Facilitating Asynchronous Idea Generation and Selection with Chatbots

Joongi Shin
Aalto University
Helsinki, Finland
joongishin@gmail.com

Ankit Khatri
Aalto University
Helsinki, Finland
ankitk.cs.21@nitj.ac.in

Michael A. Hedderich
LMU Munich & MCML
Munich, Germany
hedderich@cis.lmu.de

Andrés Lucero
Aalto University
Helsinki, Finland
lucero@acm.org

Antti Oulasvirta
Aalto University
Helsinki, Finland
antti.oulasvirta@aalto.fi

Abstract

People can generate high-quality ideas by building on each other's ideas. By enabling individuals to contribute their ideas at their own comfortable time and method (i.e., asynchronous ideation), they can deeply engage in ideation and improve idea quality. However, running asynchronous ideation faces a practical constraint. Whereas trained human facilitators are needed to guide effective idea exchange, they cannot be continuously available to engage with individuals joining at varying hours. In this paper, we ask how chatbots can be designed to facilitate asynchronous ideation. For this, we adopted the guidelines found in the literature about human facilitators and designed two chatbots: one provides a structured ideation process, and another adapts the ideation process to individuals' ideation performance. We invited 48 participants to generate and select ideas by interacting with one of our chatbots and invited an expert facilitator to review our chatbots. We found that both chatbots can guide users to build on each other's ideas and converge them into a few satisfying ideas. However, we also found the chatbots' limitations in social interaction with collaborators, which only human facilitators can provide. Accordingly, we conclude that chatbots can be promising facilitators of asynchronous ideation, but hybrid facilitation with human facilitators would be needed to address the social aspects of collaborative ideation.

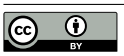
CCS Concepts

• **Human-centered computing** → **Collaborative and social computing**.

Keywords

Asynchronous ideation, conversational agent, facilitator

ACM Reference Format:
Joongi Shin, Ankit Khatri, Michael A. Hedderich, Andrés Lucero, and Antti Oulasvirta. 2024. Facilitating Asynchronous Idea Generation and Selection with Chatbots. In *36th Australasian Conference on Human-Computer Interaction (OzCHI '24), November 30–December 04, 2024, Brisbane, QLD, Australia.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3726986.3726994



OzCHI '24, Brisbane, QLD, Australia

ACM ISBN 979-8-4007-1509-9/24/11
https://doi.org/10.1145/3726986.3726994

1 INTRODUCTION

Human creativity thrives on collective intelligence [65]. By exchanging ideas and perspectives, people can diversify and converge to innovative solutions beyond individuals' consideration [3, 4, 41, 65]. One form of collaborative ideation is *asynchronous ideation*, wherein collaborators independently contribute ideas through a shared online platform [11, 49, 63]. In contrast to synchronous ideation, where collaborators ideate together in real time, asynchronous ideation offers flexibility for individuals to ideate at their own comfortable pace and method (Figure 1). This enables individuals to reflect deeply on their ideation process, which has been shown to improve idea quality [13].

However, facilitating effective idea exchange in asynchronous settings faces a practical challenge; that is, human *facilitators* cannot be available continuously to guide individuals joining at varying hours [51]. In any collaborative ideation events, trained facilitators take crucial roles in guiding collaborators to *build on each other's ideas* [20, 45]. They actively engage with individuals and lead them to generate, criticize, refine, and select a few promising ideas as a group effort. Without facilitators, collaborators may ideate ineffectively, diverging towards unrelated subjects and failing to reach a consensus [45, 69].

In this paper, we ask *how to design conversational agents to facilitate asynchronous idea generation and selection* (Figure 2). Unlike human facilitators, conversational agents such as chatbots have no restriction on being continuously available, which could provide facilitation on demand. While prior arts have shown chatbots' potential in facilitating online discussion and brainstorming [23, 30, 36, 43, 61], they focus on the synchronous discussion as a group (e.g., collaborators developing ideas by taking turns) and do not investigate asynchronous ideation (e.g., collaborators building on each other's ideas independently). Accordingly, we hypothesize that chatbots can adopt human facilitators' behaviors for *converging individuals' ideation efforts into a collaborative effort* (Figure 3). Chatbots could;

- present collaborators' ideas and opinions as *inspirations*;
- suggest *ideation methods* to build on the others' ideas and opinions; and
- request individuals to *rate* their ideas for their ideation goal.

Building chatbots has become more accessible thanks to Large Language Models (LLMs), but it is unclear how to design the interactions so that they can actually facilitate asynchronous ideation.

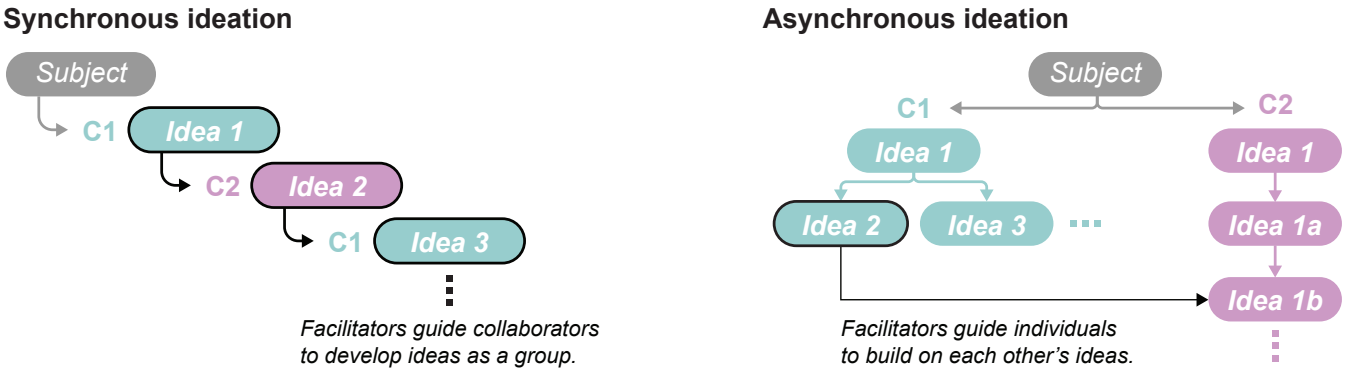


Figure 1: A conceptual model of synchronous and asynchronous ideation. Unlike synchronous ideation, where collaborators (C1 and C2) develop ideas by taking turns, we focus on asynchronous ideation where collaborators generate ideas individually, at their own pace and method (right). For example, while C1 diversifies ideas, C2 could focus on improving ideas. Here, facilitators' role will be presenting C1's idea as an inspiration to C2.

To close this gap, we distill the guidelines found in the literature on human facilitators [54, 56, 68] and propose two chatbots: One provides a structured ideation process (i.e., **structured facilitator**) and another adapt the ideation process to individuals' ideation performance (i.e., **adaptive facilitator**).

The structured facilitator guides collaborators to first diversify ideas and then improve them. This approach has been proven to be effective in generating a large number of ideas, which also increases the number of high-quality ideas [55]. During the idea selection phase, the facilitator focuses users' attention on the ideas that have been rated to be helpful [12]. We designed the adaptive facilitator inspired by how human facilitators spontaneously alter their guidance to provide tailored ideation processes to individuals [12]. The adaptive facilitator provides inspirations (i.e., similar or dissimilar ideas) and ideation methods (i.e., generate any or improved ideas) based on how well individuals generate ideas from them. During the idea selection phase, the facilitator focuses users' attention on the ideas that have 'uncertain' group opinions (i.e., ideas with fewer and more diverse opinions). By studying both approaches, we demonstrate how human facilitators' behaviors can be translated into chatbots.

The chatbots are supported by a semantic similarity classifier that can retrieve and present other collaborators' ideas by their similarity to individuals' own ideas. For this, we tested a prompt-based classifier using LLM (GPT-4) [19] and a fine-tuned classifier using a pretrained language model (DistilBERT) [6]). We found that the fine-tuned classifier can estimate the semantic similarity as accurately as the prompt-based classifier but at a faster processing speed, making it more suitable for conversational interaction with collaborators. In addition, we make the adaptive facilitator provide helpful inspirations and ideation methods based on how well users can generate quality ideas from them. For this, we used the Multi-Armed Bandit (MAB) algorithm [67].

We conducted two studies to understand the strengths and weaknesses of our chatbots in facilitating asynchronous ideation. First, we assessed the chatbots from collaborators' perspectives. We invited 48 participants to generate and select ideas by interacting with one of our chatbots. We analyzed their interaction behaviors, perception of ideating through the chatbots, and satisfaction with the resulting ideas. Then, we explored our chatbots' potential with a human facilitator, an expert who provided critical perspectives based on their facilitation practice over 25 years. We sensitized the expert by having them facilitate an in-person ideation workshop in the same structure as our chatbots facilitated and interviewed their perspectives on facilitating asynchronous ideation with chatbots.

The results show that both chatbots were helpful in building on other group members' ideas. The structured facilitator was more helpful in diversifying ideas, and the adaptive facilitator could elicit satisfaction at a similar level to the human-facilitated ideation. However, we also identified the potential limitations of chatbots such as the lack of accountability and social interaction that only human facilitators can provide. Accordingly, we conclude that chatbots can facilitate collaborative ideation in asynchronous settings, yet human experts' intermediate facilitation would still be required. With this paper, we uncover the positives and negatives of chatbots in facilitating asynchronous ideation (Table 2) and make the following contributions:

(1) We demonstrate designing two chatbot facilitators. Both chatbots adopt the literature-based guidelines for facilitating effective collaborative ideation and employ computational methods.
(2) We found the positives and negatives of the chatbot facilitators from collaborators' and an expert human facilitator's perspectives.
(3) We provide the implications of chatbot facilitators in practice. We also offer our code for implementing chatbot facilitators using our computational methods as well as sample prompts to support those who aim to implement LLM-based facilitators using our design guidelines [1].

2 RELATED WORK

Facilitators' roles in collaborative ideation have been extensively investigated in design and HCI literature [12, 18, 20, 45, 48], which

[1] https://github.com/joongishin/aiFacilitator

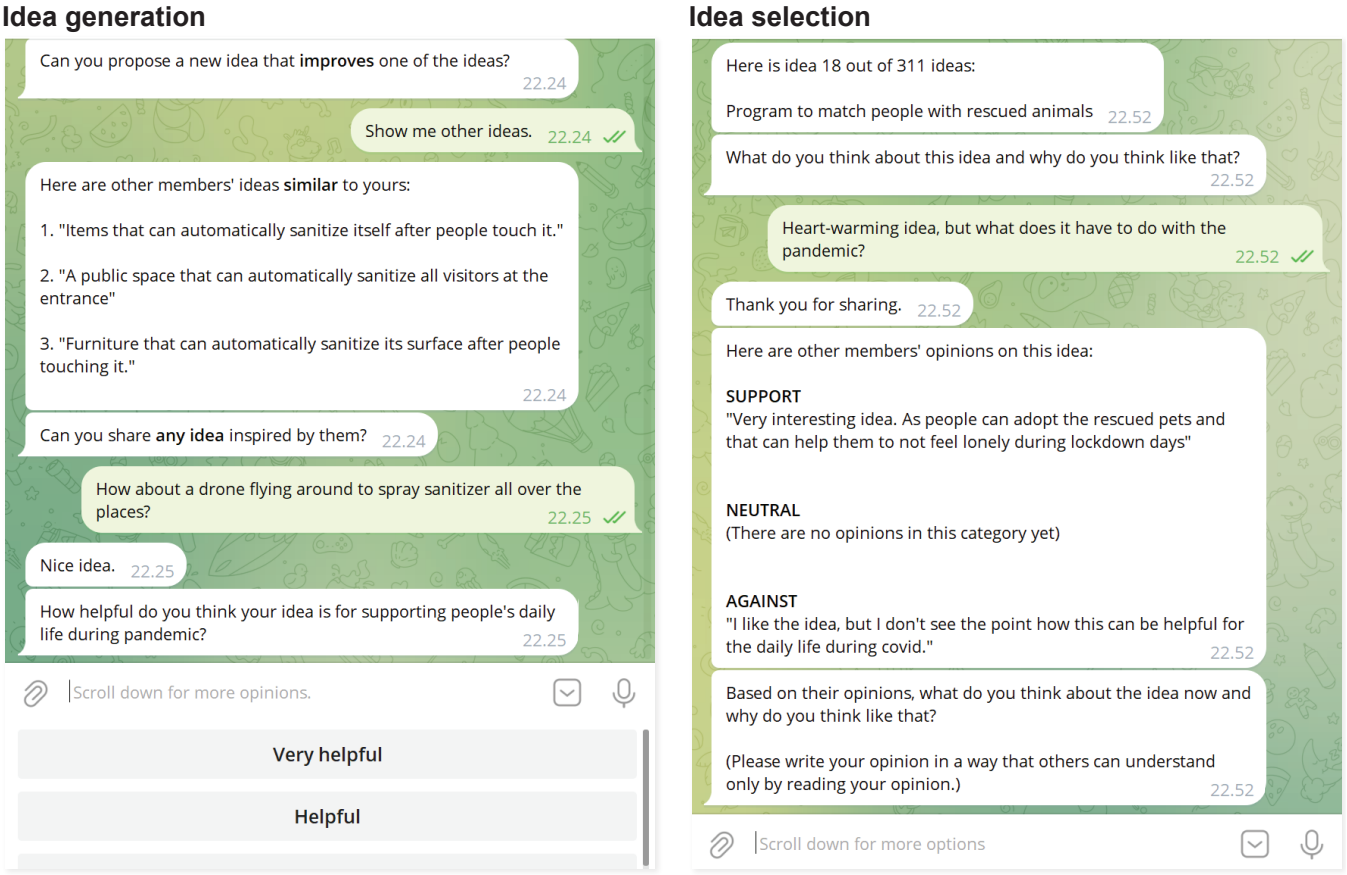


Figure 2: We designed chatbots that interact with individual collaborators and facilitate their asynchronous idea generation (left) and selection (right). We adapted guidelines from the literature on human facilitators to design our chatbots' behaviors, presenting other collaborators' ideas as inspiration and suggesting ideation methods to guide individuals to build on each other's ideas.

help train human facilitators. We build on such knowledge and propose an alternative way of facilitating asynchronous ideation with conversational agents.

2.1 Facilitators in Collaborative Ideation

Regardless of individuals' expertise in collaborative activities, facilitators take essential roles by closely engaging with participants [18, 45, 48]. Dahl and Sharma [12] identified six roles of facilitators that can enable effective idea exchange among group members, despite their varying communication skills or authority in design projects. In particular, we focus on the roles of enabler, direction setter, and users' advocate that provide direct support on ideation. The enabler assists individual group members to actively participate in their discussion, lowering the barrier of sharing their ideas and perspectives. The direction setter aligns group members' collaborative efforts with their project goals. The users' advocate represents and negotiates each user group's perspective, convincing other group members to take each other's viewpoints. These roles would be particularly important in asynchronous settings, where collaborators need to consider other group members' ideas and opinions despite having no direct conversation with each other. One remark is, all the above-mentioned facilitation is achieved through conversation.

2.2 Interactive Support for Collaborative Ideation

To generate and select ideas as a group, collaborators need a setting where they can effectively exchange their ideas. In response, diverse interactive systems have been explored to support communication among a user group. For instance, Mazalke et al. [50] developed a digital tabletop where users can spatially share their stories in person. Kennedy et al. [33] explored conducting collaborative design in a digital environment and identified that online meeting platforms enabled more balanced idea exchange among the collaborators. Despite the utility of their systems, these approaches require users to collaborate at the same time, synchronously. In contrast, we aim to enable users to collaborate asynchronously, by adapting the guidelines on facilitating effective idea generation [16, 39].

Similar to our interests, digital platforms have been investigated to enable asynchronous idea generation [37, 48, 76]. For instance,

Literature-based guidelines	Our design decisions for chatbot facilitators
Enable independent ideation to boost idea generation [54]	Engage with individual collaborators in isolated settings
Present similar / dissimilar ideas to leverage group ideation [63]	Provide inspirational ideas by their similarity
Provide instructions for building on each other's ideas [48]	Suggest ideation methods
Focus collaborators' attention on their ideation goal [12]	Request collaborators to rate their ideas against the ideation goal

Table 1: Exemplary literature-based guidelines we adopted to design our chatbot facilitators.

Siangliulue et al., [63] studied how users group multiple ideas by their semantic similarity and developed a platform where they contribute their ideas by the groups. Online platforms such as Miro[2] and Mural[3] provide diverse functionalities for representing and organizing ideas. Online petition websites can also support a population to raise socially immanent issues in discussion [29]. Most systems, however, expect collaborators to be self-motivated to discuss their ideas and do not provide adaptive guidance as human facilitators do. Without facilitation, users would miss the others' meaningful contributions to build on or diverge ideas to unrelated subjects [45, 69]. In response, we study artificial facilitators that can provide concrete guidance to individual collaborators in asynchronous settings.

2.3 Chatbots as Facilitators

The recent advancements in LLMs have created powerful chatbots (e.g., ChatGPT [4]) that can conduct human-like conversation [31]. Such LLM-based chatbots have been shown to support people's ideation, such as summarizing extensive text [28] and generating ideas [17]. However, most applications have been limited to the interaction between a single user and a chatbot [15, 40, 42, 75]. How to design chatbots (or instruct LLM-based chatbots) that facilitate collaborative ideation among people should be studied. In this paper, we demonstrate designing chatbot facilitators by adopting human facilitators' behaviors.

Diverse collaborative ideation methods (e.g., discussing ideas using physical probes [1, 25]) have been practiced for users to explore and elaborate their ideas [21, 32, 34, 71, 74]. All methods, however, are designed to be led by trained facilitators. Conducting collaborative ideation hence highly depends on facilitators' expertise and availability. In response, scholars have investigated diverse roles of conversational agents to support collaborative activities [22, 24, 26, 53, 59, 72]. For instance, Wambsganss et al. [73] developed a chatbot that guides users to improve their persuasive writing. These studies show a promising result that private discussion with a chatbot can prime users for collaboration. In alignment with our aim, a few studies developed chatbots to facilitate ideation among group members. Kim et al. [35] developed a chatbot that moderates discussion in real-time, pacing collaborators' discussion, identifying inactive collaborators, and organizing their discussion points. The authors also investigated how chatbots can facilitate idea convergence at the end of discussion [36]. Hadfi et al. [23] designed chatbots to facilitate discussion on an online forum, which guides participants to sequentially comment on the previous opinion. While these studies show the potential of facilitating collaborative ideation via chatbots when human facilitators are not available, they are limited to synchronous ideation where collaborators need to exchange ideas in turn. The most similar study to ours was conducted by Shin et al. [62], developing a rule-based chatbot that facilitates consensus-building to resolve conflicts by enabling independent exploration of different opinions. We extend the study by designing chatbot facilitators that can guide asynchronous idea generation and selection.

3 LITERATURE-BASED GUIDELINES FOR CHATBOT FACILITATION

We distilled guidelines found in the literature that demonstrate human facilitators' roles from empirical analysis, case studies, and interviews on facilitating collaborative design [12, 20, 45, 48]. The guidelines include approaches for structuring group ideation and adaptively engaging with collaborators, boosting their collective effort in generating and selecting ideas. Accordingly, we made a conceptual model of effective facilitation (Figure 3) and four design decisions commonly applied to our chatbot facilitators (Table 1).

First, chatbots should engage with individuals in isolated settings. Facilitators often divide collaborators to work alone, without interfering with each other (e.g., Nominal group technique [27]). This enables individuals to generate ideas without taking turns (i.e., production blocking) and without worrying about the others' criticism (i.e., evaluation apprehension), hence effectively increasing the number of ideas [16, 39]. In the process, facilitators pay close attention to individuals [12], fostering their independent creativity [8, 54]. We expect chatbot facilitators also to guide individuals in solitary settings (e.g., chatrooms dedicated to individuals), enabling ideation at their own pace and creativity.

Second, chatbots should present inspirational ideas by their similarity to the proposed ideas. Facilitating collaborators to combine common ideas can lead to generating impactful ideas [41]. Showing other group members' dissimilar ideas can lead to generating unique ideas [63]. Likewise, we expect showing similar ideas to inspire exploration in the related subjects or different subjects to avoid duplicates. By showing dissimilar ideas, we expect collaborators to explore the subjects that they would not have considered otherwise [3, 54]. We apply the same reasoning for idea selection: showing dissimilar opinions would make collaborators consider diverse viewpoints and make correct judgments.

Third, chatbots should suggest ideation methods. Facilitators present a set of actions that collaborators can build on each others' contributions [20, 48]. For instance, during group brainstorming, facilitators specify the ideation method that group members can use to improve

[2]https://miro.com/
[3]https://www.mural.co/
[4]https://openai.com/blog/chatgpt

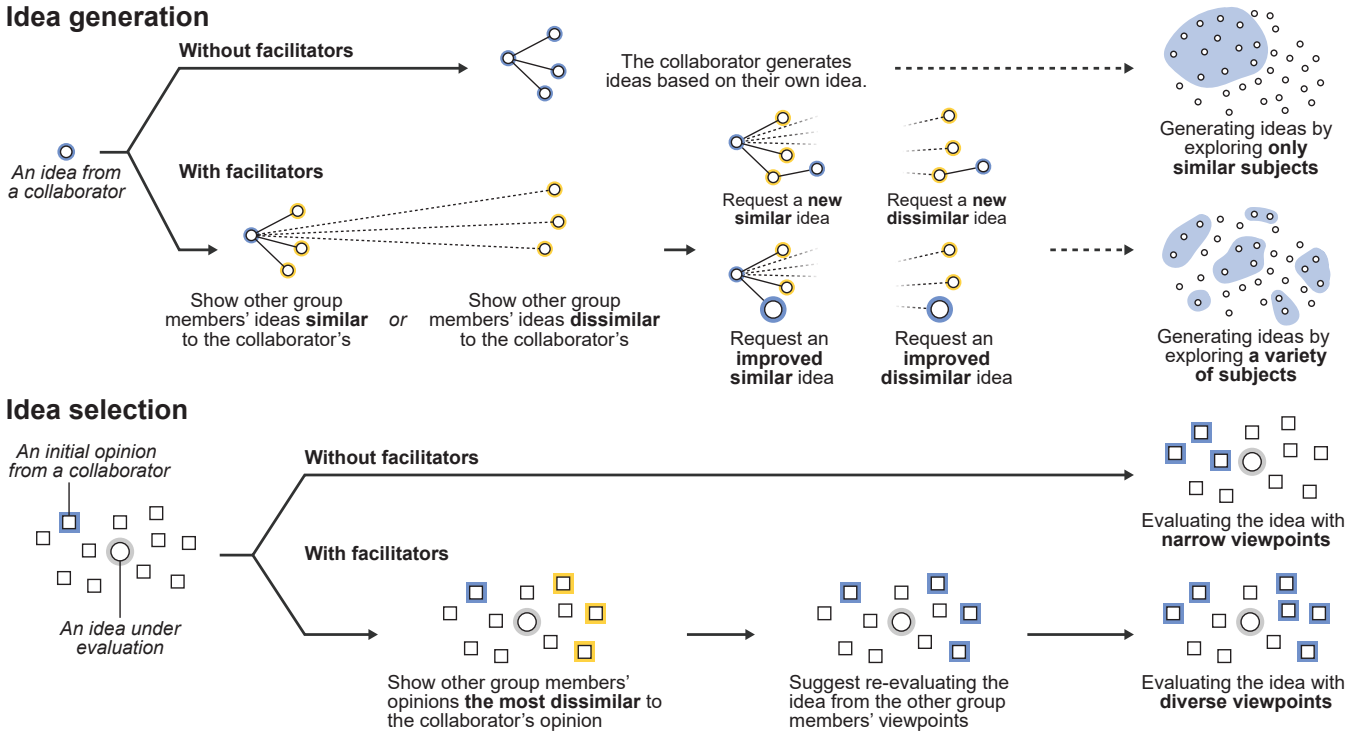


Figure 3: Collaborative idea generation (top) and selection (down) led by facilitators. By showing other group members' ideas and opinions, facilitators can guide individual collaborators to effectively generate more ideas and evaluate the ideas with more diverse viewpoints. We designed chatbot facilitators that could guide such effective ideation in asynchronous settings.

each other's ideas (e.g., "Imagine why this idea will not work" and "Imagine what would be needed to make this idea work") [70]. We adopt the behaviors and design the chatbots to suggest ideation methods in addition to providing inspiration. They suggest collaborators propose new or improved ideas during idea generation and re-evaluate the shown ideas from the others' viewpoint (i.e., perspective-taking) during idea selection.

Lastly, chatbots should request collaborators to rate their ideas regarding their ideation goals. While interacting with proposed ideas and opinions, collaborators could lose their focus and start developing unrelated ideas. To resolve this, facilitators often interfere and redirect collaborators' attention to the ideation goal [12]. Similarly, we employ the rating approach, which could quickly form group opinions and focus collaborators' attention on the ideation goal. We expect chatbots asking collaborators to rate their ideas with regard to ideation goals can help regain their focus after seeing or proposing unrelated ideas.

4 DESIGN OF CHATBOT FACILITATORS

We designed the structured and adaptive facilitators that guide individual collaborators to generate and select ideas by building on each other's contributions. Overall, both chatbots share the same conversation flow as shown in Figure 4.

The chatbots begin the idea generation phase by introducing an ideation goal and how they will support the users' ideation. They also encourage users to propose as many ideas as possible, unrestricted by current technology or resources. Afterward, the chatbots start the main cycle of idea generation. (a) The chatbots first show either three similar or dissimilar ideas to the users' latest idea (they show the common or rare ideas if the users just began the interaction, hence have not yet shared any idea). Then, (b) the chatbots suggest the users propose any or improved idea. After the users share an idea, (c) the chatbots ask them to rate their own idea using a 7-point Likert scale, considering how well it achieves the ideation goal. After the rating, (d) the chatbots repeat the three actions until the end of idea generation.

The chatbots begin the idea selection phase by reminding users about the ideation goal, highlighting the number of generated ideas, and encouraging users to review as many ideas as possible. After the introduction, the chatbots begin the cycle of idea selection. (e) The chatbots present one idea at a time and (f) ask for the users' initial opinions on the idea. Then, (g) the chatbots present other group members' opinions on the idea that are most dissimilar to the users' initial opinion. The chatbots present them in three categories (support, neutral, and against) based on the others' final ratings of the idea. (h) The chatbots then suggest users re-evaluate the idea considering the other group members' opinions. As a response, the users share either a new opinion or keep their initial one. The chatbots thank users for sharing their opinions and (i) request the users to rate how much they would like to try out the idea using a 7-point Likert scale. After collecting the users' ratings, the chatbots repeat the three actions until the idea selection phase ends.

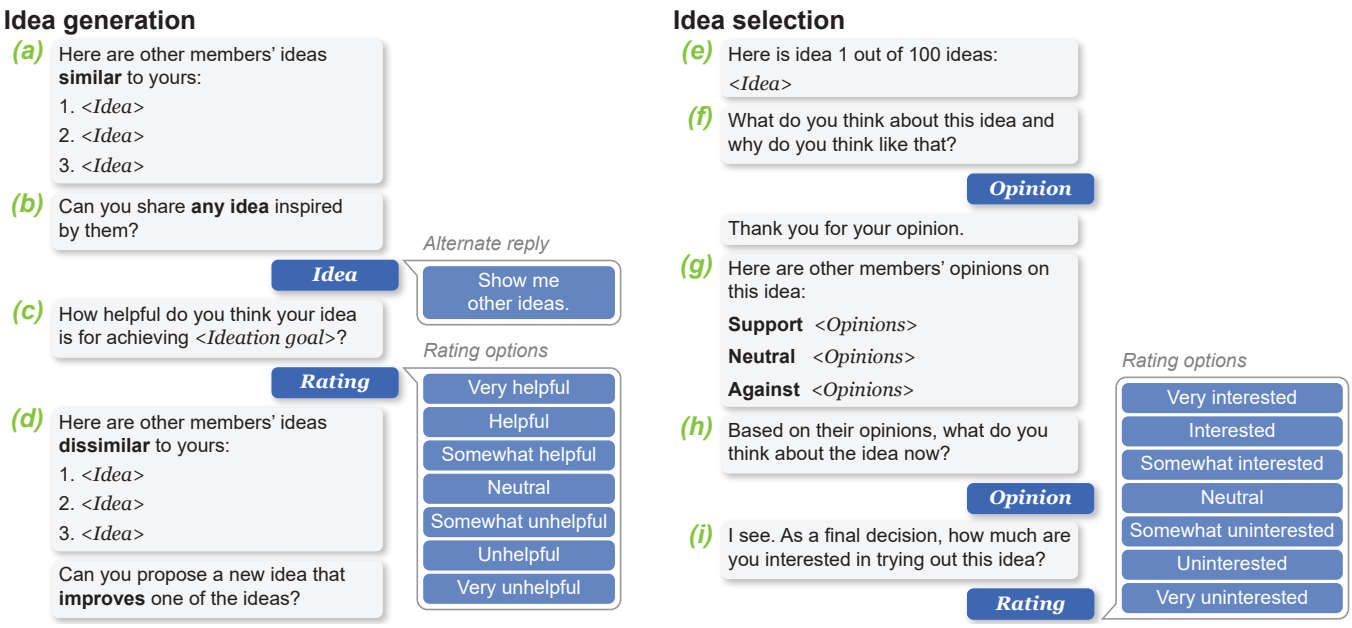


Figure 4: During idea generation (left), the chatbots repeat the cycle of presenting other group members' ideas as inspiration (a), suggesting an ideation method (b), and requesting ratings on the users' own ideas (c). Then, the chatbots show other types of inspiration and ideation methods (d). During idea selection (right), the chatbots show a collected idea (e), request users' initial opinion (f), show other group members' opinions (g), suggest re-evaluating the idea (h), and request ratings on the idea.

The structured and adaptive facilitators differ in how they decide which inspirations and ideation methods to present during idea generation and which ideas to focus users' attention on during idea selection. In particular, in the idea selection phase, we expect evaluating all generated ideas to be laborious. Therefore, we designed the chatbot facilitators to present ideas in the order that a group of individuals can reach a consensus without reviewing every idea. The following subsections provide details on the differences between the chatbot facilitators.

4.1 Structured Facilitator

The structured facilitator provides the same ideation process to all users. The facilitator divides idea generation into two parts: First diversifying ideas and then improving them. This is based on the principle that a group of users performs better when each ideation session has a single purpose, concentrating on one type of ideation at a time [54, 56, 68]. This also aligns with the 'quantity over quality' approach [55], which promotes generating a large number of ideas without worrying about their quality, hence exploring a broad spectrum of ideas. Our facilitator shows dissimilar ideas and suggests proposing any ideas to assist the divergent thinking in the first part. To assist in improving ideas, it shows similar ideas and suggests improving one of them.

During idea selection, the structured facilitator presents ideas in the descending order of their ratings received during the idea generation phase. This strategy can focus collaborators' attention on the ideas with higher potential, filtering out less interesting ones and reducing individuals' effort in reviewing all proposed ideas [16, 38, 46].

4.2 Adaptive Facilitator

During idea generation, the adaptive facilitator selects inspirations and ideation methods based on users' performance in generating ideas. For instance, if users generate more helpful ideas by responding to similar ideas and thinking about new ones, the adaptive facilitator will continue guiding the users with the same type of inspiration and ideation method. If users start to generate less helpful ideas, then the facilitator will suggest different types of inspiration and ideation methods. Whereas this approach will bring multiple types of ideation in a single phase and contradicts the principles applied to the structured facilitator, we hypothesize that the same principles may not apply when users can generate ideas at their own pace in the asynchronous setting.

During idea selection, the adaptive facilitator continuously updates which idea it should present next based on the 'uncertainty' of group opinions on the ideas. We assume that, after collecting sufficient number of opinions and ratings on an idea, users can be certain whether they would like or dislike the idea as a group. In other words, collecting more opinions on such ideas will not add more information about how collaborators think about the idea. Therefore, it would be more efficient if users review the other ideas that have not yet collected enough opinions. This can lead to distributing users' effort, which would be especially useful when there are many ideas to review.

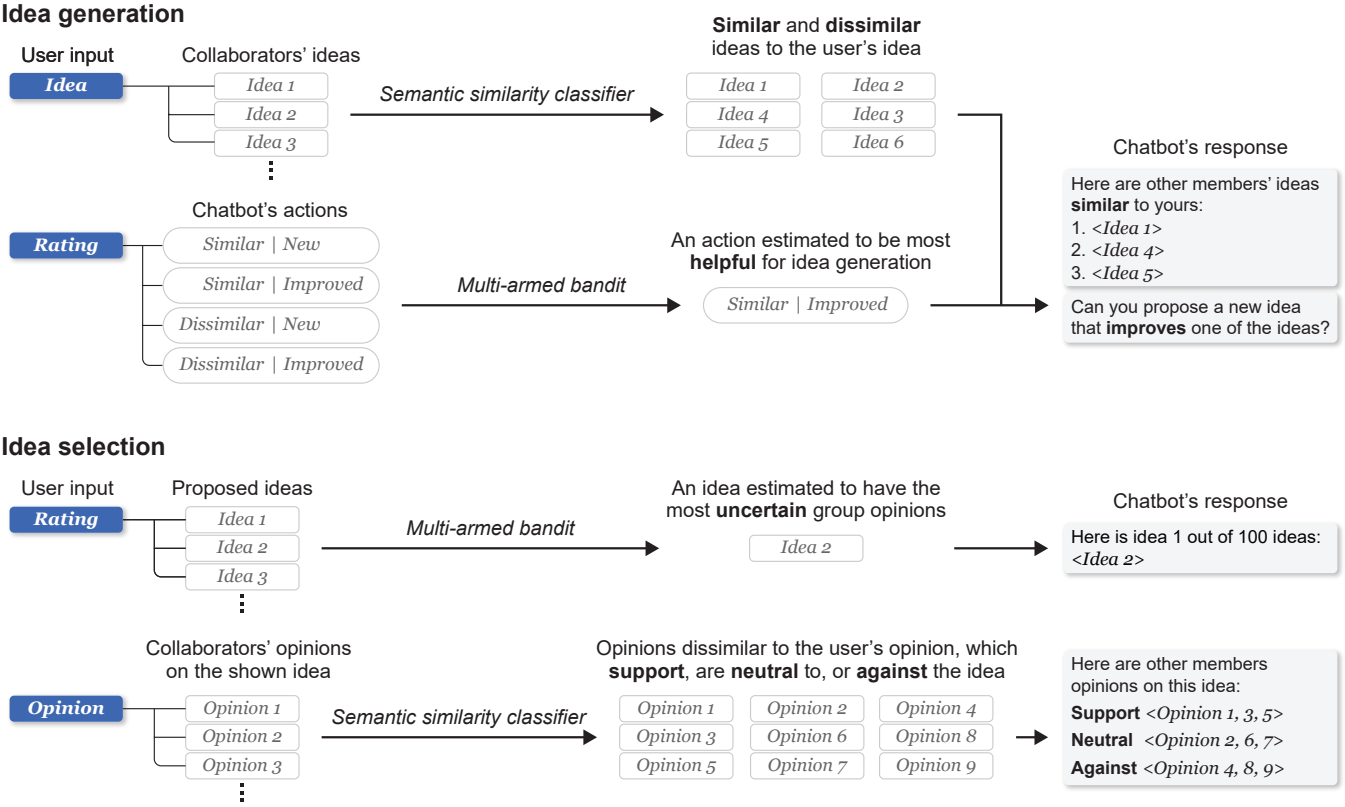


Figure 5: The system structure of the adaptive facilitator. During idea generation (top), the MAB system selects inspirations and ideation methods based on the individuals' rating of their own ideas. During idea selection (bottom), the MAB system presents ideas based on the users' collective rating of each idea (i.e., prioritizing ideas with uncertain group opinions).

5 COMPUTATIONAL APPROACHES FOR CHATBOT FACILITATORS

We implemented our chatbots' behaviors using Natural Language Processing (NLP) and probabilistic inference. This section provides an overview; detailed methods and evaluations are in Appendices A and B.

5.1 Semantic Similarity Classifier

Both chatbot facilitators present other group members' ideas and opinions by their similarity to the users' latest input. For this, they need to perform a Semantic Textual Similarity (STS) prediction, measuring how closely related two sentences are when looking at the information they convey. In the context of facilitating ideation, we aim for chatbots that compute STS aligned with humans' perception of idea similarity and at an interaction rate of speed.

In a preliminary experiment, we evaluated two classifiers for this STS task: a **prompt-based classifier** using LLM and a **fine-tuned classifier** based on a pretrained language model. For the prompt-based classifier, we used GPT-4[5], the latest model at the time of testing, and based our prompts on Gatto et al.'s instruction [19]. For the fine-tuned classifier, we used DistilBERT [58] and fine-tuned it on the SemEval2017-STS dataset [5]. Both classifiers give a higher score for more equivalent sentences (e.g., *The bird is bathing in the sink.* vs. *Birdie is washing itself in the water basin.*) and a lower score if two sentences are dissimilar (e.g., *A boat sails along the water.* vs. *The man is playing the guitar*), all definitions and examples from [5]. We measured the correlation between the classifiers' semantic similarity estimation to human annotations as well as their response latency in comparing 100 pairs of sentences. The detailed implementation and evaluation procedure are given in Appendix A.

The results show that the similarity predictions of both classifiers reflect human similarity judgments, well-achieving correlations $> 80\%$ with manually labeled ground truth data. However, the fine-tuned classifier is faster, spending on average 2.64 seconds (SD = 0.28), while the prompt-based classifier spent on average 5.16 seconds (SD = 1.20). Accordingly, we conclude the fine-tuned classifier is more suited to our interaction scenario and used it for our chatbots.

Internally, the fine-tuned classifier predicts a score on a 5-point scale. Based on comparison with human annotation, we split this core into three ranges for our application when judging pairs of user-submitted ideas: for a similarity score below 2, the ideas are marked as dissimilar, above 2 as similar, and above 3 as too similar.

[5] gpt-4-0613: https://platform.openai.com/docs/models/overview

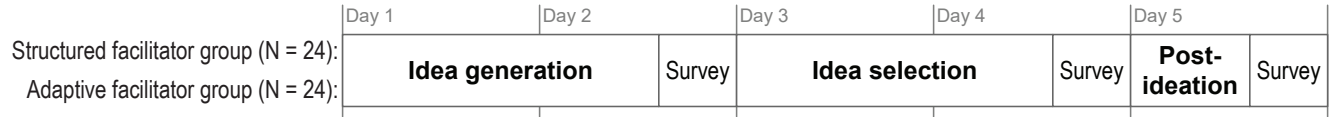


Figure 6: The structure of our user study. Two groups of participants performed idea generation and selection by interacting with the structured or adaptive facilitators. They reviewed the selected ideas on the last day. Participants shared their perceptions of chatbots' facilitation on a survey after each ideation phase.

5.2 Multi-Armed Bandit

We designed the multi-armed bandit (MAB) system for the adaptive facilitator's behaviors. Its system structure is shown in Figure 5. The name originates from a context where a gambler visits a casino to play slot machines, so-called one-armed bandits. By playing one machine (action) at a time (trial), the gambler wins a certain amount of payout (reward). The gambler's goal is to win as much reward as possible within the finite number of trials by selecting the machine that gives the highest reward. Since the gambler does not know which one gives more rewards, they need to try out different slot machines to find the most rewarding one (exploration) and mostly play the best one (exploitation), while balancing how much they should explore and exploit within the limited trials.

Our adaptive facilitator's objective aligns well with MAB. The facilitator tries one dialogue (action) at a time and receives a user reply (reward). Within the limited duration of the conversation, the facilitator needs to identify which dialogue is most helpful to the users (exploration) and perform the most helpful one (exploitation).

We use the Upper Confidence Bound (UCB) algorithm [67] for our MAB system. At every trial, the algorithm computes two types of information for each action: The estimated reward, representing *how much reward the system would get by selecting the action in the next trial*, and the uncertainty, representing *how much the rewards vary for an action across its trials*. Based on the UCB formulation, even if the MAB system exploits the most rewarding action, it will eventually select another because of its high uncertainty. A chatbot interaction guided by the MAB looks like this:

(1) the MAB system selects one of the action (e.g., present similar ideas and request a new idea);
(2) the chatbot performs the action;
(3) the user responds (e.g., generates and rates an idea);
(4) the chatbot receives a certain reward based on the quality of the user's response;
(5) the MAB system computes the estimated reward and the uncertainty of each action;
(6) the MAB system selects the next action that has the highest mean reward and uncertainty.

We evaluated how our MAB systems would explore and exploit during each collaborative ideation phase in a simulated environment. We saw that the MAB chose the actions in a reasonable way, quickly identifying a suitable inspiration method (exploitation) but also changing the inspiration method if it became less beneficial (exploration), as if the user starts generating less helpful ideas from that type of inspiration. The details of our MAB system and its evaluation are given in Appendix B.

6 EMPIRICAL EVALUATION

We conducted a user study to understand the strengths and weaknesses of our chatbots in facilitating asynchronous ideation. For this, we staged an asynchronous ideation event, where two groups of participants ideated with one of the chatbots at their own comfortable time and place (Figure 6). We set the event goal as exploring ideas that can improve people's daily life during a pandemic and recruited participants who were interested in the subject.

The ideation event consisted of three phases: idea generation, idea selection, and post-ideation. We held each idea generation and selection for two days to simulate the conventional asynchronous ideation held for multiple days, allowing participants to ideate at varying hours. The post-ideation was for showing the results (three ideas with the highest ratings) to the participants and collecting their responses to the results. Throughout the event, participants interacted with chatbots only (i.e., our chatbots mediated the idea exchange among the participants).

6.1 Apparatus

We used the Telegram messaging platform to develop our chatbots (Figure 2). Due to its popularity[6], we could easily deploy our chatbots and avoid making participants learn a new interface. We implemented our chatbots using Python and connected to the Telegram platform through its API[7]. The participants could interact with the chatbots via the Telegram application on their PC or smartphone, using their personal accounts.

In addition to the regular message input field, we implemented response buttons for participants to easily request the chatbots to show other ideas, pause and resume the interaction, and rate ideas. This also enabled the chatbots to collect users' ratings on ideas between 1-7. To identify the top three ideas in the end, we implemented a feature that computes a grand score of each idea based on its mean ratings and certainty (i.e., standard error of the mean ratings). In this approach, the chatbots avoided selecting ideas that only had a couple of maximum ratings.

6.2 Participants

We recruited 48 participants via social network (Mean age = 27.21, SD = 4.95, self-identified as men = 21 and women = 27), who had experienced any type of collaborative ideation (e.g., group discussion). The participants varied in their backgrounds such as teachers, software developers, researchers, and nurses. We randomly assigned half of them to interact with either the structured or adaptive facilitator, considering gender balance. For spending minimum 100

[6]https://telegram.org/faq#q-what-is-telegram-what-do-i-do-here
[7]https://github.com/python-telegram-bot/

minutes over five days, we compensated each participant with a 30-euro voucher. All participants received the same amount, unrelated to the number of ideas they generated or reviewed.

6.3 Task

For each day, participants chose when and for how long (minimum 20 minutes in total) to interact with their chatbot facilitators. They were free to pause and resume the interaction as they wanted. Following our chatbots' guidance, participants aimed to generate and evaluate as many ideas as possible. During the post-ideation, participants reviewed the three selected ideas and reported their satisfaction on a survey. At the end of each ideation phase, participants completed a survey to measure the usefulness of chatbots and comment on their facilitation.

6.4 Measurements

To understand the potential of chatbot facilitators, we observed 1) participants' interaction behavior with the chatbots, 2) performance in ideation, 3) perception of the chatbots' facilitators, and 4) satisfaction with the end result. For this, we recorded the moments when participants generated and selected ideas as well as the perceived quality of ideas. We measured participants' perceptions of the chatbots using the system usability scale (SUS) [2] and creative support index (CSI) [7], which also measures the helpfulness of systems on collaborative activities. We collected participants' comments using open-ended questions, inquiring about what they liked and disliked throughout the asynchronous ideation. Lastly, we measured participants' satisfaction with the selected ideas with a survey inspired by the guidelines in consensus-building [9, 66]. Even if participants believe that the selected ideas are not good enough, they can still be willing to commit to the group decisions. This is considered a successful group effort. Accordingly, we adopt the questionnaires tested in the previous study about consensus building [62], which are answered with a 7-point Likert scale (Figure 11, bottom).

6.5 Procedure

We conducted the entire study online. Prior to the study, participants were given a manual for interacting with the chatbot facilitators and following the overall study schedule. On each morning, research moderators reminded participants about the ideation via email. They sent another reminder an hour before the end of each ideation phase, informing participants to wrap up and complete the survey. All other ideation-related facilitation was led by the chatbot. During the post-ideation phase, the chatbots informed that three ideas with the highest ratings were selected and presented them with exemplary opinions from three categories (support, neutral, and against). At the end of each phase, the chatbots conducted the surveys by explicitly informing participants to answer only based on their experience during the ideation, not considering the survey as a part of the interaction.

7 Results

The participants reported both chatbot facilitators to be helpful in generating and selecting ideas. Each group of participants generated 474 and 395 ideas by interacting with the structured facilitator (i.e., structured group) and the adaptive facilitator (i.e., adaptive group), respectively (Figure 7). Half of the participants from the structured group and nine participants from the adaptive group generated more than 20 ideas in total. The chatbots filtered out too similar, repetitive ideas and prompted the structured group to review 311 ideas and the adaptive group to review 200 ideas. Each participant reviewed between 21 and 68 ideas.

The top three ideas from the structured group were *"More widely available cheap solutions to grow food indoors", "Some kind of online collaborative art project would cheer up people", and "More communitarian activities within different parts of town to improve local services and advance a strengthened sense of togetherness". The adaptive group's ideas were "Development of more flexible, comfortable tools to test if we caught some type of flu would be better", "Moisturizer that sanitizes your hands at the same time", and "A simple website that has the rules and guidelines by officials that are effective now and a bot to answer people's questions".*

7.1 Interaction with the Chatbot Facilitators

To understand how participants collectively ideated over multiple days, we plotted their moments of interactions on a timeline (Figure 8). The data shows that both chatbot facilitators were interacting with some participants almost all the time. The individuals' data shows that they were indeed contributing their ideas on their own time, pausing and resuming their ideation. For instance, P24 initiated the idea generation around 4 pm, shared ten ideas, took a break for 10 minutes, and shared three more ideas in 10-minute intervals. He came back to the chatbot after an hour to share four more ideas on the first day. The data also shows an increasing number of notable ideas proportional to the number of total ideas. Accordingly, we conclude that asynchronous ideation was achieved via the chatbot facilitators.

7.2 Idea Generation

The goal of idea generation was creating as many ideas as possible by proposing new and improved ideas. To observe potential trade-offs between the chatbots' facilitation, we compared the results between the structured and adaptive group (between-group) and between the days (within-group). For this, we performed two-by-two factorial analysis, performing Mixed ANOVA ($p < 0.05$) on IBM SPSS statistics software.

7.2.1 Ideas similarity. We analyzed idea similarity by comparing all ideas to a single idea, which we added an initial inspiration (*"A clear mask that allows people to see your expressions"*). The structured and

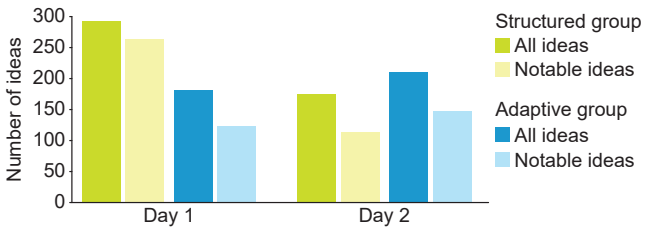


Figure 7: Number of all and notable ideas generated during the idea generation phase. Ideas that were not filtered out by our chatbots were considered notable.

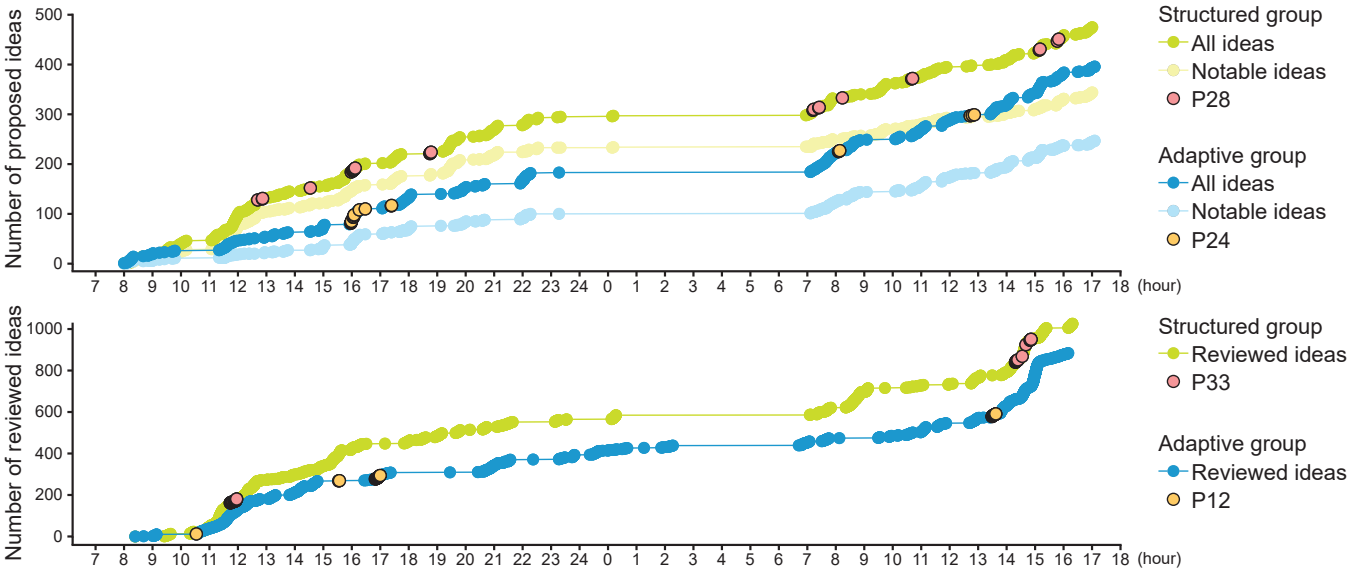


Figure 8: The moments that participants proposed ideas (top) and selected ideas (bottom). We highlight exemplary participants' interaction data. The data shows the participants' asynchronous collective effort spent at their own comfortable time.

adaptive group's mean idea similarity was 0.93 (SD = 0.4) and 1.41 (SD = 0.4), respectively (Figure 9, left). According to our semantic similarity classifier, both groups mostly generated *dissimilar* ideas (i.e., similarity score below 2).

The statistical analysis shows that there was a significant main effect of chatbot facilitators on the similarity of ideas ($F(1, 358) = 60.93$, $p < 0.01$). This indicates that if we ignore on which day the ideas are generated, the structured group generated more dissimilar ideas. There was no significant main effect of days ($F(1, 358) = 0.26$, $p = 0.61$)). Yet, there was a significant interaction between the types of chatbots and the days ($F(1, 358) = 6.01$, $p = 0.02$). The similarity distribution shows that the structured group generated less dissimilar ideas on the second day while the adaptive group generated more dissimilar ideas on the second day.

7.2.2 Self-rating on the generated ideas. To focus participants' attention on the subject, both chatbots asked the participants to rate the helpfulness of their ideas in achieving the ideation goal (Figure 9, right). The statistical analysis shows no significant main effect of chatbot facilitators ($F(1, 358) = 1.93$, $p = 0.17$) or the interaction effect between the chatbot types and the days ($F(1, 358) = 0.92$, $p = 0.34$) Yet, there was a significant main effect of days on the self-ratings ($F(1, 358) = 5.69$, $p = 0.02$), which indicates that both groups considered their ideas more helpful on the second day.

7.3 User Ratings

The results of CSI and SUS are shown in Figure 10, grouped by the ideation phases. We assumed that the participants' perception of the chatbots could be changed based on their satisfaction with the end result. Accordingly, we compared the user response between the chatbots (between-group) and between the ideation phases (within-group), performing a two-by-three factorial analysis.

7.3.1 Creativity Support Index. The mean CSI score of the structured facilitator was 70.36 (SD = 2.67), 67.82 (SD = 3.28), and 68.86 (SD = 3.13) for idea generation, idea selection, and post-ideation, respectively. The mean CSI score of the adaptive facilitator was 64.80 (SD = 2.73), 66.13 (SD = 3.35), and 67.88 (SD = 3.20). No statistically significant effects were observed across all comparisons, which indicates that participants' perception of chatbots did not change throughout the ideation or differed by the chatbots. Cherry and Latulipe [7] state that the CSI score below 50 means that the system does not support creative work and above 90 means excellent support. Accordingly, we conclude that both chatbots well facilitated the participants' creativity in the collaborative ideation.

7.3.2 System Usability Scale. The mean SUS of the structured facilitator was 77.92 (SD = 2.56), 73.65 (SD = 3.44), and 75.73 (SD = 2.98) for each phase. The adaptive facilitator's mean SUS was 68.04 (SD = 2.61), 74.67 (SD = 3.51), and 73.91 (SD = 3.04). The same as the CSI result, there was no statistical significant effects across all comparisons. Brooke [2] states that the SUS score above 68 means above-average usability. Therefore, we conclude that both chatbots were considerably usable for asynchronous ideation.

7.3.3 Result Satisfaction. The participants' satisfaction with the selected ideas is shown in Figure 11, bottom. Both groups expressed above-average satisfaction. Observing individuals' responses revealed that six and five out of 24 participants from the structured and adaptive group disliked the selected ideas. We performed one-way ANOVA to compare the user response between the groups. There was no significant effect of the chatbots in all survey items.

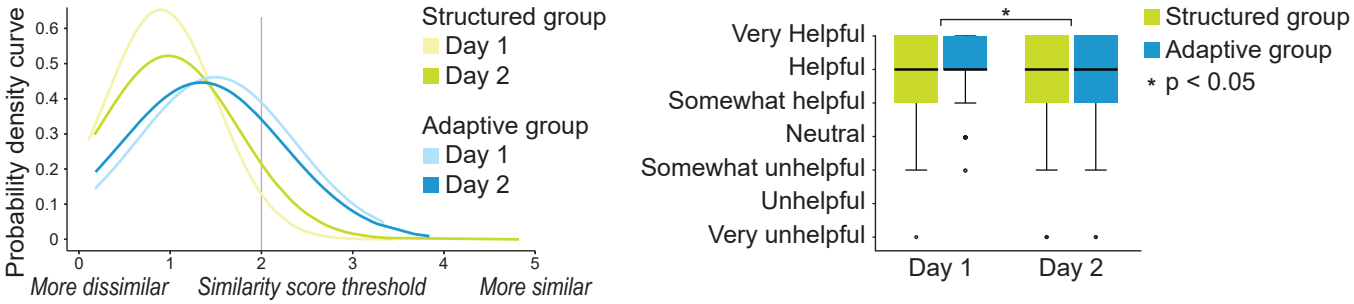


Figure 9: The similarity of ideas that participants generated by interacting with the structured and adaptive facilitators on the first and the second day (left) and the participants' self-rating of own ideas (right).

7.4 Experience of Chatbot Facilitators

We analyzed the participants' comments via affinity diagraming [47]. Despite the differences between the ideation phases and the chatbots, both groups of participants shared similar positive and negative comments. Below, we grouped them by our chatbot designs.

7.4.1 Inspiration. The participants mostly commented on the inspirations brought by the chatbots. They considered reading others' ideas and opinions helped them diversify ideas, propose more creative ideas, and more correctly evaluate ideas. For instance, P27 commented, *"It showed ideas from other collaborators and spiked my creativity."* Likewise, P48 commented that the dissimilar ideas inspired new ideas that he would not have considered. Based on such responses, we confirm that the chatbots facilitated participants to build on each other's contributions in the asynchronous setting.

Yet, negative aspects were also reported that building on the others' ideas was difficult when they were written unclearly. P34 commented, *"Sometimes the participants didn't express or describe their ideas properly. If there was someone modifying the errors, then show to other participants, that would be better."* Other participants (P2 and 25) reported that some opinions were out of context (e.g., *"I agree with the second opinion."*) and stunted their own thought processes.

7.4.2 Ideation method. The participants reported that the ideation method made them think more productively with a specific aim. For instance, P36 commented that the chatbot's suggestion guided her to improve the ideas rather than quickly criticizing them. Likewise, P21 commented, *"I particularly liked how I could give my opinion on an idea before hearing others opinion... I could correct myself with the opinions from the group."* No negative responses were reported about the chatbots suggesting ideation method.

7.5 Experience of Chatbot-Facilitated Asynchronous Ideation

The analysis also revealed the strengths and weaknesses of asynchronous ideation facilitated by the chatbots. Below, we present the five themes.

7.5.1 Anonymity. The participants appreciated that the collaborative ideation was performed anonymously. They reported that being unaware of who the others were encouraged them to share even less practical but more creative ideas, without worrying about the others' criticism. The most representative response was from P28, *"I liked that the ideation was anonymous so it was easier to suggest even slightly odd ideas."* The anonymity was also considered helpful during idea selection. The participants reported that they could evaluate the ideas with more objective viewpoints, instead of hiding their true opinions to not offend the others. For instance, P20 commented that collaboration without being scared of hurting someone's feelings was helpful. After seeing the selected ideas, P1 remarked, *"This tool was very democratic in showing ideas, which helped developing ideas without being biased by whom it came from."*

7.5.2 Asynchrony. For all participants, this was their first time attending asynchronous ideation. In response, the participants acknowledged the benefit of asynchronous settings and the chatbot faciliation. They reported that the chatbots enabled contribution at their own comfortable hours and pace. P19 commented, *"It was easy to share ideas, it gave me a chance to work and collaborate with my preferred times and rhythm."* The others also remarked that they could spend as many hours as they needed for generating ideas and appreciated that they did not have to come up with all ideas in a single meeting.

However, the participants reported that the lack of immediate feedback lowered the immersiveness of collaborative ideation. P40 commented that he would have enjoyed the idea generation more if there was profound connection with the other group members. Similarly, P47 expressed the urge for meeting the people who generated the ideas that he liked. Another limitation was the unequal amount of available inspiration between the participants. In the beginning of each ideation phase, there were relatively fewer ideas and opinions. In response, 11 participants reported seeing the same inspirations repetitively, which demotivated their ideation.

7.5.3 Accessibility. Most participants interacted with the chatbots via their smartphones, except for seven participants who used laptops. In response, the participants reported that they ideated in diverse places such as a couch, bed, desk, park, and public transportation systems. Since the participants could interact with the chatbots at any place and time, they acknowledged the chatbots' readiness for collecting ideas and inspiring their ideation. P25 and P4 commented that they could swiftly interact with other group

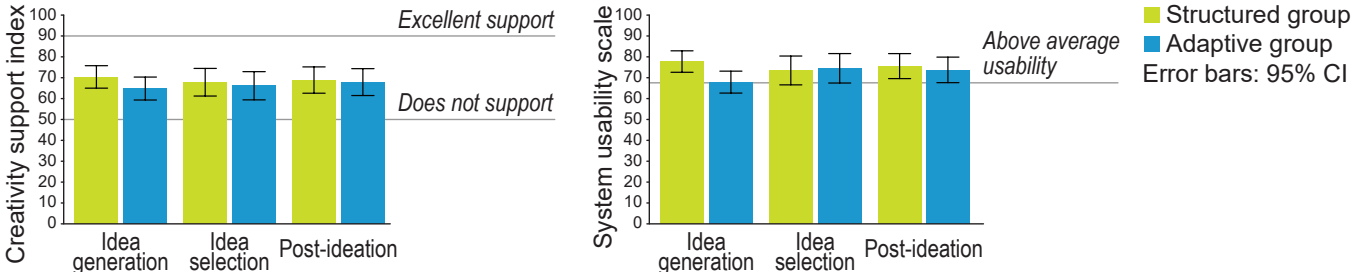


Figure 10: User response on the Creativity Support Index (left) and System Usability Scale (right).

members' opinions. P1 commented, *"I liked that the bot was constantly present. I did not have to wait long to see other's ideas."* Likewise, P27 reported that being able to track his ideas whenever he needed was beneficial to resume the ideation.

7.5.4 Text only ideation. Mostly negative responses were reported about the text-only ideation. The participants commented difficulties in describing and developing ideas without any other modalities such as drawings or video. P2 commented, *"I like brainstorming visually, so writing purely by text was a little stiff at times."* Likewise, P25 and 9 commented that understanding others' ideas and thought processes only based on their writings was difficult.

7.5.5 Non-human facilitator. The participants highlighted the benefits of talking to the chatbots, instead of talking to the other group members or human faciliator. For instance, they appreciated having no social pressure in making the 'listener' wait and the chatbots' polite responses regardless of what they contribute. At the same time, the participants shared the limitations of being a chatbot such as the lack of expression, repetitive dialogue, and the shallow conversation. These attributes were considered to make the interaction boring, lowering their commitment to the ideation. P35 commented that his perception of interacting with the chatbot made him less serious during the ideation.

8 EXPERT EVALUATION

The empirical evaluation study revealed the pros and cons of the chatbots' facilitation from the collaborators' perspective. To extend the findings, we conducted an expert evaluation study to explore chatbots' potential from human facilitators' viewpoints. For this, we recruited an expert with a unique profile who has studied and facilitated collaborative ideation for over 25 years (age = 49, self-identified as a woman). In particular, the expert has experience in facilitating collaborative ideation events with hundreds of participants both in in-person and asynchronous settings, which can provide critical insights about human facilitators' challenges and the design of chatbot facilitators.

To prime the expert for reviewing the chatbots' facilitation, we had the expert first facilitate an ideation workshop independently, without using chatbots. The workshop was held for an hour with 11 participants (Mean age = 27.18, SD = 7.55, self-identified as men = 8 and women = 4), following the same ideation structure and goal as our chatbots facilitated. Afterward, we demonstrated our chatbots and conducted an interview to collect the expert's perspectives. We also recorded the participants' ideas and their satisfaction with them using the same survey in the earlier study. We used them as a baseline to further analyze the chatbot-facilitated ideation.

The comparisons between the chatbot- and human-facilitated groups are shown in Figure 11. We compared the *proportion of notable ideas* using the chi-square test of independence with a Bonferroni correction. We compare the groups' *idea similarity* and *satisfaction with the selected ideas* using the Kruskal-Walis H test with Dunn's test as a post-hoc analysis. Below is the summary of the results:

- Both chatbot-facilitated groups' proportion of notable ideas was statistically similar to the human-facilitated group's (both $p > 0.1$).
- The structured group's idea similarity was statistically similar to the human-facilitated group's ($p = 0.80$), while the adaptive group's idea similarity was statistically higher, indicating less idea diversification ($p < 0.01$).
- The adaptive group's satisfaction with selected ideas was statistically similar to the human-facilitated group's, except for their preference for the selected ideas. Both chatbot-facilitated groups had statistically lower preferences than the human-facilitated group (both $p < 0.01$).

The results suggest that the structured and adaptive facilitator has their own strengths, achieving idea diversification and satisfaction similar to human-facilitated ideation, respectively.

The interview with the expert highlighted pain points in facilitating multiple groups of collaborators. The expert described that human facilitators get tired throughout facilitation, making it challenging to pay closer attention to individuals. The expert added that she also did not have enough energy to take care of less active participants during this workshop. Yet, hiring more trained facilitators is expensive.

Accordingly, the expert remarked on the potential hybrid facilitation between human and chatbot facilitators that complement each other's limitations. She remarked on the capability of chatbot facilitators for (i) cross-pollinating ideas, (ii) supporting ideation with facts, and (iii) moderating progress. While ideating with the participants, the expert frequently referenced ideas from the other group as inspiration. She noted that chatbots could have supported this process. She said, *"I tried to cross-pollinate ideas by bringing ideas between groups, but I could not do that systematically because I cannot hold all of them in my head. This is something that the chatbots can do."* In addition, she commented that there were moments when knowing more facts about certain ideas could have helped the participant improve ideas, which chatbots could proactively provide.

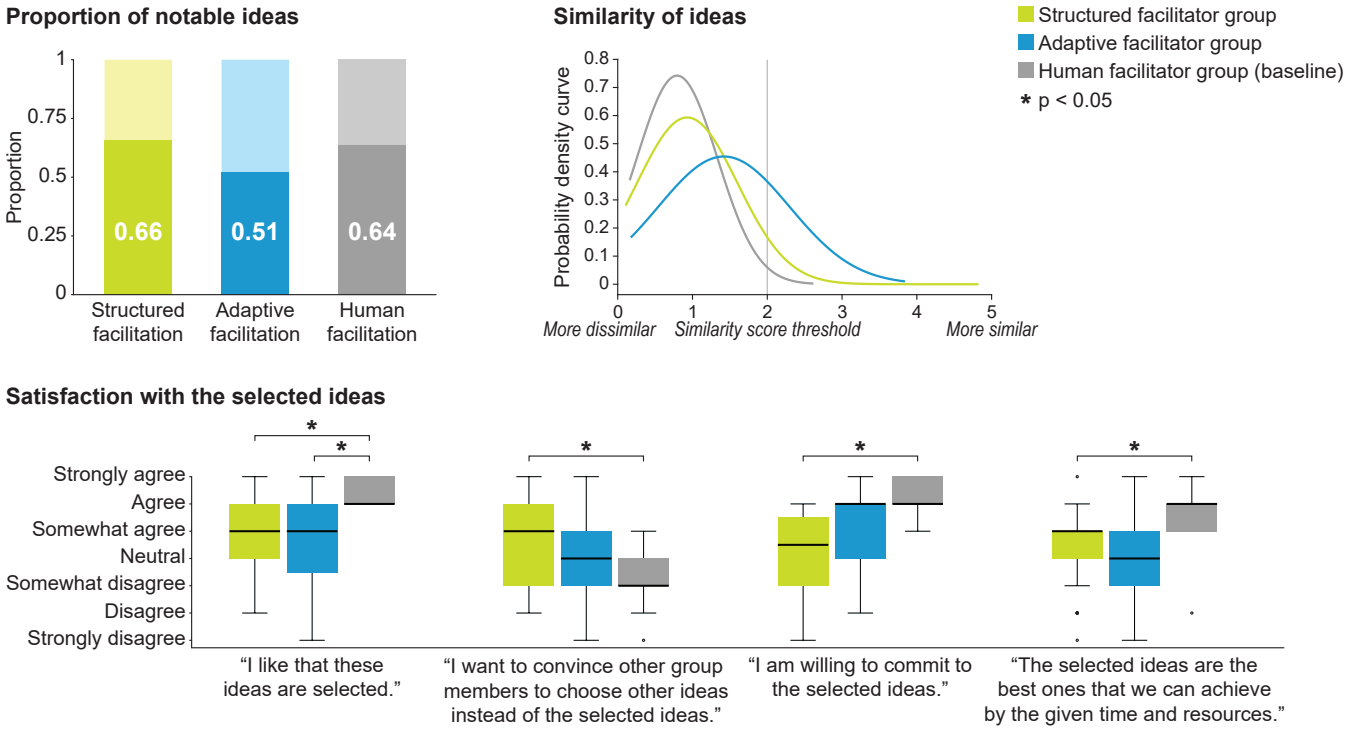


Figure 11: The comparisons between the human-facilitated and chatbot-facilitated ideation.

We also observed that the participants often searched for images on their smartphones to enrich their discussion. Lastly, she expected the chatbot to moderate the progress of ideation with prompts. She commented that chatbots could help teams understand where they are in the ideation process and keep them on track, preventing participants from just focusing on their own ideas.

The expert also highlighted the shortcomings of chatbot facilitators including (a) the lack of embodied interaction, (b) experience sharing, and (c) multi-modality. First, the expert remarked that facilitation is a performance and collaboration does not only happen at the idea level. She added that social interactions during collaborative ideation demand the whole body and spatial aspects that chatbots cannot replace. Second, the human facilitator commented that anonymous collaboration through chatbots could demotivate participants. She remarked that people value meeting others and sharing experiences, which strongly motivates them to participate in collaborative ideation events. Lastly, the human facilitator pointed out text-only ideation. She commented that ideation using only text is feasible in the early stage, but it would limit the development of concrete subjects or prototypes.

9 DISCUSSION

Developing conversational agents that adopt human behaviors has become more accessible thanks to LLMs [52, 60]. Still, this requires an understanding of what could be adopted from humans, instructing (e.g., prompting) behaviors that make the best out of conversational agents. In this paper, we make a case for facilitating asynchronous ideation by leveraging conversational agents. We presented two designs of chatbot facilitators based on the literature on human facilitators. We investigated their potential from the viewpoints of collaborators (Study 1) and an expert facilitator (Study 2), as shown in Table 2. Study 1 showed that both chatbots enable asynchronous ideation among collaborators, allowing them to build on each other's contributions at their own pace. Study 2 suggests that the structured facilitator has an advantage in diversifying ideas and the adaptive facilitator can help collaborators choose satisfying ideas similar to human-facilitated ideation. However, the two studies also exposed a shortcoming of chatbot facilitators: they do not moderate social interaction among participants more broadly. In the following, we elaborate on this finding and reflect on its implications.

9.1 Applying Chatbots in Collaborative Activities

Our studies suggest that chatbots can facilitate collaborative ideation in asynchronous settings. We propose three practical suggestions on how to deploy chatbot facilitators to benefit collaborative efforts in creative projects such as design. First, we suggest using *chatbots for aligning collaboration efforts*. In collaborative activities, it is essential that group members aim for the same ideation goal. However, human facilitators can experience challenges in closely engaging with individuals and redirecting their attention. Instead, as we observed in Study 1, chatbots can engage with all members to focus on joint goals and get them to reflect on their own contributions as a group.

Chatbot-facilitated asynchronous ideation	Positive	Negative
Showing inspirations	Building on each other's contribution	Unclearly written ideas and opinions
Suggesting ideation methods	Avoid early criticism and be more productive	-
Requesting ratings	-	-
Anonymity	Objective ideation unbiased by social interaction	Difficult to socially bond with the others
Asynchrony	Ideation at comfortable time and pace	Lack of immediate feedback from the others
Accessibility	Propose ideas as soon as they appear	-
Text only	-	Difficult to ideate and describe ideas
Non-human facilitator	No stress by talking to a person	Repetitive dialogue

Table 2: Summary of our findings on the positives and negatives of asynchronous ideation facilitated by chatbots.

Second, we suggest using *chatbots for guiding ideation toward unexplored subjects*. To diversify ideas, collaborators need to know what kinds of ideas they have created as a group. However, reviewing all generated ideas and identifying unexplored subjects is laborious even for multiple human facilitators. Instead, chatbots with advanced NLP methods could cluster ideas and identify less discussed subjects. Then, similar to our chatbots suggesting inspirations, chatbots could prioritize overlooked subjects to promote effective idea exploration.

Third and last, we suggest using *chatbots for moderating collaborators understanding of each other*. Study 2 showed that collaborative ideation is not just generating or selecting ideas together. We observed that the participants actively shared their personal experiences to empathize with each other and convey their perspectives. Our study showed that anonymous idea exchange through chatbots is helpful in sharing personal thoughts without worrying about criticism. Previous studies also show that users are more willing to share their personal stories with chatbots than other human beings, partially because chatbots appear less judgmental [10, 44]. This could help collaborators share their thoughts more directly and understand each other more correctly.

9.2 Understanding Differences of Chatbot Facilitators

Our studies provide directions for adapting human facilitators' behaviors to chatbot facilitators. We learned that chatbots can lead structured facilitation to the generation of more diverse ideas. In particular, the structured facilitator guided the participants to think about new ideas on the first day only. In response, the participants could have focused on diversification without the need to consider their quality. In contrast, the adaptive facilitator suggested both diversifying and improving ideas in a single day. This might have encouraged the participants to consider both the dissimilarity and quality of ideas, overburdening their ideation process. This assumption aligns well with guidelines of effective brainstorming, having a single purpose of ideation at a time and suspending criticism [54, 56, 68]. We also learned that chatbots can facilitate adaptively, achieving a similar level of satisfaction with final ideas as human-facilitated ideation. We assume that guiding collaborators to review ideas with uncertain group opinions helped them strengthen their consensus. Potentially, combining both structured and adaptive facilitation into a single chatbot could be more effective, guiding one ideation method per idea generation and adaptively prioritizing ideas to review during idea selection.

9.3 Toward Hybrid Facilitation

Based on our results, we suggest directions toward a 'hybrid model' of human–AI facilitation that exploits their unique strengths. First, we note that chatbot facilitators can be in a secondary role (an assistant) to a human facilitator. In Study 2, the human facilitator reported the moments when she could have applied an extra facilitator. She reported that moderating more than one group of participants was challenging and she often forgot to encourage less active participants. We expect chatbots to provide such support. For instance, while a human facilitator engages with one group of participants, chatbots could mimic the human facilitator's dialogue and guide the rest of the participants. Moreover, it is vital for the human facilitator to retain leadership and accountability. Second, the techniques presented for chatbot facilitation so far in the literature suggest that chatbots are best when deployed in collaborative ideation. Chatbots can help human facilitators to monitor the progress and learn from it. Human facilitators could also use this period to focus on designing the next stage of collaborative ideation, identifying opportunities to include collaborators with different perspectives or exploring other areas of the design space [64]. Lastly, chatbot facilitators could help human facilitators better understand the participants after the first round. Chatbot facilitators can identify the characteristics or preferences of participants for human facilitators by continuously engaging with each individual collaborator throughout collaborative ideation activities.

9.4 Limitations and Future Work

We report four limitations in our study and propose corresponding future work. First, we focused on facilitating asynchronous ideation in the context of idea generation and selection. Whereas diversifying and converging ideas are the fundamental activities in the ideation process, how chatbots can facilitate other collaborative scenarios, such as problem-solving and deliberate discussion, in asynchronous settings remains unexplored. Future work could look into the different facilitation requirements that arise in different scenarios and examine how chatbots could adopt them. Second, we studied chatbots that used text as the medium. In Study 1, four participants reported that ideating and describing ideas using only

text was challenging. In the future, chatbots could facilitate multimodal communication. Chatbots are already capable of integrating sketches, images, videos, and audio into dialogue. Recent advances in multi-modal machine learning could provide the basis for chatbot facilitators to understand users' non-text contributions [57]. Third, our expert evaluation study is based on a single case study. While we recruited the expert with unique profile, we acknowledge that our expert's viewpoints may not fully generalize to other facilitators. To address this, future work could include experts with diverse backgrounds and competencies to uncover the broader perspectives on using chatbots as facilitators. Last, we studied asynchronous and anonymous ideation. Despite our findings on the benefits of anonymity, having no direct communication and not knowing each other could demotivate collaborators. Future work needs to take this into account and study artificial facilitators that can also promote social interaction among collaborators.

10 CONCLUSION

In this paper, we designed chatbot facilitators to guide asynchronous idea generation and selection among collaborators. We designed the structured and adaptive facilitators by adopting the guidelines found in the literature about human facilitators. Our structured facilitator guided individuals by providing a structured ideation process and our adaptive facilitator guided them by adapting to their ideation performance. Our studies suggest the strengths and limitations of chatbot facilitators. Both chatbots were found to be helpful, especially when helping collaborators build on each other's contributions. Each structured and adaptive facilitator had their strengths in diversifying ideas and achieving satisfaction with selected ideas similar to human-facilitated ideation. With our findings, we discuss the potential of a 'hybrid model' where human and chatbot facilitators complement each other's strengths. We conclude that chatbots can be promising alternative facilitators of asynchronous ideation, providing continuous guidance in the absence of human facilitators.

Acknowledgments

This research was supported by the Research Council of Finland grant 357578 (Subjective Functions); the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A6A3A03056886); and IFI program of the German Academic Exchange Service (DAAD).

References

[1] Pernille Viktoria Kathja Andersen and Wafa Said Mosleh. 2021. Conflicts in co-design: engaging with tangible artefacts in multi-stakeholder collaboration. *CoDesign* 17, 4 (2021), 473–492. doi:10.1080/15710882.2020.1740279 arXiv:https://doi.org/10.1080/15710882.2020.1740279

[2] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[3] Vincent R. Brown and Paul B. Paulus. 2002. Making Group Brainstorming More Effective: Recommendations From an Associative Memory Perspective. *Current Directions in Psychological Science* 11, 6 (2002), 208–212. doi:10.1111/1467-8721.00202 arXiv:https://doi.org/10.1111/1467-8721.00202

[4] Hernan Casakin and Georgi V. Georgiev. 2021. Design creativity and the semantic analysis of conversations in the design studio. *International Journal of Design Creativity and Innovation* 9, 1 (2021), 61–77. doi:10.1080/21650349.2020.1838331 arXiv:https://doi.org/10.1080/21650349.2020.1838331

[5] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. doi:10.18653/v1/S17-2001

[6] Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of Semantic Similarity—A Survey. *ACM Comput. Surv.* 54, 2, Article 41 (feb 2021), 37 pages. doi:10.1145/3440755

[7] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.* 21, 4, Article 21 (jun 2014), 25 pages. doi:10.1145/2617588

[8] Aeran Choi and Brian Hand. 2020. Students' construct and critique of claims and evidence through online asynchronous discussion combined with in-class discussion. *International Journal of Science and Mathematics Education* 18, 6 (2020), 1023–1040. doi:10.1007/s10763-019-10005-4

[9] Peter T Coleman, Morton Deutsch, and Eric C Marcus. 2014. *The handbook of conflict resolution: Theory and practice*. John Wiley & Sons.

[10] Christopher Collins, Simone Arbour, Nathan Beals, Shawn Yama, Jennifer Laffier, and Zixin Zhao. 2022. Covid Connect: Chat-Driven Anonymous Story-Sharing for Peer Support. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 301–318. doi:10.1145/3532106.3533545

[11] Lauren E. Coursey, Belinda C. Williams, Jared B. Kenworthy, Paul B. Paulus, and Simona Doboli. 2020. Divergent and Convergent Group Creativity in an Asynchronous Online Environment. *The Journal of Creative Behavior* 54, 2 (2020), 253–266. doi:10.1002/jocb.363 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.363

[12] Yngve Dahl and Kshitij Sharma. 2022. Six Facets of Facilitation: Participatory Design Facilitators' Perspectives on Their Role and Its Realization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 484, 14 pages. doi:10.1145/3491102.3502013

[13] Aaron Davis, Wallace Niki, Joseph Langley, and Gwilt Ian. 2021. Low-contact co-design: considering more flexible spatiotemporal models for the co-design workshop. *Strategic Design Research Journal* 14, 1 (2021).

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. doi:10.18653/v1/n19-1423

[15] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-Based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 623–627. doi:10.1145/3527927.3535197

[16] Haakon Faste, Nir Rachmel, Russell Essary, and Evan Sheehan. 2013. Brainstorm, Chainstorm, Cheatstorm, Tweetstorm: New Ideation Strategies for Distributed HCI Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1343–1352. doi:10.1145/2470654.2466177

[17] Stefano Filippi. 2023. Measuring the Impact of ChatGPT on Fostering Concept Generation in Innovative Product Design. *Electronics* 12, 16 (2023). doi:10.3390/electronics12163535

[18] Rosendy Galabo, Badziili Nthubu, Leon Cruickshank, and David Perez. 2020. Redesigning a workshop from physical to digital: Principles for designing distributed co-design approaches. (2020).

[19] Joseph Gatto, Omar Sharif, Parker Seegmiller, Philip Bohlman, and Sarah Masud Preum. 2023. Text Encoders Lack Knowledge: Leveraging Generative LLMs for Domain-Specific Semantic Textual Similarity. arXiv:2309.06541 [cs.CL]

[20] Chiara Del Gaudio, Carlo Franzato, and Alfredo Jefferson de Oliveira. 2020. Co-design for democratising and its risks for democracy. *CoDesign* 16, 3 (2020), 202–219. doi:10.1080/15710882.2018.1557693 arXiv:https://doi.org/10.1080/15710882.2018.1557693

[21] Elisa Giaccardi, Pedro Paredes, Paloma Díaz, and Diego Alvarado. 2012. Embodied Narratives: A Performative Co-Design Technique. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2317956.2317958

[22] Yoshiko Goda, Masanori Yamada, Hideya Matsukawa, Kojiro Hata, and Seisuke Yasunami. 2014. Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education* 13, 1 (2014), 1–7. doi:10.12937/ejsise.13.1

[23] Rafik Hadfi, Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. 2021. Argumentative Conversational Agents for Online Discussions. *Journal of Systems Science and Systems Engineering* 30, 4 (01 Aug 2021), 450–464. doi:10.1007/s11518-021-5497-1

[24] Rafik Hadfi, Jawad Haqbeen, Sofia Sahab, and Takayuki Ito. 2021. Argumentative conversational agents for online discussions. *Journal of Systems Science and Systems Engineering* 30, 4 (2021), 450–464. doi:10.1007/s11518-021-5497-1

[25] Kim Halskov and Peter Dalsgård. 2006. Inspiration Card Workshops. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) *(DIS '06)*. Association for Computing Machinery, New York, NY, USA, 2–11. doi:10.1145/1142405.1142409

[26] Jawad Haqbeen, Takayuki Ito, Rafik Hadfi, Tomohiro Nishida, Zoia Sahab, Sofia Sahab, Shafiq Roghmal, and Ramin Amiryar. 2020. Promoting discussion with AI-based facilitation: Urban dialogue with Kabul city. In *Proceedings of the 8th ACM Collective Intelligence, ACM Collective Intelligence Conference Series, Boston (Virtual Conference), South Padre Island, TX, USA*, Vol. 18.

[27] Nichole Harvey and Colin A Holmes. 2012. Nominal group technique: An effective method for obtaining group consensus. *International Journal of Nursing Practice* 18, 2 (2012), 188–194. doi:10.1111/j.1440-172X.2012.02017.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1440-172X.2012.02017.x

[28] Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2023. Z-Code++: A Pre-trained Language Model Optimized for Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 5095–5112. doi:10.18653/v1/2023.acl-long.279

[29] White House. 2019. Petition the White House on the Issues that Matter to You.

[30] Takayuki Ito, Rafik Hadfi, and Shota Suzuki. 2022. An Agent that Facilitates Crowd Discussion. *Group Decision and Negotiation* 31, 3 (01 Jun 2022), 621–647. doi:10.1007/s10726-021-09765-8

[31] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120. doi:10.1073/pnas.2208839120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2208839120

[32] Anu Kankainen, Kirsikka Vaajakallio, Vesa Kantola, and Tuuli Mattelmäki. 2012. Storytelling Group – a co-design method for service design. *Behaviour & Information Technology* 31, 3 (2012), 221–230. doi:10.1080/0144929X.2011.563794 arXiv:https://doi.org/10.1080/0144929X.2011.563794

[33] Alison Kennedy, Catherine Cosgrave, Joanna Macdonald, Kate Gunn, Timo Dietrich, and Susan Brumby. 2021. Translating Co-Design from Face-to-Face to Online: An Australian Primary Producer Project Conducted during COVID-19. *International Journal of Environmental Research and Public Health* 18, 8 (2021). doi:10.3390/ijerph18084147

[34] Finn Kensing and Andreas Munk-Madsen. 1993. PD: Structure in the Toolbox. *Commun. ACM* 36, 6 (1993), 78–85.

[35] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. *Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376785

[36] 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, Article 87 (apr 2021), 26 pages. doi:10.1145/3449161

[37] Scott R. Klemmer, Mark W. Newman, Ryan Farrell, Mark Bilezikjian, and James A. Landay. 2001. The Designers' Outpost: A Tangible Interface for Collaborative Web Site. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (Orlando, Florida) *(UIST '01)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/502348.502350

[38] John Knight, Dan Fitton, Charlie Phillips, and Dylan Price. 2019. Design Thinking for Innovation. Stress Testing Human Factors in Ideation Sessions. *The Design Journal* 22, sup1 (2019), 1929–1939. doi:10.1080/14606925.2019.1594950 arXiv:https://doi.org/10.1080/14606925.2019.1594950

[39] Stefan Werner Knoll and Graham Horton. 2010. Changing the Perspective: Improving Generate thinkLets for Ideation. In *2010 43rd Hawaii International Conference on System Sciences*. 1–10. doi:10.1109/HICSS.2010.103

[40] A. Baki Kocaballi. 2023. Conversational AI-Powered Design: ChatGPT as Designer, User, and Product. arXiv:2302.07406 [cs.HC]

[41] Nicholas W. Kohn, Paul B. Paulus, and YunHee Choi. 2011. Building on the ideas of others: An examination of the idea combination process. *Journal of Experimental Social Psychology* 47, 3 (2011), 554–561. doi:10.1016/j.jesp.2011.01.004

[42] Bishal Lamichhane. 2023. Evaluation of ChatGPT for NLP-based Mental Health Applications. arXiv:2303.15727 [cs.CL]

[43] Franc Lavrič and Andrej Škraba. 2023. Brainstorming Will Never Be the Same Again—A Human Group Supported by Artificial Intelligence. *Machine Learning and Knowledge Extraction* 5, 4 (2023), 1282–1301. doi:10.3390/make5040065

[44] Sung-Chul Lee, Jaeyoon Song, Eun-Young Ko, Seongho Park, Jihee Kim, and Juho Kim. 2020. *SolutionChat: Real-Time Moderator Support for Chat-Based Structured Discussion.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376609

[45] Yanki Lee. 2008. Design participation tactics: the challenges and new roles for designers in the co-design process. *CoDesign* 4, 1 (2008), 31–50. doi:10.1080/15710880701875613 arXiv:https://doi.org/10.1080/15710880701875613

[46] Marcela Litcanu, Octavian Prostean, Cosmin Oros, and Alin Vasile Mnerie. 2015. Brain-Writing Vs. Brainstorming Case Study For Power Engineering Education. *Procedia - Social and Behavioral Sciences* 191 (2015), 387–390. doi:10.1016/j.sbspro.2015.04.452 The Proceedings of 6th World Conference on educational Sciences.

[47] Andrés Lucero. 2015. Using Affinity Diagrams to Evaluate Interactive Prototypes. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 231–248.

[48] Andrés Lucero, Kirsikka Vaajakallio, and Peter Dalsgaard. 2012. The dialogue-labs method: process, space and materials as structuring elements to spark dialogue in co-design events. *CoDesign* 8, 1 (2012), 1–23. doi:10.1080/15710882.2011.609888 arXiv:https://doi.org/10.1080/15710882.2011.609888

[49] Yossi Maaravi, Ben Heller, Yael Shoham, Shay Mohar, and Baruch Deutsch. 2021. Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science* 15, 6 (01 Aug 2021), 1431–1464. doi:10.1007/s11846-020-00400-5

[50] Ali Mazalek, Claudia Winegarden, Tristan Al-Haddad, Susan J. Robinson, and Chih-Sung Wu. 2009. Architales: Physical/Digital Co-Design of an Interactive Story Table. In *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction* (Cambridge, United Kingdom) *(TEI '09)*. Association for Computing Machinery, New York, NY, USA, 241–248. doi:10.1145/1517664.1517716

[51] Nicolas Michinov and Corine Primois. 2005. Improving productivity and creativity in online groups through social comparison process: New evidence for asynchronous electronic brainstorming. *Computers in Human Behavior* 21, 1 (2005), 11–28. doi:10.1016/j.chb.2004.02.004

[52] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

[53] SangAh Park, Yoon Young Lee, Soobin Cho, Minjoon Kim, and Joongseek Lee. 2021. *"Knock Knock, Here Is an Answer from Next Door": Designing a Knowledge Sharing Chatbot to Connect Residents: Community Chatbot Design Case Study.* Association for Computing Machinery, New York, NY, USA, 144–148. https://doi.org/10.1145/3462204.3481738

[54] Paul B Paulus, Jonali Baruah, and Jared B Kenworthy. 2018. Enhancing collaborative ideation in organizations. *Frontiers in psychology* 9 (2018), 2024.

[55] Paul B Paulus, Nicholas W Kohn, and Lauren E Arditti. 2011. Effects of Quantity and Quality Instructions on Brainstorming. *The Journal of Creative Behavior* 45, 1 (2011), 38–46. doi:10.1002/j.2162-6057.2011.tb01083.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2162-6057.2011.tb01083.x

[56] Gerard J. Puccio, Cyndi Burnett, Selcuk Acar, Jo A. Yudess, Molly Holinger, and John F. Cabra. 2020. Creative Problem Solving in Small Groups: The Effects of Creativity Training on Idea Generation, Solution Creativity, and Leadership Effectiveness. *The Journal of Creative Behavior* 54, 2 (2020), 453–471. doi:10.1002/jocb.381 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.381

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[58] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108

[59] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond Dyadic Interactions: Considering Chatbots as Community Members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300680

[60] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (01 Nov 2023), 493–498. doi:10.1038/s41586-023-06647-8

[61] Donghoon Shin, Sangwon Yoon, Soomin Kim, and Joonhwan Lee. 2021. BlahBlahBot: Facilitating Conversation between Strangers Using a Chatbot with ML-Infused Personalized Topic Suggestion. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 409, 6 pages. doi:10.1145/3411763.3451771

[62] Joongi Shin, Michael A. Hedderich, AndréS Lucero, and Antti Oulasvirta. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 78, 13 pages. doi:10.1145/3526113.3545671

[63] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 937–945. doi:10.1145/2675133.2675239

[64] Marc Steen, Menno Manschot, and Nicole De Koning. 2011. Benefits of co-design in service design projects. *International Journal of Design* 5, 2 (2011).
[65] James Surowiecki. 2004. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations* 296, 5 (2004).
[66] Lawrence E Susskind, Sarah McKearnen, and Jennifer Thomas-Lamar. 1999. *The consensus building handbook a comprehensive guide to reaching agreement.* SAGE, Thousand Oaks, Calif. ;.
[67] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.
[68] Katja Thoring and Roland M. Müller. 2011. Understanding the Creative Mechanisms of Design Thinking: An Evolutionary Approach. In *Procedings of the Second Conference on Creativity and Innovation in Design* (Eindhoven, Netherlands) *(DESIRE '11)*. Association for Computing Machinery, New York, NY, USA, 137–147. doi:10.1145/2079216.2079236
[69] Kirsikka Vaajakallio, Jung-Joo Lee, and Tuuli Mattelmäki. 2009. "It Has to Be a Group Work!": Co-Design with Children. In *Proceedings of the 8th International Conference on Interaction Design and Children* (Como, Italy) *(IDC '09)*. Association for Computing Machinery, New York, NY, USA, 246–249. doi:10.1145/1551788.1551843
[70] Andrew Van De and Andre L. Delbecq. 1971. Nominal Versus Interacting Group Processes for Committee Decision-Making Effectiveness. *Academy of Management Journal* 14, 2 (1971), 203–212. doi:10.5465/255307 arXiv:https://doi.org/10.5465/255307
[71] Froukje Sleeswijk Visser, Pieter Jan Stappers, Remko van der Lugt, and Elizabeth B-N Sanders. 2005. Contextmapping: experiences from practice. *CoDesign* 1, 2 (2005), 119–149. doi:10.1080/15710880500135987 arXiv:https://doi.org/10.1080/15710880500135987
[72] Greg Walsh and Eric Wronsky. 2019. AI + Co-Design: Developing a Novel Computer-Supported Approach to Inclusive Design. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 408–412. doi:10.1145/3311957.3359456
[73] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 683, 13 pages. doi:10.1145/3411764.3445781
[74] Bo Xie, Allison Druin, Jerry Fails, Sheri Massey, Evan Golub, Sonia Franckel, and Kiki Schneider. 2012. Connecting generations: developing co-design methods for older adults and children. *Behaviour & Information Technology* 31, 4 (2012), 413–423. doi:10.1080/01449291003793793 arXiv:https://doi.org/10.1080/01449291003793793
[75] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. doi:10.1145/3490099.3511105
[76] Roshanak Zilouchian Moghaddam, Brian P. Bailey, and Christina Poon. 2011. IdeaTracker: An Interactive Visualization Supporting Collaboration and Consensus Building in Online Interface Design Discussions. In *Human-Computer Interaction – INTERACT 2011*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 259–276.

APPENDIX

A SEMANTIC SIMILARITY CLASSIFIER

Using a semantic similarity classifier allows us to compare two user-provided ideas and automatically judge how similar or dissimilar they are. This removes the need for a human facilitator or for the participants to manually check the similarity of ideas and categorize them. Given the speed of the semantic similarity classifier, it also provides the chatbot with near-instant information on the similarity, avoiding long wait times and allowing for a more fluent conversation with the user. For this, we tested a **prompt-based classifier** based on LLM and a **fine-tuned classifier** based on a language model.

We implemented the prompt-based classifier using GPT-4 (gpt-4-0613), the latest model at the time of testing, and followed the prompt design introduced by Getto et al. [19]. The authors evaluated the competence of two LLMs (Llama2-7b and gpt-3.5-turbo-0301) on a semantic textual similarity prediction using different prompt designs. Among them, we followed the best-performing design (**Prompt**: *Output a number between 0 and 1 describing the semantic similarity between the following two sentences: Sentence 1:* $< Text1 >$ *Sentence 2:* $< Text2 >$).

We implemented the fine-tuned classifier by fine-tuning DistilBERT [58]. We chose DistilBERT as it is a compressed version of the popular BERT [14] model and, thus, reduces hardware requirements. We take the "base-uncased" variant and train it on the English portion of the SemEval2017-STS dataset [5]. We train for 5 epochs with a batch size of 32, fp16 precision, and mean-squared-error loss, selecting the weights of the best epoch according to the development set.

A.1 Evaluation of Semantic Similarity Estimation

We evaluated the classifiers by measuring the correlation of their predictions with the manually annotated development and test set of SemEval2017-STS. Fine-tuned classifier achieved a Pearson/Spearman correlation of 87/87% on the development set and of 82/81% on the test set. Prompt-based classifier achieved 84/85% and 80/81%, respectively. While both not beating the state-of-the-art in NLP, this is in range with modern classifiers. Sanh et al., [58] report, e.g., a correlation of 91% on the development set. Accordingly, we conclude that both classifiers can provide high-quality semantic similarity estimation.

A.2 Evaluation of Response Latency

We evaluated our classifiers' response latency by measuring how long it takes them to estimate the similarity between an input sentence and another 100 sentences. This resembles our chatbots' facilitation scenario, where a user submits an idea and the chatbots present the three most similar or dissimilar ideas from all other collaborators. The fine-tuned classifier performs the 100 comparisons in a single run, outputting a set of similarity scores. In the case of the prompt-based classifier, we adjusted the prompt to instruct LLM to fetch the three most similar and dissimilar sentences from 100 sentences in relation to an input sentence (Prompt: *Here is a set of sentences: Sentences:* $< 100 sentences >$ *Select three most similar and*

dissimilar sentences to the sentence below: < 1*sentence* >). We made this adjustment as Gatto et al.'s prompt design can not perform multiple comparisons in a single prompt [19] (i.e., running prompts multiple times takes longer).

We tested each classifier 100 times and computed their mean response latency. The results showed that the fine-tuned classifier is faster (Mean = 2.64 seconds, SD = 0.28) than the prompt-based classifier (Mean = 5.16 seconds, SD = 1.20). Accordingly, we conclude that the fine-tuned classifier is more suitable for our interaction scenario.

A.3 Estimating the Threshold of the Semantic Similarity Classifier

We evaluated how well the fine-tuned classifier works in our idea-generation context and mapped its similarity scale to a categorization that is useful for us (i.e., similar or dissimilar with regard to user-provided ideas). For this;

(1) We generated 100 ideas written in colloquial styles that represent the result of our chatbot interaction (e.g. *"I think more comfortable masks need to be designed"*).
(2) We obtained similarity scores for all sentence pairs from our classifier.
(3) Since the distribution of scores was skewed towards many dissimilar pairs, We randomly sampled 12 pairs from each 0.5 interval between 0 to 3.5 (7 range, 84 pairs in total): We excluded the range above 3.5 as such a score was only reached when the sentences were the same.
(4) 10 participants (Mean age = 24.80, SD = 4.26, 5 males) reviewed the selected pairs in a randomized order and rated their similarity using a 5-point Likert-scale (1 = very dissimilar and 5 = very similar): We excluded the pairs that were labeled as "neutral" and grouped together "very similar" and "similar" into a single "similar" label (and respectively for "dissimilar") since the distinction was not relevant for our use case.
(5) Based on their ratings, we thresholded the classifier's output score. We classified everything below and equal to the threshold as dissimilar and above as similar. We computed agreements between classifier and human labels via accuracy, i.e., the number of annotations where the classifier and human agree / the total number of annotations.
(6) The highest agreement to 'similar' is reached for a threshold of 2.0 with an accuracy of 83%. The highest agreement to 'very similar' is reached for 3.0 with 84%.

Accordingly, our chatbot facilitators presented other collaborators' ideas with a similarity score below 2.0 as dissimilar, above 2.0 as similar, and above 3.0 as very similar, which are filtered out as being considered repetitive ideas.

B MULTI-ARMED BANDIT

We implemented a Multi-Armed Bandit (MAB) system for our adaptive facilitator. The MAB system selects an action (e.g., our chatbot's facilitator behaviors), receives a reward (e.g., users' rating on their own ideas), and selects the next action to try.

B.1 Upper Confidence Bound Algorithm

For the MAB, we used the Upper Confidence Bound (UCB) algorithm:

$$A_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right] \quad (1)$$

Where:

- A_t = action selected at trial t.
- $Q_t(a)$ = estimated reward of action a at trial t.
- $N_t(a)$ = number of times that action a has been selected up to trial t.
- $\log t$ = total number of trials that the MAB system has performed.
- c = constant that controls the level of exploration.

The estimated reward ($Q_t(a)$) is computed with the following formula:

$$Q_t(a) = m_t + \frac{r_t - m_t}{N_t(a)} \quad (2)$$

Where:

- r_t = reward received at trial t.
- m_t = the mean reward of action a up to trial t.

Based on the expressions, the algorithm can be understood as;

> 'for action a, compute the estimated reward ($Q_t(a)$) and the uncertainty ($\sqrt{\frac{\log t}{N_t(a)}}$) for trial t. Compute them for all the actions and select the one (A_t) that has the highest upper confidence bound; the weighted sum of estimated reward and uncertainty ($\arg\max_a \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$)'.

C ACTIONS OF THE ADAPTIVE FACILITATOR

For each idea generation and selection phase, we developed an MAB system according to the adaptive facilitator's behaviors. During the idea generation phase, the adaptive facilitator has four actions as the combination of two inspirations and two ideation methods:

- "Here are other members' ideas *similar* to yours... Can you propose *any* idea?"
- "Here are other members' ideas *similar* to yours... Can you propose an *improved* idea?"
- "Here are other members' ideas *dissimilar* to yours... Can you propose *any* idea?"
- "Here are other members' ideas *dissimilar* to yours... Can you propose an *improved* idea?"

Responding to an action, a user will either generate an idea or request a different inspiration if the user could not generate an idea based on the action. When the user generates an idea, the chatbot asks the user to rate how helpful the idea is for achieving the design goal using 7-point Likert scale. Accordingly, the chatbot receives a reward between 1-7. When the user requests different inspirations, the chatbot automatically receives the lowest reward of 1. This means that the more helpful ideas that the user generates, the higher reward the chatbot receives, which will make the chatbot try that action more than the other actions.

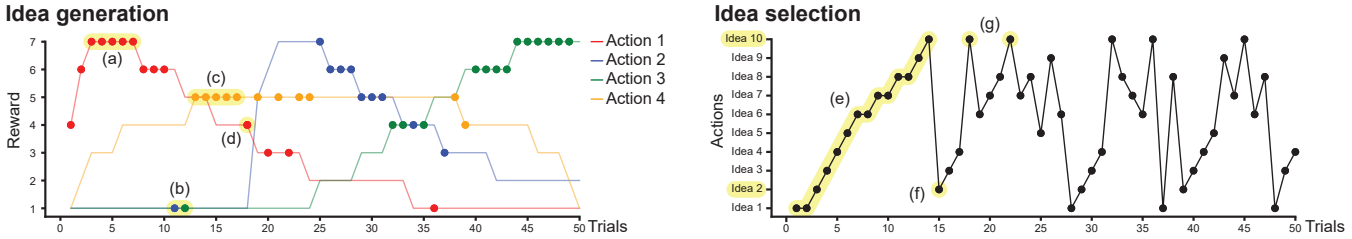


Figure 12: The evaluation of the MAB system in the simulated environment. We observed that the MAB system adaptively finds the most rewarding action during idea generation (left) and adaptively prioritizes the most uncertain ideas during idea selection (right).

During the idea selection phase, the notable ideas collected from the idea generation phase become the actions (Actions = [idea 1, idea 2, idea 3,..., idea n]). When the chatbot presents an idea, users rate the idea using the 7-point Likert scale. The chatbot will then compute how diverse users' opinions are and how many users have rated this idea, which is the standard error of the ratings (more details in Supplementary.C). For instance, if one idea has more diverse and a fewer number of ratings, its standard error will be higher. Accordingly, the chatbot will receive a higher reward and try to show that idea first to the other.

D REWARDS FOR THE ADAPTIVE FACILITATOR'S ACTIONS

During each collaborative ideation phase, our MAB system receives different rewards according to users' ratings:

- **Idea generation**: r_t = a user's rating of their own idea between 1 (very unhelpful) and 7 (very helpful). For instance, if a user;
  - rates their idea 'very helpful' at trial t, the MAB system receives a reward $r_t = 7$.
  - rates their idea 'very unhelpful' at trail t, the MAB system receives a reward $r_t = 1$.
- **Idea selection**: r_t = standard error (SE) of users' ratings between 1 (not interested at all) and 7 (very interested). For instance, idea A has received one user's rating $\{7\}$. If the next user;
  - rates idea A 'very interested' at trial t, the MAB system receives a reward $r_t = SE$ of $\{7, 7\}$.
  - rates idea A 'not interested at all' at trial t, the MAB system receives a reward $r_t = SE$ of $\{7, 1\}$.

The formula for the SE is:

$$r_t = \frac{\sigma(X_t)}{\sqrt{n(X_t)}} \tag{3}$$

Where:

- X_t = set of user ratings for the proposed idea up to trial t.
- $\sigma(X_t)$ = standard deviation of X_t.
- $n(X_t)$ = number of samples in X_t.

D.1 Simulated Evaluation

We evaluated how our MAB systems would explore and exploit during each collaborative ideation phase in a simulated environment. We created an environment that simulates a user's rating behavior during the idea generation phase (Figure 12). We assume that a user's ideation performance (i.e., how well they could generate helpful ideas responding to each inspiration and ideation method) would change over time. For instance, users might generate more helpful ideas by building on similar ideas at the beginning. As they run out of ideas by thinking about a similar subject, they might generate ideas more easily from dissimilar ideas later on. Accordingly, we predefined a set of rewards for four actions that would be most rewarding at the beginning, middle, end, or relatively lower than the other three actions throughout the trials. Assuming that users would not generate ideas forever, we tested the MAB system for up to 50 trials.

The simulation showed that the MAB system mostly identified the best actions. For instance, in the first 10 trials (Figure 12.a), the MAB system exploited Action 1 as it was the most rewarding. Then, responding to the decreased estimated reward of action 1 and increased uncertainty of the other actions, the MAB system explored Action 2 and 3 (Figure 12.b), which gave the lowest rewards. In the following 5 trials (Figure 12.c), the MAB system identified that Action 4 was most rewarding and exploited the action while exploring Action 1 (Figure 12.d), which previously gave higher rewards than the other actions.

We created another environment that simulates a situation where users review and rate ideas during the idea selection phase (Figure 12). We evaluated how the MAB system would adaptively reorder the ideas based on the ratings they receive. In the simulation, idea 2 and 10 had the highest standard error of the ratings. By selecting the idea 2 and 10, the chatbot would receive higher rewards. The simulation showed that the MAB system was prioritizing the two ideas throughout the trials. In the first 14 trials, the MAB system was mostly exploring, collecting users' ratings on each idea. Then, it identified that idea 2 was giving the highest reward, hence it prioritized idea 2 in the 15th trial. On the 18th and 22nd trials, the MAB system also prioritized idea 10 as it gave the higher reward.