as illustrated in Figures 4A and 4B, respectively). Such an increase can arise through two mechanisms: an intensive margin effect, in which the least active contributors share more facts, and an extensive margin effect, in which otherwise non-participating members (those contributing no new facts) contribute at least one fact. When restricting to groups in which all members contribute at least one fact, we do not observe differences between treatments in the minimum number of facts shared ($F(3,277) = 1.37$, $p = 0.25$), suggesting limited differences in minimum engagement among already-contributing members.

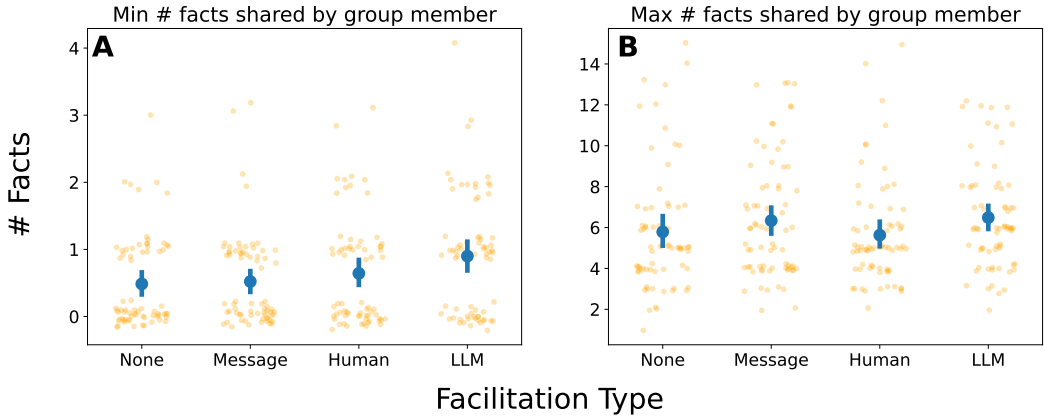



Fig. 4. **LLM facilitation increased the minimum individual contribution to fact aggregation.** (A) and (B) show the minimum and maximum number of distinct facts surfaced by any individual group member, respectively. Error bars indicate 95% confidence intervals, and may be smaller than the marker. Scatter markers each represent a single group. Fact counts are discrete, but vertical jitter is added for visibility.

However, while only 39%, 44%, and 47% of groups achieved contribution from all members in the "None", "Message", and "Human" conditions respectively, 61% of LLM-facilitated groups had all members contribute at least one new fact ($\chi^2(3, N = 281) = 8.10$, p = 0.044), providing suggestive evidence that gains in information sharing may arise from the extensive margin.

### 4.3 Information aggregation did not translate into UWLM-consistent decisions

Defining a group's collective choice as the option chosen by a majority, Eldoron (the optimal choice) was chosen by 31% of unfacilitated groups, 21% of groups presented with the one-time message, 30% of groups facilitated by a human, and 23% of groups facilitated by an LLM; however, these proportions are not statistically distinguishable from one another ($\chi^2(3, N = 281) = 2.85$, p = 0.42). While unaffected by the facilitative treatments, these proportions of optimal outcomes are in line with those reported for the hidden profile task upon which ours is based [4]. If groups strictly followed UWLM to make their decision, Eldoron would have been chosen by 60% of unfacilitated groups, 48% of groups presented with the one-time message, 47% of groups facilitated by a human, and 63% of groups facilitated by an LLM, yet these differences between treatments are also indistinguishable ($\chi^2(3, N = 281) = 5.6$, p = 0.13). Similarly, the proportion of groups who would have chosen Eldoron under UWLM but did not actually do so does not vary significantly across treatments ($\chi^2(3, N = 281) = 7.52$, p = 0.057).

Beyond the stringent assumptions of strict UWLM decision-making, differential weighting of information across options may partially explain deviations from UWLM decisions. As estimated by a multinomial logit model with Myloria (the "bait") as the baseline option, the odds of choosing Eldoron (the optimal option) increased as its score increased ($\beta = 0.395$, 95% CI [0.245, 0.545], $p < 0.001$) and decreased with increasing Myloria score ($\beta = -0.734$, 95% CI [-0.919, -0.548], $p < 0.001$), while the score of Cragnio did not influence this contrast ($\beta = 0.054$, 95% CI [-0.170, 0.279], p = 0.64). Similarly, the odds of choosing Cragnio increased as its score increased ($\beta = 0.417$, 95% CI [0.136, 0.698], p = 0.004) and decreased with increasing Myloria score ($\beta = -0.649$, 95% CI [-0.843, -0.455], $p < 0.001$), while the score of Eldoron did not influence this contrast ($\beta = -0.043$, 95% CI [-0.236, 0.150], p = 0.66). In both comparisons, the magnitude of the effect of Myloria's score exceeded that

of the alternative city by 86% (vs. Eldoron) and 56% (vs. Cragnio), suggesting that group decisions are more strongly associated with information about Myloria than with information about the other options.

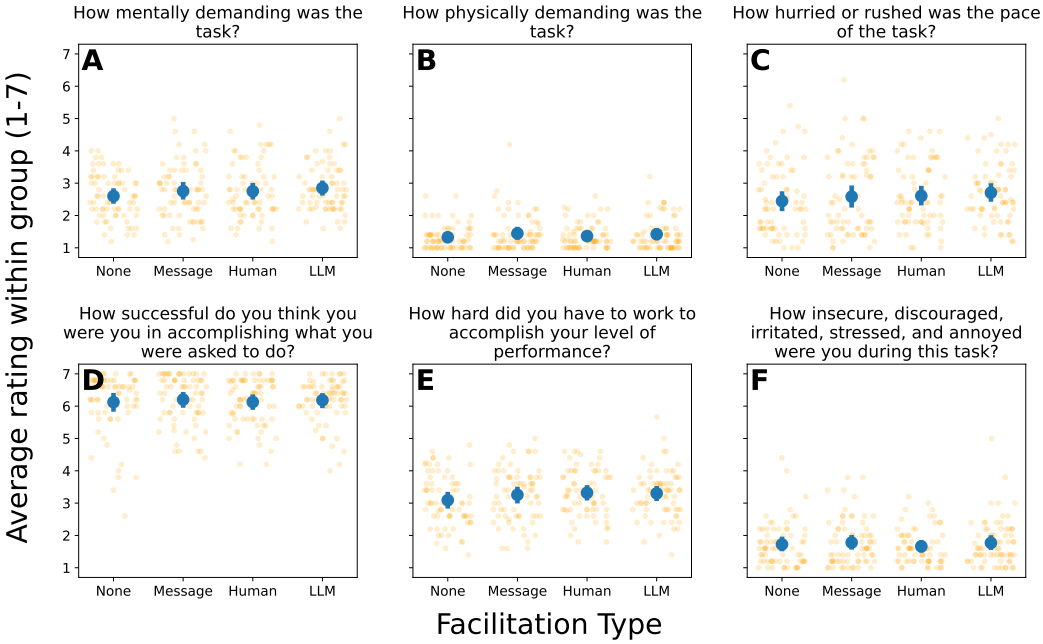



Fig. 5. **NASA Task Load Index (TLX).** The TLX measures perceptions of a workload along six dimensions: mental demand (A), physical demand (B), temporal demand (C), performance (D), effort (E), and frustration (F). Participants individually answered each question on a discrete scale of 1 ("very little") to 7 ("very much"). Responses were averaged within each group, and each scatter marker represents a single group. Error bars indicate 95% confidence intervals, and may be smaller than the marker.

### 4.4 Perceptions of the task, group, and facilitator

As measured by the NASA TLX, group-level mean responses (averaging across members within each group) indicate that sentiment towards the task within groups was generally positive and indistinguishable across experimental conditions, with participants reporting generally low levels of mental, physical, and temporal demand, as well as effort and irritation (Figure 5A-C, E-F respectively), and high levels of perceived success in the task assigned to them (Figure 5D). Similarly, groups reported high levels of perceived personal contribution, individual influence over the decision, productivity, and structure (Figure 6A-D, respectively), with no statistically meaningful differences between conditions. Group cohesion—the mean proportion of group members who said they would want to repeat the task with the same group, as opposed to a random group—ranged from 77% in the "Message" condition to 80% in the "Human" condition, with insignificant differences between conditions ($F(3,277) = 0.21$, $p = 0.89$). Taken together, these results suggest that participants experienced the task and group interaction similarly across conditions, despite differences in facilitation.

In both the "Human" and "LLM" conditions, facilitators were rated positively for their ability to elicit information, summarize the discussion, and keep the group on task (Figure 6E-G, respectively).

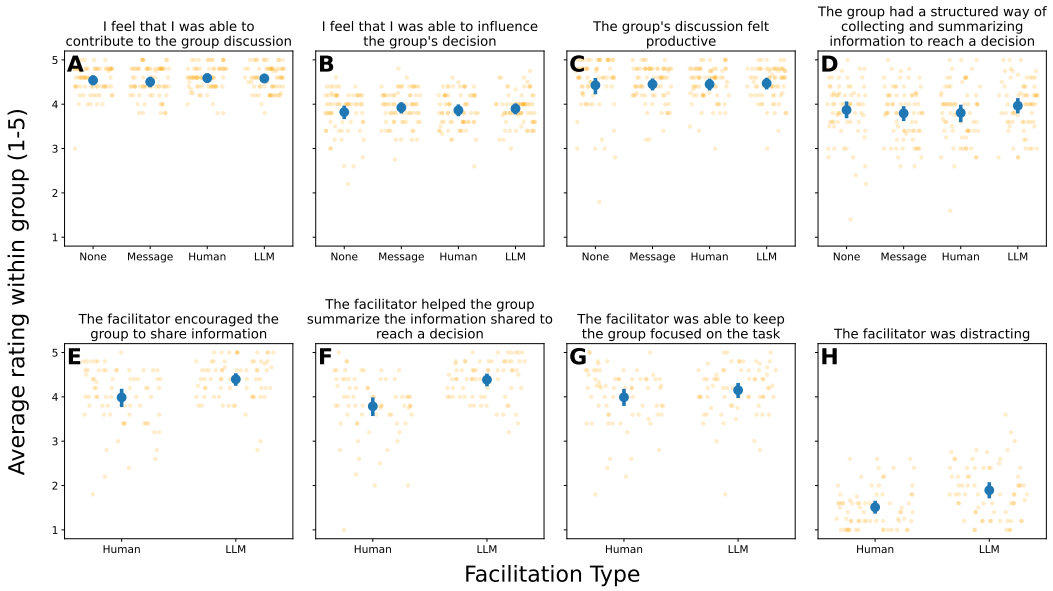



Fig. 6. **Subjective evaluations of group dynamics and facilitator effectiveness.** Participants were individually asked to evaluate their group (A-D) and facilitator (E-H) by rating their agreement with each statement on a discrete scale of 1 ("strongly disagree") to 5 ("strongly agree"). Responses were averaged within each group, and each scatter marker represents a single group. Error bars indicate 95% confidence intervals, and may be smaller than the marker.

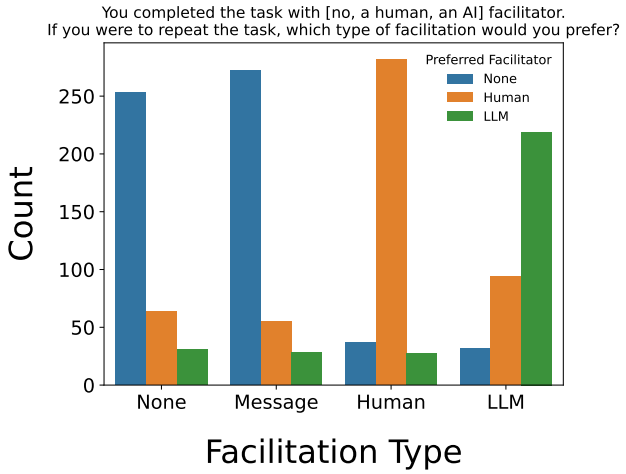



Fig. 7. **Participant preferences for future facilitation.** Participants individually reported their preferred facilitation type if they were to hypothetically repeat the task, grouped in the figure according to their initial facilitation experience (None, Message, Human, or LLM).

The mean group rating of LLM facilitators was higher for both elicitation ($\Delta = 0.41$, 95% CI [0.22, 0.60], $t(138) = 4.18$, $p < 0.001$) and summarization ($\Delta = 0.59$, 95% CI [0.39, 0.79], $t(138) = 5.87$, $p < 0.001$), which was echoed in free-text feedback: 102 participants (across 57 of 70 LLM-facilitated groups) commented positively on elicitation and/or summarization. For example, one participant noted that the facilitator was "good for summarizing the data and prompting participation…", while another described it as helping the group "stay on goal and make sure we were prgoressing [sic] in an organized manner."

Both facilitators were rated as generally non-distracting (Figure 6H), but the LLM was rated as more distracting ($\Delta = 0.38$, 95% CI [0.21, 0.56], $t(138) = 4.29$, $p < 0.001$), with participants describing recurring issues such as repetitive interventions, disrupted flow, and a sense of being rushed. Direct call-outs were particularly divisive. Some participants found them effective for increasing participation (e.g., "did a good job making people who didn't talk much try to talk more," and "it reminded me to start sharing more"), while others experienced them as disruptive ("a little annoying and disruptive to the flow of conversation") or socially uncomfortable ("it made me feel a bit pressured… like it singled me out"). Overall, participants tended to prefer the facilitation style (or lack thereof) they experienced (Figure 7), suggesting that preferences may be shaped by direct experience.

## 5 Discussion

### 5.1 What value did LLM facilitation add to the deliberative process?

*5.1.1 Information aggregation.* LLM facilitation increased the volume and breadth of information shared within groups, without extending the length of discussions, indicating a shift toward more information-dense deliberation. Importantly, these gains appear to arise, at least partially, from the extensive margin: rather than amplifying already active contributors, the LLM increased the minimum level of participation within groups, reducing instances of non-contribution. This distinction is critical in hidden profile tasks, where uniquely held information may reside with a single participant, and failure to elicit that information can prevent the group from accessing the full set of relevant facts.

By increasing the likelihood of contribution and, consequently, the likelihood that privately held information enters the discussion, the LLM facilitator can improve the reliability of information elicitation. At the same time, this intervention targets availability rather than use: while more information is surfaced, how that information is interpreted, weighted, and integrated into a final decision remains unchanged. As such, LLM facilitation addresses one necessary condition for effective collective decision making—information aggregation—without resolving downstream challenges of information integration.

*5.1.2 Decision-making.* Despite increased information sharing, LLM-facilitated groups were no more likely to select the UWLM-optimal option, suggesting a disconnect between information availability and use. In one sense, this is not surprising: the model hinges on the strong assumptions that information is perceived as intended, weighted consistently, aggregated linearly, and tabulated exactly to identify the score-maximizing option. However, this disconnect is not necessarily problematic, as the goal of facilitation is not to induce strictly UWLM-consistent decisions. Rather, UWLM serves as a diagnostic benchmark for assessing whether failures arise from insufficient information aggregation or from how that information is ultimately used. In this light, prior work summarized by Lu et al. [41], reviewing 24 empirical HPT studies, shows that groups are substantially more likely to select the correct option when profiles are manifest (i.e., all members begin with access to the full set of information), suggesting that UWLM-consistent behavior may be plausible for this task structure.

In contrast to the manifest condition, hidden profile settings introduce asymmetries in initial information that can generate early impressions, which in turn shape conversational dynamics, attention, and the interpretation of subsequently shared information. As a result, even when an intervention increases information coverage, groups may not fully recover the benefits of a manifest profile, as newly surfaced information is filtered through these initial biases. This perspective is consistent with prior work arguing that HPT performance is driven more by information *use* than *coverage* [80]; as Dennis puts it, "you can lead a group to information, but you can't make it think" [15]. Notably, while the LLM increased information coverage, this shift remains modest relative to the full information set, leaving groups in a partially hidden regime where early asymmetries can continue to shape how information is interpreted and used.

In our setting, although the information weighting we observe is directionally consistent with UWLM decision-making (higher normative scores increase the odds of selection), we find that information supporting the "bait" option (Myloria) is more strongly associated with selection than information supporting the other options. While our design does not identify the mechanism underlying this asymmetry, this difference may reflect the individual-level preference bias commonly documented in the HPT paradigm [20, 28, 50, 60]. Because each participant initially sees Myloria as the most appealing option in their individual information set, they may remain inclined to focus on Myloria in their *use* of the covered facts, thereby influencing group-level decision making.

*5.1.3 Subjective experience.* Despite meaningfully altering the structure of group discussions, LLM facilitation did not negatively impact participants' experience of the task or their group. Across measures of workload and group dynamics, we observe no meaningful differences between conditions, suggesting that the gains in information sharing were not achieved at the expense of increased cognitive burden or reduced group cohesion.

Relative to the human facilitator, the LLM outperformed in eliciting and summarizing information, despite relying on a relatively simple, domain-agnostic setup. However, participant feedback also highlights trade-offs in how these interventions were experienced, particularly around disruption and direct targeting. Taken together, these findings empirically underscore the need for facilitation systems that address not only what to say, but when and to whom–an emphasis reflected in emerging frameworks for group-AI intervention design [19, 44, 54].

## 5.2 Designing more effective LLM facilitators

*5.2.1 Process intervention beyond information aggregation.* Our findings suggest that while our instance of LLM facilitation improves the availability of information, it may not meaningfully alter how that information is interpreted, weighted, and integrated into a final decision. This gap is, in part, a consequence of our design choices: the LLM intervened at fixed intervals with group-level messages, operating only over the shared transcript and general instructions to elicit and organize information. As a result, the facilitator primarily influenced what information entered the discussion, without shaping how that information was subsequently processed.

One implication is that more targeted process interventions could influence how information is perceived as the discussion unfolds. For example, rather than continuously aggregating information, facilitation could segment deliberation into cycles of aggregation and reflection, enabling groups to periodically synthesize and evaluate what has been shared before introducing additional information. More broadly, the appropriate structure of information aggregation is likely to be context-dependent. While the LLM in our study encouraged more balanced discussion across options, such holistic coverage may be suboptimal in time-constrained settings where dominated alternatives should be quickly eliminated. This suggests value in allowing users to specify or adapt the facilitation process to the demands of the task. Process interventions may also be applied at the decision-making stage.

Introducing structured procedures such as staged evaluation, ranking, or variants of the Delphi method, could support more systematic comparison of alternatives and reduce the influence of early impressions on final decisions.

Importantly, we deliberately constrained the LLM to a process-oriented role, avoiding direct recommendations to preserve the group's sense of ownership over the decision. Future systems that more directly shape the decision-making process (eg., by pointing to critical trade-offs or contributing judgment) may help guide more effective information use, but introduce new risks around bias and over-reliance. The "Deliberative AI" system described by Ma et al. [42] integrates both process and content interventions in a single-user system, and suggests a promising scaffold to be extended to multi-user settings.

*5.2.2 Managing coordination.* The subjective evaluations of the LLM facilitator point to a trade-off between coordination and attention management in group-AI interactive settings. Although participants appreciated the LLM's role in aggregating and organizing information, they also perceived it as more distracting than a human facilitator. Feedback describing excessive or disruptive interjections, and a sense of being rushed, suggests that facilitation competes for participant attention within a shared communication channel.

In addition to the content of an LLM intervention, its form may shape this trade-off. In our study, LLM facilitator messages were substantially longer than those written by human facilitators and were delivered repeatedly throughout a fixed 10-minute discussion. This format introduces both attentional and temporal pressure. Longer facilitator messages occupy more of the visible chat window, potentially crowding out participant messages and interrupting conversational flow. They also require more time to process, potentially contributing to the sense of feeling rushed and reducing the time group members have to share, discuss, and interpret information to make a decision.

Design strategies such as constraining message length, adjusting intervention frequency, or shifting toward more conversational, incremental interactions may reduce this burden. However, such changes may also affect the effectiveness of facilitation, introducing a tension between informativeness and intrusiveness that warrants further study. A complementary approach is to diversify the channels through which LLM facilitators engage participants. Rather than addressing the entire group simultaneously, facilitative systems could also interact with participants individually. This would enable the facilitator to privately deliver feedback (as studied by Do et al. [19]), discuss concerns or motivations, or provide tailored prompts aligned with each participant's level of engagement, potentially reducing distraction in the shared discussion space while preserving the benefits of targeted process intervention.

*5.2.3 Managing emotional tension.* Unlike human facilitators, LLMs may more readily directly address individual group members. Because these call-outs are visible to the entire group, they can increase social pressure to participate, potentially reducing social loafing [27]. However, participant feedback suggests that these interventions can also induce discomfort. Prior work by Houtti et al. [25] suggests that when interacting with AI facilitators, group members may respond to such discomfort by rationalizing away or dismissing critical feedback, potentially undermining the effectiveness of the intervention.

This suggests another dynamic trade-off: while direct interventions may boost immediate participation, they may also erode receptivity to future interventions if they generate sufficient discomfort. As a result, the effectiveness of facilitation may depend not only on whether interventions occur, but on how they shape participants' willingness to engage with subsequent guidance, necessitating careful calibration of when and how individuals are targeted.

### 5.3 Limitations

The effects of LLM facilitation on information sharing, decision-making, and subjective experience are shaped by several design choices in our experimental setting. Beyond the implementation of the LLM facilitator and its interface, as discussed in Section 5.2, contextual factors such as time pressure [28, 29], group size [14], and communication modality (chat, video, audio, face-to-face) [30, 34, 63] can significantly alter the attentional landscape of the task, in turn influencing the efficacy of LLM facilitation.

Holding the design of the LLM facilitator and task context constant, the characteristics of the hidden profile task itself may further shape these effects. For example, an HPT's difficulty [8, 59] and total information load [68] may moderate both a facilitator's ability to elicit information and the decision-making value of any additional information that is surfaced. More fundamentally, the HPT in our study is a collaborative task in a neutral domain; effectively managing divergent viewpoints and incentives in contentious [43] or competitive [74, 77] settings may require more sophisticated facilitation strategies than those considered here.

In sum, the range of factors influencing LLM-facilitated group interactions necessitate a systematic and integrative approach to future research [3]. By carefully examining these dimensions jointly, the field can advance toward more robust and generalizable insights into effective group-AI facilitation.

### 5.4 Open data and open-source software to support group-AI interaction research

Progress in understanding group-AI interaction is constrained by limited access to shared datasets and experimental infrastructure. While paradigms such as the HPT provide a canonical setting for studying how groups aggregate distributed information, results are often difficult to compare or build upon due to differences in implementation and the lack of accessible, process-level data. Advancing toward a cumulative science of group-AI interaction requires both richly instrumented datasets and reusable experimental platforms.

To this end, we release our dataset capturing 14,343 messages shared by 1,475 participants across 281 groups, providing a rich resource for researchers studying facilitative technologies and group decision-making. Alongside the transcripts, we provide all LLM requests and responses, along with associated metadata and LLM-generated rationales, enabling researchers to conveniently perform post-hoc analyses or simulate alternative facilitation prompts.

In addition to the dataset, we introduce the Group-AI Interaction Laboratory (GRAIL), our open-source experimental platform based on the Empirica framework [2], joining a growing library of group-AI deliberation research tools such as the Deliberate Lab platform [57]. GRAIL is designed to address key practical challenges in conducting group-based behavioral experiments, including coordinating synchronous participation, managing real-time interaction, and integrating LLMs into shared communication environments. By providing a flexible interface for specifying task structure, participant-specific stimuli, and LLM intervention logic, GRAIL allows researchers to prototype and deploy new experimental designs without re-implementing the underlying system.

## 6 Conclusion

Our study demonstrates the practical capacity of LLMs to *actively* support group decision making discussions by directly engaging group members and prompting participation as a discussion evolves, as well as effectively summarizing rapid, information-dense communication. By raising the minimum level of individual task engagement within groups, the LLM facilitator increased both the breadth of information at a group's disposal, and the likelihood that *all* group members

contributed information to the decision—an outcome that is critical in settings with asymmetric information, where a single member's private knowledge may be pivotal.

Decision-making, whether performed individually or collaboratively, is a complex, multi-stage process. Although the increased information sharing observed did not translate directly into better final decisions in our experiment, this outcome highlights the inherent complexity and context dependence of group decision-making. It also underscores the potential value of integrating computational support that focuses distinctly on each stage of the deliberative process, from initial information gathering through to evaluation and consensus formation.

By encouraging broader participation within groups, the LLM facilitator helps groups consider a wider range of perspectives, particularly from group members who might otherwise remain passive. In practical scenarios analogous to hidden profile tasks—such as hiring committees—this form of facilitation may play an important role in mitigating biases through promoting conversational equity. Moreover, continued development of computational facilitation technologies could expand access to these organizational techniques, making them available to communities unable to afford professional facilitation services.

Our work builds upon an extensive body of research within the CSCW and broader HCI communities studying the effectiveness of group decision support systems, which have historically incorporated a wide range of computational methods. The novel affordances of LLMs present an opportunity to increase the fidelity and adaptivity of these systems, and it is our hope that the robust empirical evidence of this potential that we have presented here inspires future work in this direction, and that our open-source experimental platform enables researchers to systematically explore and extend these potential benefits.

## 7 Data, code, and institutional review

Experimental data and analysis code are available on the study's Open Science Framework repository (https://doi.org/10.17605/OSF.IO/ERVNB), and the GRAIL platform is available through its GitHub repository (https://github.com/microsoft/group_ai_lab). The study was approved by the Microsoft Research Institutional Review Board (MSR IRB; Record #10883), and all participants provided explicit consent to participate.

## Acknowledgments

We are grateful to Daniel Ames for his helpful insight into the hidden profile task paradigm.

## References

[1] Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, and Duncan J Watts. 2021. Task complexity moderates group synergy. *Proc. Natl. Acad. Sci. U. S. A.* 118, 36 (7 Sept. 2021).

[2] Abdullah Almaatouq, Joshua Becker, James P Houghton, Nicolas Paton, Duncan J Watts, and Mark E Whiting. 2021. Empirica: a virtual lab for high-throughput macro-level experiments. *Behav. Res. Methods* 53, 5 (Oct. 2021), 2158–2171.

[3] Abdullah Almaatouq, Thomas L Griffiths, Jordan W Suchow, Mark E Whiting, James Evans, and Duncan J Watts. 2022. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* 47 (21 Dec. 2022), e33.

[4] Daniel Ames. [n. d.]. Grogan Air: Team Decision Exercise. https://caseworks.business.columbia.edu/caseworks/grogan-air-team-decision-exercise. Accessed: 2025-3-3.

[5] 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. *Proc. Natl. Acad. Sci. U. S. A.* 120, 41 (10 Oct. 2023), e2311627120.

[6] Robert S Baron. 2005. So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. In *Advances in Experimental Social Psychology*. Advances in experimental social psychology, Vol. 37. Elsevier, 219–253.

[7] Y Benjamini and Y 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.

[8] Felix C Brodbeck, Rudolf Kerschreiter, Andreas Mojzisch, Dieter Frey, and Stefan Schulz-Hardt. 2002. The dissemination of critical, unshared information in decision-making groups: the effects of pre-discussion dissent. *Eur. J. Soc. Psychol.* 32, 1 (Jan. 2002), 35–56.

[9] Felix C Brodbeck, Rudolf Kerschreiter, Andreas Mojzisch, and Stefan Schulz-Hardt. 2007. Group decision making under conditions of distributed knowledge: The information asymmetries model. *Acad. Manage. Rev.* 32, 2 (April 2007), 459–479.

[10] Xinyue Chen, Lev Tankelevitch, Rishi Vanukuru, Ava Elizabeth Scott, Payod Panda, and Sean Rintel. 2025. Are we on track? AI-assisted active and passive goal reflection during meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–22.

[11] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-assisted group decision making through LLM-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, Vol. 1. ACM, New York, NY, USA.

[12] Nicholas A Christakis and James H Fowler. 2008. The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* 358, 21 (22 May 2008), 2249–2258.

[13] Elijah L Claggett, Robert E Kraut, and Hirokazu Shirado. 2025. Relational AI: Facilitating intergroup cooperation with socially aware conversational support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–22.

[14] Michael G Cruz, Franklin J Boster, and Jóse I Rodríguez. 1997. The impact of group size and proportion of shared information on the exchange and integration of information in groups. *Communic. Res.* 24, 3 (June 1997), 291–313.

[15] Alan R Dennis. 1996. Information exchange and use in group decision making: You can lead a group to information, but you can't make it think. *MIS Q* 20, 4 (Dec. 1996), 433.

[16] Gerardine DeSanctis and R Brent Gallupe. 1987. A foundation for the study of group decision support systems. *Manage. Sci.* 33, 5 (May 1987), 589–609.

[17] Stephan Diederich, University of Göttingen, Germany, Alfred Benedikt Brendel, Stefan Morana, Lutz Kolbe, TU Dresden, Germany, Saarland University, Germany, and University of Göttingen, Germany. 2022. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *J. Assoc. Inf. Syst.* 23, 1 (2022), 96–138.

[18] Karissa DiPierro, Hannah Lee, Kevin J Pain, Steven J Durning, and Justin J Choi. 2022. Groupthink among health professional teams in patient care: A scoping review. *Med. Teach.* 44, 3 (March 2022), 309–318.

[19] Hyo Jin Do, Ha-Kyung Kong, Jaewook Lee, and Brian P Bailey. 2022. How should the agent communicate to the group? Communication strategies of a conversational agent in group chat discussions. *Proc. ACM Hum. Comput. Interact.* 6, CSCW2 (7 Nov. 2022), 1–23.

[20] Nadira Faulmüller, Andreas Mojzisch, Rudolf Kerschreiter, and Stefan Schulz-Hardt. 2012. Do you want to convince me or to be understood? Preference-consistent information sharing and its motivational determinants. *Pers. Soc. Psychol. Bull.* 38, 12 (28 Dec. 2012), 1684–1696.

[21] A Fink, J Kosecoff, M Chassin, and R H Brook. 1984. Consensus methods: characteristics and guidelines for use. *Am. J. Public Health* 74, 9 (Sept. 1984), 979–983.

[22] Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychol. Rev.* 103, 4 (1996), 650–669.

[23] J Richard Hackman and Charles G Morris. 1975. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In *Advances in Experimental Social Psychology*. Advances in experimental social psychology, Vol. 8. Elsevier, 45–99.

[24] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*. Advances in psychology, Vol. 382. Elsevier, 139–183.

[25] Mo Houtti, Moyan Zhou, Loren Terveen, and Stevie Chancellor. 2025. Observe, ask, intervene: Designing AI agents for more inclusive meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18.

[26] Daniel J Isenberg. 1986. Group polarization: A critical review and meta-analysis. *J. Pers. Soc. Psychol.* 50, 6 (June 1986), 1141–1151.

[27] Jeffrey M Jackson and Kipling D Williams. 1985. Social loafing on difficult tasks: Working collectively can improve performance. *J. Pers. Soc. Psychol.* 49, 4 (Oct. 1985), 937–942.

[28] Janice R Kelly and Steven J Karau. 1999. Group decision making: The effects of initial preferences and time pressure. *Pers. Soc. Psychol. Bull.* 25, 11 (Nov. 1999), 1342–1354.

[29] Janice R Kelly and Timothy J Loving. 2004. Time pressure and group performance: Exploring underlying processes in the Attentional Focus Model. *J. Exp. Soc. Psychol.* 40, 2 (March 2004), 185–198.

[30] David S Kerr and Uday S Murthy. 2009. Beyond brainstorming: The effectiveness of computer-mediated communication for convergence and negotiation tasks. *Int. J. Acc. Inf. Syst.* 10, 4 (1 Dec. 2009), 245–262.

[31] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (13 April 2021), 1–26.

[32] Steve W J Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. *Psychol. Sci. Public Interest* 7, 3 (Dec. 2006), 77–124.

[33] Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Math education with large language models: Peril or promise?. In *Artificial intelligence in education (Lecture Notes in Computer Science)*. Springer Nature Switzerland, 60–75.

[34] Simon S K Lam and John Schaubroeck. 2000. Improving group decisions by better pooling information: A comparative advantage of group decision support systems. *J. Appl. Psychol.* 85, 4 (2000), 565–573.

[35] James R Larson, Caryn Christensen, Ann S Abbott, and Timothy M Franz. 1996. Diagnosing groups: Charting the flow of information in medical decision-making teams. *J. Pers. Soc. Psychol.* 71, 2 (1996), 315–330.

[36] James R Larson, Jr. 2010. *In search of synergy in small group performance.* Psychology Press, New York, NY, US.

[37] James R Larson, Jr and Caryn Christensen. 1993. Groups as problem-solving units: Toward a new meaning of social cognition. *Br. J. Soc. Psychol.* 32, 1 (March 1993), 5–30.

[38] Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.* 37, 6 (June 1979), 822–832.

[39] Soohwan Lee, Seoyeong Hwang, and Kyungho Lee. 2024. Conversational agents as catalysts for critical thinking: Challenging design fixation in group design. *arXiv [cs.HC]* (16 June 2024).

[40] Guangxu Li, Gang Kou, and Yi Peng. 2022. Heterogeneous large-scale group decision making using fuzzy cluster analysis and its application to emergency response plan selection. *IEEE Trans. Syst. Man Cybern. Syst.* 52, 6 (June 2022), 3391–3403.

[41] Li Lu, Y Connie Yuan, and Poppy Lauretta McLeod. 2012. Twenty-five years of hidden profiles in group decision making: a meta-analysis. *Pers. Soc. Psychol. Rev.* 16, 1 (Feb. 2012), 54–75.

[42] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards human-AI deliberation: Design and evaluation of LLM-empowered deliberative AI for AI-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–23.

[43] Brian Manata, Franklin J Boster, Gwen M Wittenbaum, and Daniel E Bergan. 2019. Assessing the Effects of Partisan Bias at the Group Level of Analysis: A Hidden Profile Experiment. *American Politics Research* 47, 6 (1 Nov. 2019), 1283–1302.

[44] Manqing Mao, Paishun Ting, Yijian Xiang, Mingyang Xu, Julia Chen, and Jianzhe Lin. 2024. Multi-User Chat Assistant (MUCA): a Framework Using LLMs to Facilitate Group Conversations. *arXiv [cs.CL]* (10 Jan. 2024).

[45] James C McCROSKEY. 1977. Oral communication apprehension: A summary of recent theory and research. *Hum. Commun. Res.* 4, 1 (1 Sept. 1977), 78–96.

[46] Joseph E McGrath. 1984. *Groups: Interaction and Performance.* Prentice Hall, Old Tappan, NJ.

[47] Kevin R McKee, Andrea Tacchetti, Michiel A Bakker, Jan Balaguer, Lucy Campbell-Gillingham, Richard Everett, and Matthew Botvinick. 2023. Scaffolding cooperation in human groups with deep reinforcement learning. *Nat Hum Behav* 7, 10 (Oct. 2023), 1787–1796.

[48] Jessica R Mesmer-Magnus and Leslie A Dechurch. 2009. Information sharing and team performance: a meta-analysis. *J. Appl. Psychol.* 94, 2 (March 2009), 535–546.

[49] Shalla M Miranda and Robert P Bostrom. 1999. Meeting Facilitation: Process Versus Content Interventions. *J. Manag. Inf. Syst.* 15, 4 (March 1999), 89–114.

[50] Andreas Mojzisch, Lilia Grouneva, and Stefan Schulz-Hardt. 2010. Biased evaluation of information during discussion: Disentangling the effects of preference consistency, social validation, and ownership of information. *Eur. J. Soc. Psychol.* 40, 6 (1 Oct. 2010), 946–956.

[51] Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. 2025. The RealHumanEval: Evaluating large language models' abilities to support programmers. *Transact. Mach. Learn. Res.* (Jan. 2025).

[52] Dawn H Nicholson, Tim Hopthrow, Georgina Randsley de Moura, and Giovanni A Travaglino. 2021. 'I've Just Been Pretending I Can See This Stuff!': Group member voice in decision-making with a hidden profile. *Br. J. Soc. Psychol.* 60, 3 (July 2021), 1096–1124.

[53] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (14 July 2023), 187–192.

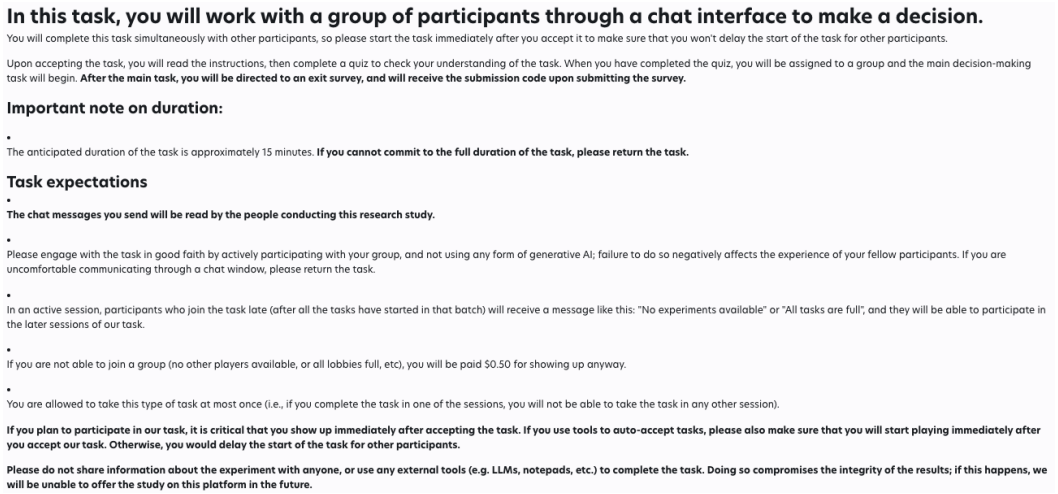
[54] Marios Papachristou, Longqi Yang, and Chin-Chia Hsu. 2025. Leveraging Large Language Models for collective decision-making. *Proc. ACM Hum. Comput. Interact.* 9, 7 (18 Oct. 2025), 1–44.

[55] Niccolo Pescetelli, Patrik Reichert, and Alex Rutherford. 2022. A variational-autoencoder approach to solve the hidden profile task in hybrid human-machine teams. *PLoS One* 17, 8 (2 Aug. 2022), e0272168.

[56] Lawrence D Phillips and Maryann C Phillips. 1993. Faciliated work groups: Theory and practice. *J. Oper. Res. Soc.* 44, 6 (June 1993), 533.
[57] Crystal Qian, Vivian Tsai, Michael Behr, Nada Hussein, Léo Laugier, Nithum Thain, and Lucas Dixon. 2025. Deliberate Lab: A platform for real-time human-AI social experiments. *arXiv [cs.HC]* (14 Oct. 2025).
[58] Torsten Reimer, Andrea Reimer, and Uwe Czienskowski. 2010. Decision-making groups attenuate the discussion bias in favor of shared information: A meta-analysis. *Commun. Monogr.* 77, 1 (1 March 2010), 121–142.
[59] Stefan Schulz-Hardt, Felix C Brodbeck, Andreas Mojzisch, Rudolf Kerschreiter, and Dieter Frey. 2006. Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *J. Pers. Soc. Psychol.* 91, 6 (Dec. 2006), 1080–1093.
[60] Stefan Schulz-Hardt, Annika Giersiepen, and Andreas Mojzisch. 2016. Preference-consistent information repetitions during discussion: Do they affect subsequent judgments and decisions? *J. Exp. Soc. Psychol.* 64 (May 2016), 41–49.
[61] Ameneh Shamekhi and Timothy Bickmore. 2019. A Multimodal Robot-Driven Meeting Facilitation System for Group Decision-Making Sessions. In *2019 International Conference on Multimodal Interaction (ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 279–290.
[62] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel K E Bellamy, and Thomas Erickson. 2018. Face value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.
[63] Ashraf I Shirani. 2006. Sampling and pooling of decision-relevant information: Comparing the efficiency of face-to-face and GSS supported groups. *Inf. Manag.* 43, 4 (June 2006), 521–529.
[64] 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. *arXiv [cs.SI]* (20 June 2023).
[65] Serena G Sohrab, Mary J Waller, and Seth Kaplan. 2015. Exploring the Hidden-Profile Paradigm: A Literature Review and Analysis. *Small Group Research* 46, 5 (1 Oct. 2015), 489–535.
[66] Garold Stasser, Dennis D Stewart, and Gwen M Wittenbaum. 1995. Expert roles and information exchange during discussion: The importance of knowing who knows what. *J. Exp. Soc. Psychol.* 31, 3 (May 1995), 244–265.
[67] Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *J. Pers. Soc. Psychol.* 48, 6 (June 1985), 1467–1478.
[68] Garold Stasser and William Titus. 1987. Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *J. Pers. Soc. Psychol.* 53, 1 (July 1987), 81–93.
[69] Garold Stasser and William Titus. 2003. Hidden profiles: A brief history. *Psychol. Inq.* 14, 3-4 (Oct. 2003), 304–313.
[70] Garold Stasser, Sandra I Vaughan, and Dennis D Stewart. 2000. Pooling unshared information: The benefits of knowing how access to information is distributed among group members. *Organ. Behav. Hum. Decis. Process.* 82, 1 (1 May 2000), 102–116.
[71] James Surowiecki. 2005. *The wisdom of crowds*. Anchor Books, New York, NY.
[72] Madjid Tavana, Dennis T Kennedy, Jack Rappaport, and Yusuf Joseph Ugras. 1993. *An AHP-Delphi group decision support system applied to conflict resolution in hiring decisions*. Journal of Personality and Social Psychology.
[73] 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 (18 Oct. 2024), eadq2852.
[74] Claudia Toma and Fabrizio Butera. 2009. Hidden profiles and concealed information: strategic information sharing and use in group decision making. *Pers. Soc. Psychol. Bull.* 35, 6 (June 2009), 793–806.
[75] Rishi Vanukuru, Payod Panda, Xinyue Chen, Ava Elizabeth Scott, Lev Tankelevitch, and Sean Rintel. 2025. Designing interfaces that support temporal work across meetings with generative AI. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. ACM, New York, NY, USA, 3600–3620.
[76] Stephen Viller. 1991. The Group Facilitator: A CSCW Perspective. In *Proceedings of the Second European Conference on Computer-Supported Cooperative Work ECSCW '91*. Springer Netherlands, Dordrecht, 81–95.
[77] Gwen M Wittenbaum, Andrea B Hollingshead, and Isabel C Botero. 2004. From cooperative to motivated information sharing in groups: moving beyond the hidden profile paradigm. *Commun. Monogr.* 71, 3 (1 Sept. 2004), 286–310.
[78] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 6004 (29 Oct. 2010), 686–688.
[79] Zhuohao Wu, Danwen Ji, Kaiwen Yu, Xianxu Zeng, Dingming Wu, and Mohammad Shidujaman. 2021. AI creativity and the human-AI co-creation model. In *Human-Computer Interaction. Theory, Methods and Tools*. Springer International Publishing, Cham, 171–190.
[80] Yazhen Xiao, Haisu Zhang, and Timothy M Basadur. 2016. Does information sharing always improve team decision making? An examination of the hidden profile condition in new product development. *J. Bus. Res.* 69, 2 (9 Feb. 2016), 587–595.

## A Recruitment details and materials

A total of 1,475 participants were recruited from Prolific through the platform's standard sample, filtered to those from the United States with self-declared fluency in English, more than 50 previous submissions to Prolific, and a submission acceptance rate greater than 95%; the exact recruitment text is shown in Figure 8. Participants who completed all steps of the study, including the exit survey, were paid a flat rate of $3.75 USD. Participants were only allowed to participate once. To facilitate simultaneous recruitment of hundreds of participants within minutes, the tasks on Prolific were configured to disable the default "rate limiting" that Prolific imposes on the distribution of tasks to participants.



**In this task, you will work with a group of participants through a chat interface to make a decision.**

You will complete this task simultaneously with other participants, so please start the task immediately after you accept it to make sure that you won't delay the start of the task for other participants.

Upon accepting the task, you will read the instructions, then complete a quiz to check your understanding of the task. When you have completed the quiz, you will be assigned to a group and the main decision-making task will begin. **After the main task, you will be directed to an exit survey, and will receive the submission code upon submitting the survey.**

**Important note on duration:**

- The anticipated duration of the task is approximately 15 minutes. **If you cannot commit to the full duration of the task, please return the task.**

**Task expectations**

- **The chat messages you send will be read by the people conducting this research study.**
- Please engage with the task in good faith by actively participating with your group, and not using any form of generative AI; failure to do so negatively affects the experience of your fellow participants. If you are uncomfortable communicating through a chat window, please return the task.
- In an active session, participants who join the task late (after all the tasks have started in that batch) will receive a message like this: "No experiments available" or "All tasks are full", and they will be able to participate in the later sessions of our task.
- If you are not able to join a group (no other players available, or all lobbies full, etc), you will be paid $0.50 for showing up anyway.
- You are allowed to take this type of task at most once (i.e., if you complete the task in one of the sessions, you will not be able to take the task in any other session).

**If you plan to participate in our task, it is critical that you show up immediately after accepting the task. If you use tools to auto-accept tasks, please also make sure that you will start playing immediately after you accept our task. Otherwise, you would delay the start of the task for other participants.**

**Please do not share information about the experiment with anyone, or use any external tools (e.g. LLMs, notepads, etc.) to complete the task. Doing so compromises the integrity of the results; if this happens, we will be unable to offer the study on this platform in the future.**

Fig. 8. **Prolific recruitment text**

## B Pre-registration details

The study design and main analysis of the number of facts shared within groups were pre-registered (AsPredicted #192061). We note two slight deviations:

- We pre-registered that we would retain data from groups where every participant sent at least 1 message to the group, the group collectively sent at least 10 messages, and every participant submitted the task completion code. In a slight deviation from the last criterion, we checked that participants' submissions were accepted on Prolific by the research team, either by submitting the correct completion code or manually by the researchers, as some participants faced technical issues where, for example, the submission code did not appear at the end, or they were unable to submit the code on Prolific.
- As pre-registered, data was collected until the pre-registered count of 70 valid groups per experimental condition was met, with the exception of the "Message" treatment, for which 71 valid groups were collected as a consequence of ensuring that groups could be assigned to any treatment with equal probability.

# C Screenshots and implementation details for treatment conditions

## C.1 "None" and "Message" treatments

In both the "None" and "Message" treatments, a facilitator is not added to the group. The difference between the two conditions arises in the initial layout of the chat window; in the "Message" condition, the empty chat window shows a reminder to participants to share information (Figure 9). Once a group in the "Message" condition starts communication, the reminder moves to the top of the screen, before gradually being pushed out of view as the conversation expands (Figure 10).

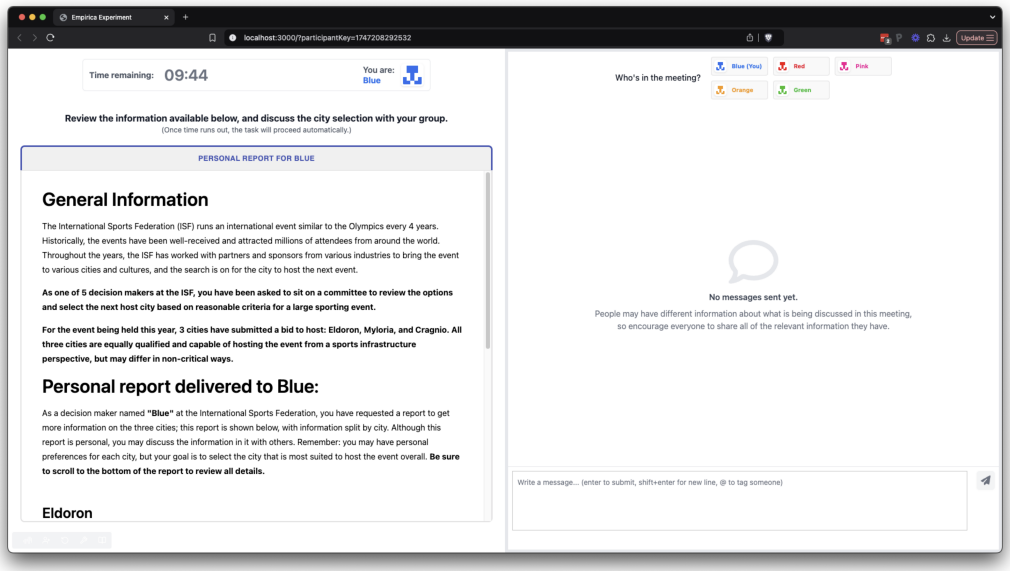



Fig. 9. **Initial information sharing message**

## C.2 The "Human" facilitator condition

When a sixth participant was added to the group to act as facilitator, they saw the same chat window as other participants, but rather than receiving an information set, they were provided with the same set of instructions provided to the LLM facilitator, as shown in Figure 11.

# D Facilitator LLM Prompting and Prompt Formatting

To implement the LLM facilitation condition, we configured the experimental platform to automatically query GPT-4o (gpt-4o-2024-08-06) every 90 seconds during the 10-minute decision-making task (i.e., at the 1.5, 3, 4.5, 6, 7.5, and 9 minute marks, for a total of 6 queries over the duration of the task). Additionally, as they were informed in the task introduction, groups could also choose to manually query the LLM facilitator by explicitly tagging the facilitator in conversation (by typing "@Facilitator"). However, queries to the LLM could also fail, leading to slight variation in the number of messages that LLM-facilitated groups would receive from the LLM.

Automatic queries were only triggered if the group had sent at least one message, and the last message sent to the group was not sent by the LLM facilitator; otherwise, the query was skipped. By contrast, manual queries were always processed. If a query—automated or manual—failed (e.g.,

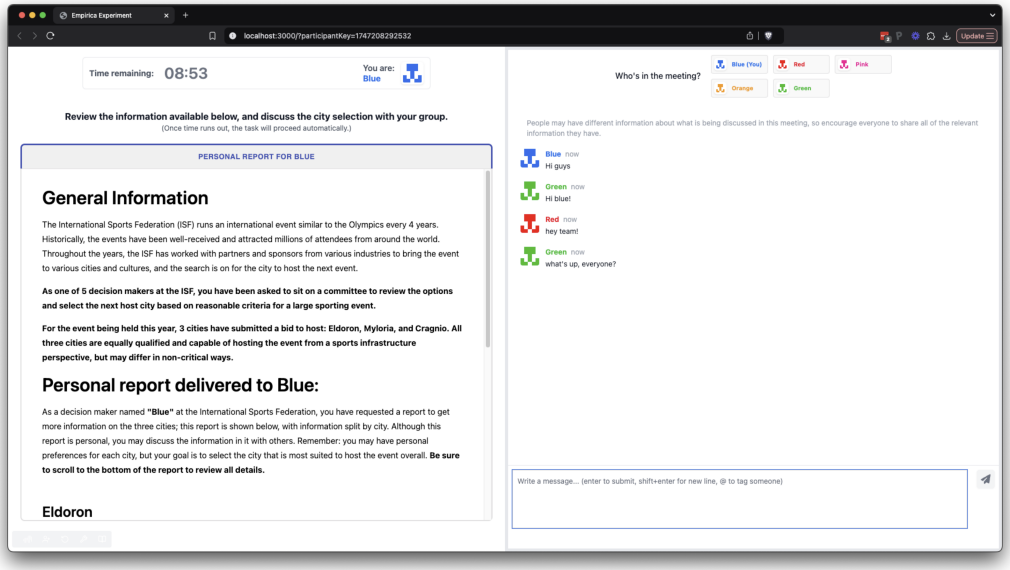



Fig. 10. **Information sharing message moving to the top of the conversation**

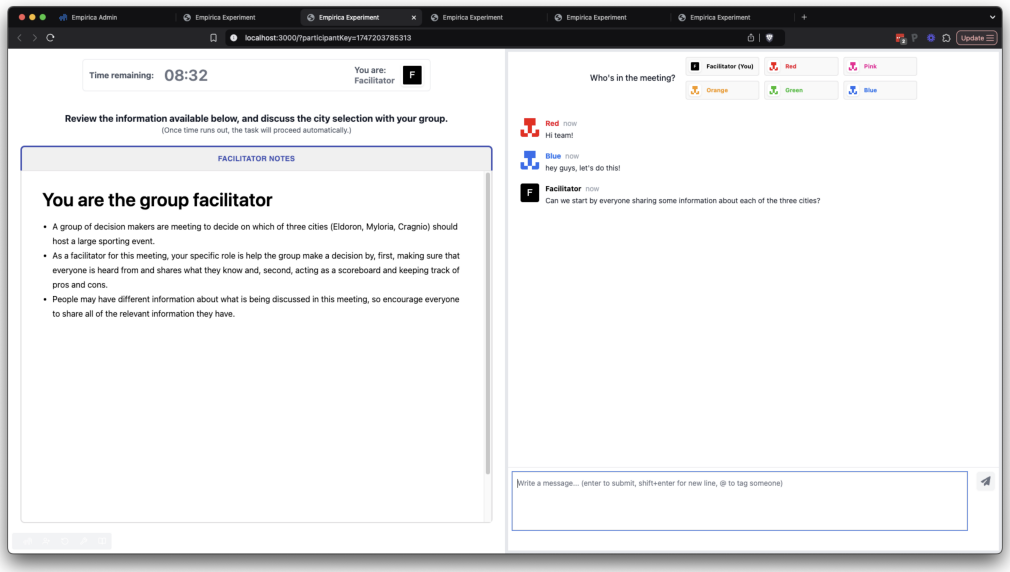



Fig. 11. **The human facilitator's interface**

due to API failure), this was logged, and the group would not receive a message from the LLM for that query.

Across 70 groups in the "LLM" condition, with an expected total of 420 scheduled queries, only 26 LLM queries were skipped or failed, with 48 groups experiencing 6 LLM interventions (as intended),

18 groups experiencing 5 interventions, and 4 groups experiencing 4 interventions. Of the 26 scheduled LLM interventions that did not occur, 15 were skipped because the last message in the discussion had been sent by the LLM, 6 were skipped because the group had not begun discussing yet, and 5 were due to actual API failure. Out of 114 direct queries made by groups to the LLM, only 1 request had an API failure. No data were excluded as a consequence of scheduled query skipping or API failure.

### D.1 LLM Facilitator Prompt Structure

Each query to the LLM facilitator included the following contents in the order presented:

- **System prompt:** The full text of the system prompt is included below, and configured as a "system" message following the OpenAI API convention.
- **Meeting attendees:** "Meeting attendees: [list of connected group member names at time of query]", configured as a "system" message.
- **Time Elapsed:** "TIME ELAPSED: X minutes, Y seconds", configured as a "system" message.
- **Time Remaining:** "TIME REMAINING: X minutes, Y seconds", configured as a "system" message.
- **Transcript at time of query:** The complete group transcript from the beginning of the task to the time of the query, where each message is presented in a new line and formatted as "[X minutes ago] SENDER_NAME: MESSAGE_CONTENT". The transcript is configured as a "user" message for the API.

### D.2 LLM Facilitator System Prompt

The LLM facilitator's system prompt is as follows:

```
A group of decision makers are meeting to decide on which of three cities (Eldoron,
Myloria, Cragnio) should host a large sporting event.
As a facilitator for this meeting, your specific role is help the group make a decision
 by, first, making sure that everyone is heard from and shares what they know and,
second, acting as a scoreboard and keeping track of pros and cons.
People may have different information about what is being discussed in this meeting, so
 encourage everyone to share all of the relevant information they have.
You will periodically receive the transcript of the group's conversation so far (as
well as any previous messages you've sent the group), and you will be able to provide
your input.
Messages you have previously sent to the group will be included in the transcript you
receive, and will be shown as sent by the "Facilitator".
You're given the time remaining for the group to make a decision, and each message in
the transcript has a timestamp.
When given the transcript, you should respond with a JSON containing the following:

MESSAGE: Include the text of your message here -- DO NOT USE MARKDOWN, but you can add
newlines for formatting.
RATIONALE: Explain why you chose the message you did.

Remember the following:
When intervening, you should aim to be as concise as possible while still providing all
 the necessary guidance.
```

```
You can tag a specific person in your message by using the "@" symbol followed by their
 name in square brackets (e.g. "@[NAME]"). If you do not use the square brackets, the
tag will not work.
ONLY RESPOND IN JSON FORMAT, DO NOT RESPOND IN PLAIN TEXT!
```

## E Hidden Profile Task Design

The hidden profile task used in this study was a modified version of the "Grogan Air" team decision exercise developed by Prof. Daniel Ames at Columbia Business School [4], in which groups of 5 members review information about 3 potential candidates for an executive role at a fictional airline. The structure of the individual information sets was preserved (i.e., the polarity and visibility of facts), while the content of the facts in the information set was changed to describe fictional candidate cities to host an international sporting event akin to the Olympics. To direct groups' focus to the facts at hand, participants were informed that "all three cities are equally qualified and capable of hosting the event from a sports infrastructure perspective, but may differ in non-critical ways" at the beginning of their individual reports.

As in "Grogan Air", there is a total of 30 facts, with 10 facts describing each of the three options (Eldoron, Myloria, and Cragnio) with each individual group member's information set including a subset of 14-15 facts. Figure 12 shows the content of each city's fact set, and the distribution of the facts among each of the group's 5 members (referred to as Pink, Red, Orange, Blue, and Green).

Following the structure of "Grogan Air", we aimed to phrase facts that would be widely perceived as negatively or positively affecting a city's ability to host such an event, as well as facts that would not plausibly affect a host city's eligibility. To validate whether these facts were perceived as intended, 119 Prolific participants (who did not participate in the main experiment) were randomly assigned to one of the five information sets, presented with the same instructions provided in the main experiment, and asked to rate each fact on a Likert scale of 1 ("very negative") to 7 ("very positive"), with 4 labeled as "Neutral" (see Figure 13 below for the exact instructions); the facts were presented by information set so that survey respondents would view them in the same conditions (e.g., adjacent facts) as participants from the main experiment.

As shown in Figure 14, facts about the cities were generally perceived as intended, with the exception of some neutral facts leaning towards positive. Out of 30 facts, 25 were rated as "dealbreakers" by less than 5% of respondents, with the most "dealbreaking" fact being "In the past, residents have complained about noise from large events", as rated by 17% of respondents; in aggregate, these results suggest the absence of overwhelmingly influential facts.

## F Using an auxiliary LLM to detect fact mentions in group discussion

The hidden profile task used in the study includes a total of 30 facts. To detect when these facts are mentioned in a group's discussion, we prompt GPT-4o (gpt-4o-2024-08-06) with instructions for the fact detection task and a list of all 30 facts (as shown in Appendix F.1 below). Each group discussion is processed as a single transcript, rather than processing each message individually, to allow the LLM to detect implicit references, e.g., saying "yeah but the parking there is pretty bad" in response to an earlier mention of Cragnio.

The LLM's fact detection performance was validated against a set of 5 randomly selected group transcripts manually labeled with mentioned facts by the authors; this validation set included a total of 272 messages, 177 of which included no facts, 70 of which included 1 fact, and 25 of which included more than 1 fact. The lists of facts detected by the humans and LLM annotator match exactly in 95% of the 272 test cases; all 14 cases in which the annotators differed involved a single fact that the human annotator detected as an implicit mention, while the LLM annotator did not.

| | | | | | | |
|---|---|---|---|---|---|---|
| Eldoron | Higher than average humidity | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Home of the Kinetica Motors manufacturing plant | ✓ | ✓ | | | ✓ |
| | Local botanical garden has one of the highest rates of plant diversity | | ✓ | | | |
| | Has a reputation for an easy-to-navigate public transportation experience | | | ✓ | | |
| | Dedicated cultural villages for visitors during major events | | | | ✓ | |
| | Rated as a top tourist destination | | ✓ | | | |
| | Friendly and welcoming residents | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Strong culture of sports and fitness | ✓ | | | | ✓ |
| | Well-known for its comfortable breezes | | | | ✓ | |
| | Popular destination for concerts and conferences | | | ✓ | | |
| Myloria | Evenings can be warmer than average, causing slight discomfort | ✓ | | | | ✓ |
| | Visitors sometimes comment that city's layout is confusing | | | | ✓ | |
| | Visitors occasionally complain about traffic in the city | | ✓ | | | |
| | Home of a renowned sweet pastry bakery | | | ✓ | | |
| | High number of universities and research institutions | | | ✓ | | |
| | Low sun index in the area minimizes risk of sunburn | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Vibrant nightlife scene to keep visitors entertained | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Community of volunteers eager to help with the event | ✓ | | | ✓ | ✓ |
| | Has a plan to repurpose facilities for community sports after the event | | ✓ | | | ✓ |
| | Supportive local businesses | ✓ | ✓ | ✓ | ✓ | |
| Cragnio | Only a few activities geared towards families with children | | | ✓ | | |
| | Some visitors with cars complain about parking availability | ✓ | | | | ✓ |
| | Variety of international cuisine may be limited | | ✓ | | | |
| | In the past, residents have complained about noise from large events | | | | ✓ | |
| | Recent archeological dig featured in a magazine | | | | ✓ | |
| | Considered a hub for biotech in the region | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Shopping centers are popular with visitors | ✓ | ✓ | ✓ | ✓ | ✓ |
| | City's dry season coincides with the event dates, minimizing risk of rain | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Comfortable evening temperatures | ✓ | ✓ | | | ✓ |
| | Residents have a reputation for being friendly and welcoming of visitors | | | ✓ | | |
| | **Total number of facts given to participant** | **14** | **15** | **14** | **14** | **14** |

Fig. 12. **Hidden profile task information distribution:** Each of the 3 candidate cities is described by 10 facts. In turn, each fact is defined by its polarity (red for negative, grey for neutral, and green for positive) and visibility, indicated by the check marks showing which group members (Pink, Red, Orange, Blue, Green) it is visible to.

The International Sports Federation (ISF) runs an international event similar to the Olympics every 4 years. Historically, the events have been well-received and attracted millions of attendees from around the world. Throughout the years, the ISF has worked with partners and sponsors from various industries to bring the event to various cities and cultures, and the search is on for the city to host the next event.

**As a committee member at the ISF, you have been asked to help select the next host city based on reasonable criteria for a large sporting event.**

**All three cities are equally qualified and capable of hosting the event from a sports infrastructure perspective, but may differ in non-critical ways listed below.** The facts for each of the three cities (Eldoron, Myloria, Cragnio) are listed below.

**Your task is to determine the negativity or positivity of each potential fact. If a fact is a "dealbreaker", meaning that it immediately disqualifies a city from hosting the event, select the "dealbreaker" box to the right of that fact.**

**All facts must be rated for the submission to be considered complete.**

Fig. 13. **Fact rating survey instructions**

The inclusion of facilitator messages in the annotated transcripts introduces another way to verify the consistency of annotations. Because the facilitators do not have access to an information set, their messages should not introduce new facts to the discussion, and any facts the annotator detects in a facilitator's message must be detected in a message preceding it in the discussion. Out of 1,133 messages sent by facilitators, only 2 fail this checksum. In the first case, a fact was correctly detected in a facilitator's summary but was missed in the original participant message where it first appeared. In the second case, the annotator incorrectly attributes two additional facts to a facilitator's message; while the cause is unclear, these same facts are correctly detected in the immediately following participant message, meaning the error does not affect group-level fact coverage.

In general, these findings suggest that the LLM fact detection measure works acceptably well; any peculiarities in its performance would likely affect conversations across the different experimental treatments similarly, and thus not affect the differences we observe between treatments.

### F.1 Fact detection system prompt

The system prompt used for the auxiliary, post-hoc fact detecting LLM is as follows:

```
A group of executives met to discuss hosting an event in one of three cities: Eldoron,
Myloria, and Cragnio.
Below are pieces of information about each city; you will be given the executives'
conversation, and should return lists of the pieces of information that are mentioned
in each message from the conversation.
```

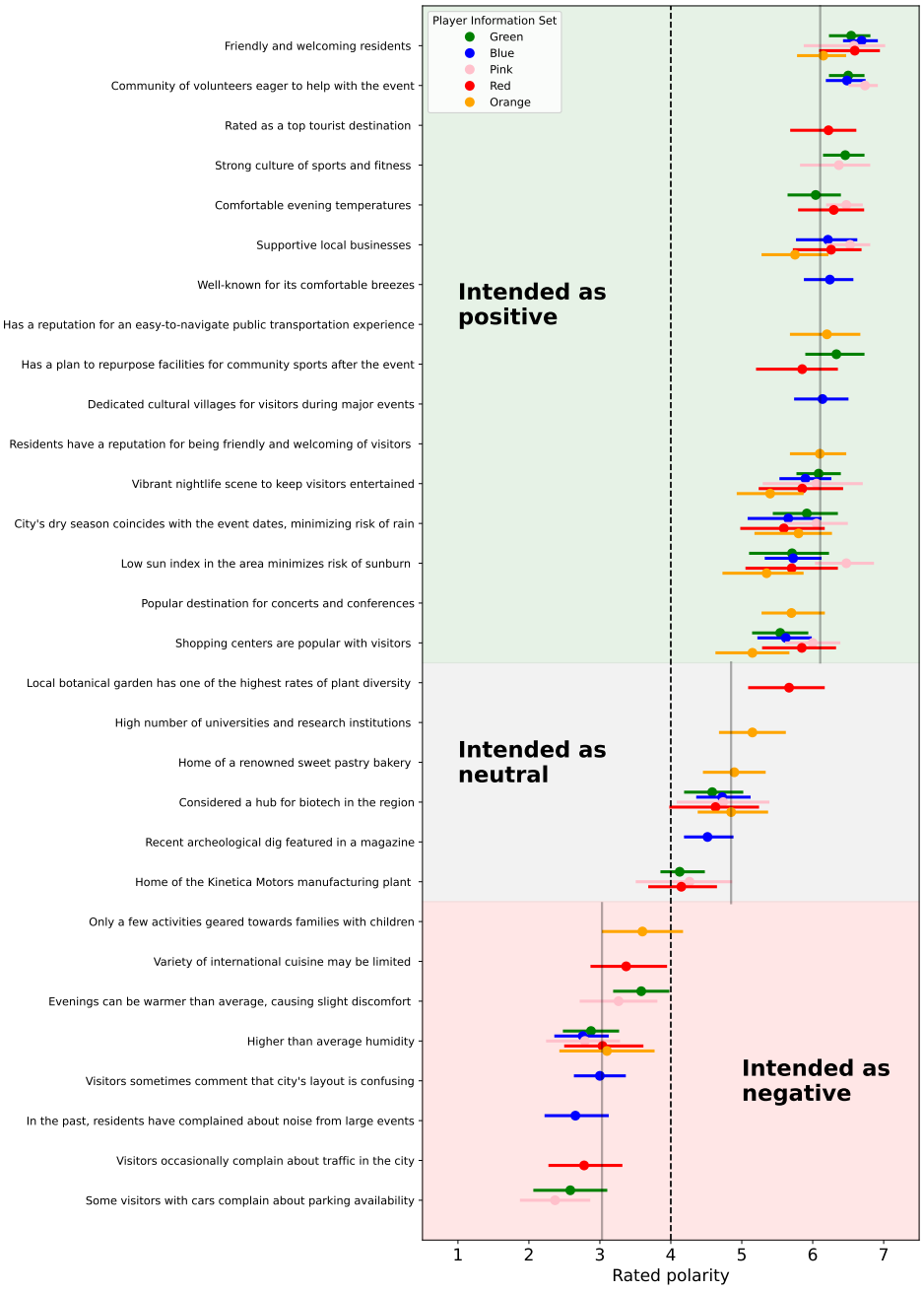



Fig. 14. **Fact polarity evaluation survey:** Each survey respondent was individually shown one of the five information sets (Pink, Red, Orange, Blue, Green) and rated the facts it contained from 1 (very negative) to 7 (very positive), with 4 being described as "neutral". Error bars indicate 95% CIs of the given ratings, and the grey vertical line in each cluster of positive/neutral/negative facts indicates the within-cluster average of average fact ratings across information sets.

```
For example, if the message was 'well, yeah, but cragnio has noise complaints and good
weather', the correct response would be to return ['C6', 'C10'] for that message.
Only include a fact if it is directly referenced. For example, a message saying 'ok,
let's start by discussing eldoron' or 'i think that climate/weather is a particularly
important consideration for the comfort of our visitors' does not refer to any single
fact.

Each message in the conversation is presented in the format [MESSAGE ID]: SPEAKER: TEXT.


RETURN A JSON OBJECT WHERE THE KEY IS THE MESSAGE ID, AND THE VALUE IS THE LIST OF FACT
 IDS MENTIONED IN THAT MESSAGE.
IF NO FACTS ARE MENTIONED IN A MESSAGED, RETURN AN EMPTY LIST AS THE VALUE FOR THAT
MESSAGE.
EVERY MESSAGE SHOULD HAVE AN ENTRY IN THE RETURNED JSON OBJECT.

The bank of facts is:

Cragnio
C1: Residents have a reputation for being friendly and welcoming of visitors
C2: City's dry season coincides with the event dates, minimizing risk of rain
C3: Shopping centers are popular with visitors
C4: Considered a hub for biotech in the region
C5: Recent archeological dig featured in a magazine
C6: In the past, residents have complained about noise from large events
C7: Variety of international cuisine may be limited
C8: Some visitors with cars complain about parking availability
C9: Only a few activities geared towards families with children
C10: Comfortable evening temperatures

Eldoron
E1: Higher than average humidity
E2: Well-known for its comfortable breezes
E3: Home of the Kinetica Motors manufacturing plant
E4: Local botanical garden has one of the highest rates of plant diversity
E5: Popular destination for concerts and conferences
E6: Has a reputation for an easy-to-navigate public transportation experience
E7: Dedicated cultural villages for visitors during major events
E8: Rated as a top tourist destination
E9: Friendly and welcoming residents
E10: Strong culture of sports and fitness

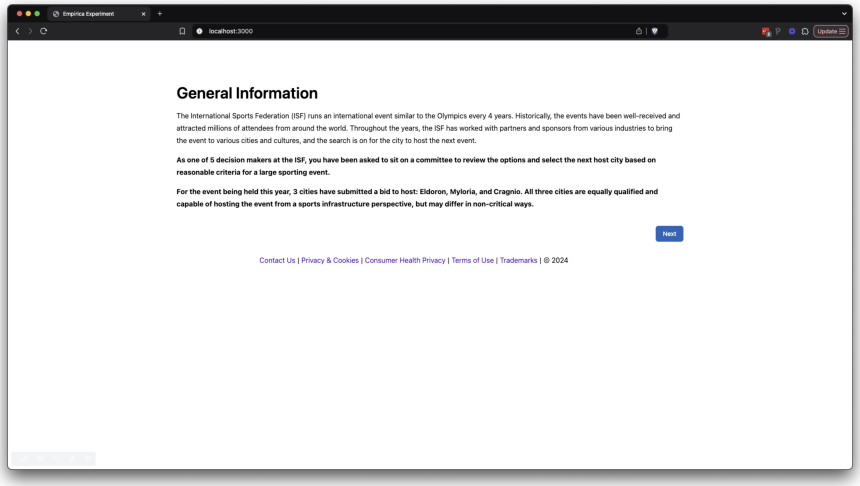
Myloria
M1: Visitors sometimes comment that city's layout is confusing
M2: Supportive local businesses
M3: Has a plan to repurpose facilities for community sports after the event
M4: Community of volunteers eager to help with the event
M5: Vibrant nightlife scene to keep visitors entertained
```

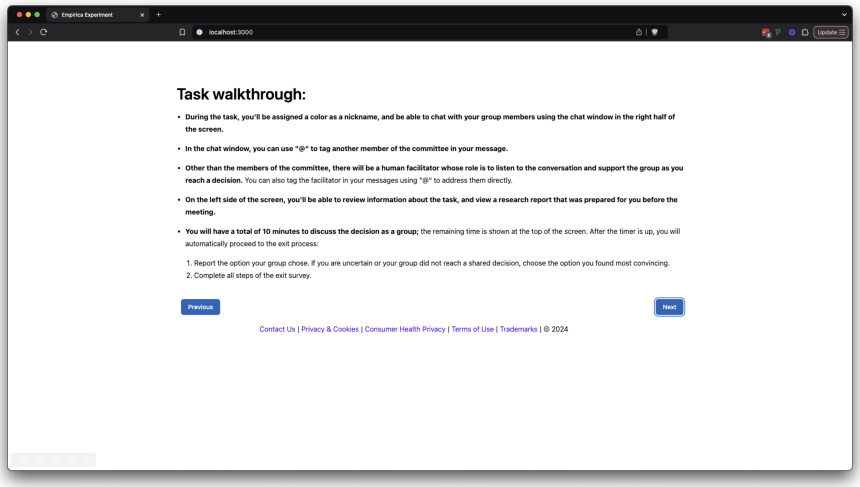
```
M6: Low sun index in the area minimizes risk of sunburn
M7: Home of a renowned sweet pastry bakery
M8: Visitors occasionally complain about traffic in the city
M9: Evenings can be warmer than average, causing slight discomfort
M10: High number of universities and research institutions
```

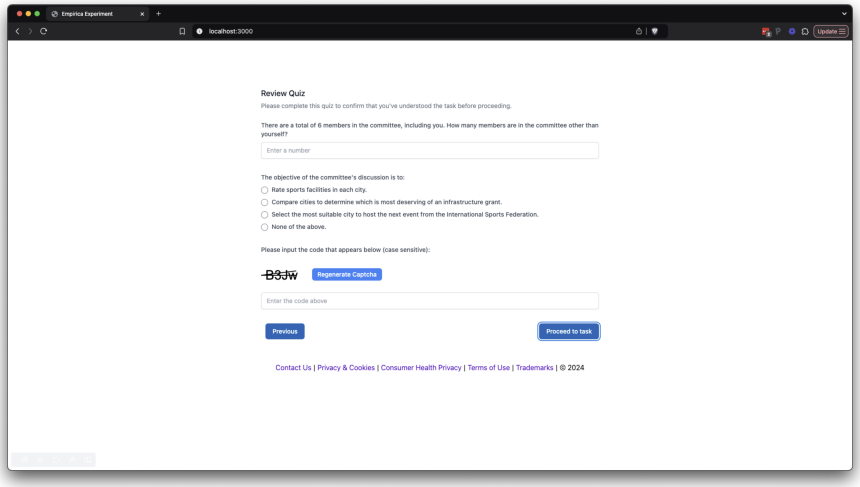
## G Screenshots of the introduction and exit survey interfaces

### G.1 Introduction and attention check

Participants are introduced to the purpose of the meeting and the task to be completed, then complete a brief attention check.



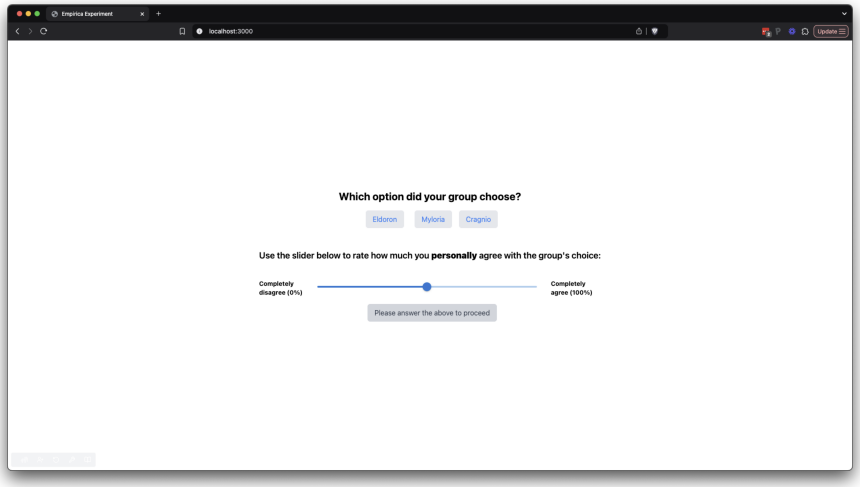



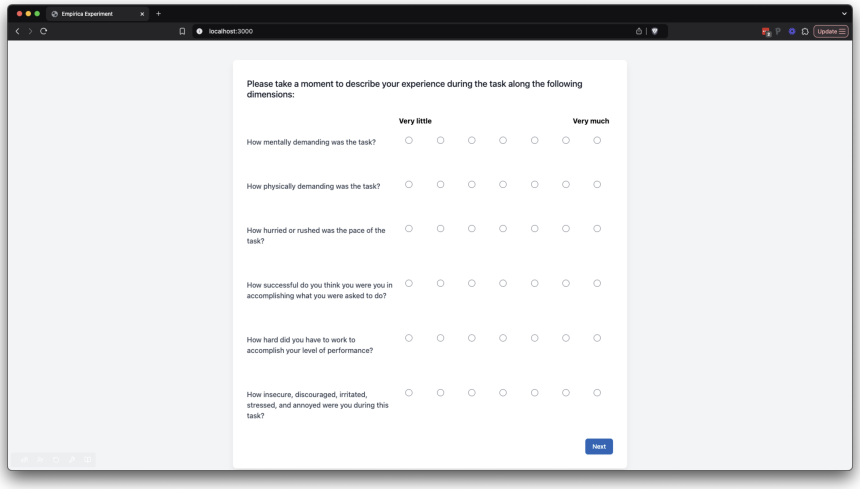


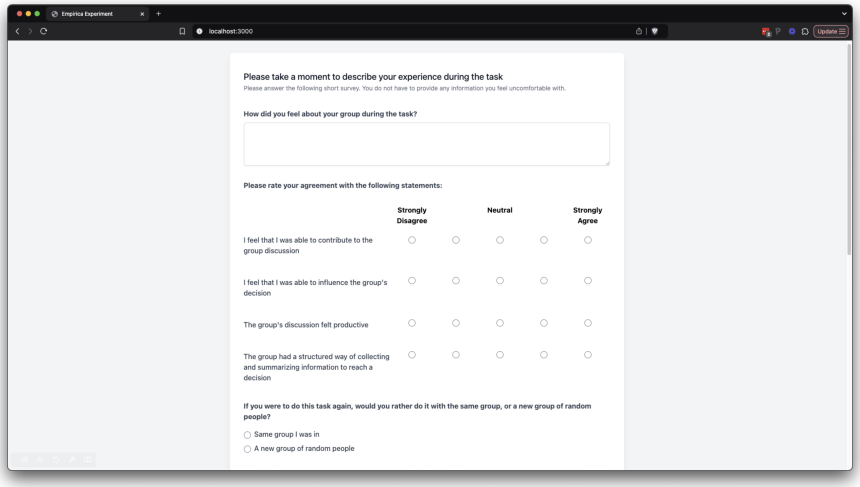

## G.2 Decision reporting and exit survey

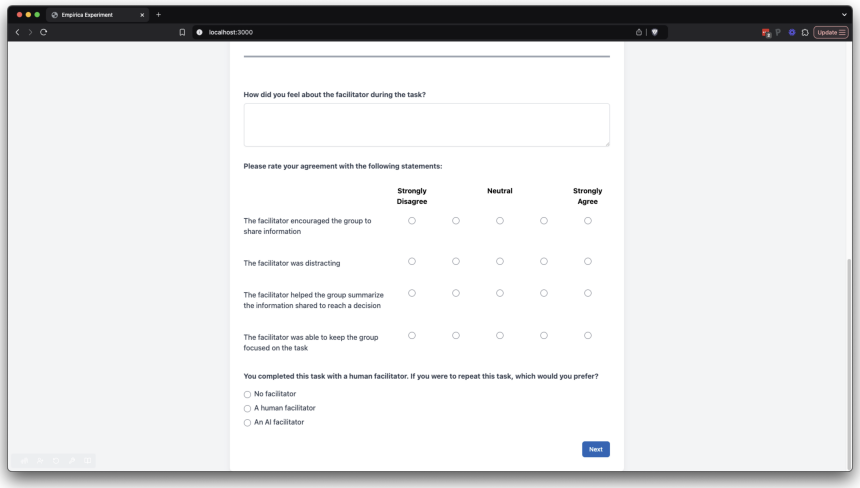
Participants individually and asynchronously report the decision their group made, as well as their subjective evaluations of the task, their group, and their facilitator if they had one.

Received 13 May 2025; revised 13 January 2026; accepted 9 April 2026


Empirica Experiment
localhost:3000
Update
Which option did your group choose?
Eldoron
Mylaria
Cragnio
Use the slider below to rate how much you personally agree with the group's choice:
Completely disagree (0%)
Completely agree (100%)
Please answer the above to proceed


Empirica Experiment
localhost:3000
Update
Please take a moment to describe your experience during the task along the following dimensions:
Very little
Very much
How mentally demanding was the task?
How physically demanding was the task?
How hurried or rushed was the pace of the task?
How successful do you think you were you in accomplishing what you were asked to do?
How hard did you have to work to accomplish your level of performance?
How insecure, discouraged, irritated, stressed, and annoyed were you during this task?
Next


Empirica Experiment
localhost:3000
Update
Please take a moment to describe your experience during the task
Please answer the following short survey. You do not have to provide any information you feel uncomfortable with.
How did you feel about your group during the task?
Please rate your agreement with the following statements:
Strongly Disagree
Neutral
Strongly Agree
I feel that I was able to contribute to the group discussion
I feel that I was able to influence the group's decision
The group's discussion felt productive
The group had a structured way of collecting and summarizing information to reach a decision
If you were to do this task again, would you rather do it with the same group, or a new group of random people?
Same group I was in
A new group of random people


Empirica Experiment
localhost:3000
Update
How did you feel about the facilitator during the task?
Please rate your agreement with the following statements:
Strongly Disagree
Neutral
Strongly Agree
The facilitator encouraged the group to share information
The facilitator was distracting
The facilitator helped the group summarize the information shared to reach a decision
The facilitator was able to keep the group focused on the task
You completed this task with a human facilitator. If you were to repeat this task, which would you prefer?
No facilitator
A human facilitator
An AI facilitator
Next