1. Evaluation Methodology
   1. Goal Setting

Improving user experience is our main goal doing this evaluation for we can discover bugs and problems like the lack of guidance on certain functions. Additionally, we want to assess the impact to ensure our website effectively meets its objectives, such as helping users form eco-friendly habits after using it.

* 1. Participants identification

We are trying our best to find more participants who are of different age, sex and academic background. Additionally, the mobile device usage habits of the participants and their willingness to protect the environment are also taken account. Furthermore, we make sure there are long term users and short term users in our evaluation. All in all, we attach great importance to the diversity of the participants in the evaluation process.

If we have more time and budget to bring our website to market, we would definitely include a broader base of participants to our project to make the evaluation results more valid.

* 1. Handling Consent and Anonymity

In the beginning, we intended to prepare information sheet and consent forms to the participants. But due to the limitation and budget, we can’t make this happen. Even though, we still verbally informed the participants about the purpose of the study and ensured their anonymity. They are also told their participation is voluntary and they are free to quit at anytime.

* 1. Triangulation

Triangulation refers to a methodology aims to improve the reliability and while research and evaluation.

To ensure a robust evaluation, we employed two types of triangulation:

* + 1. Data triangulation: Different participants will evaluate the chatbot at different times.
    2. Methodological triangulation: We will use a combination of qualitative and quantitative data gathering techniques.
  1. Phased Evaluation

Because we want to know if our development process was on the right track, we decided to conduct the evaluation in two phases. Thanks to the 2 CS Testathons support provided by the programme office, we are able to have our project tested by the students.

The first evaluation took place on July 3rd, by which time we just completed the project demo and luckily we gained a lot of valuable suggestions on user experience. The suggestions are mainly about the aesthetic design and the functions of the project. We are advised to refactor the layout of the whole page and add welcome page and error page for better user's experience. Additionally, the practical effects of our website were also questioned—specifically, whether it can truly help users form an eco-friendly lifestyle. This reminded us to make this issue the focus of our second evaluation. This initial evaluation aims to identify any issues and help refine the project.

The second evaluation happened on August 7th, by which time we had already finished all the project and deployed it to the cloud. This comprehensive evaluation focused on user experience and assessing the overall effectiveness and impact, and is the main focus of this evaluation phase.

1. Evaluation Techniques

Besides the basic information sheet we designed to collet partipant data and a questionnaire specially tailored for the long term users, we focused more on the established and scientifically proven evaluation techniques listed below:

* 1. Qualitative Evaluation
     1. Think Aloud

Think aloud is a research method that a participant speak out what they think after completing a task with some observers recording critical moment. It is a valid way of collecting and interpreting participant thinking.

* 1. Quantitative Evaluation
     1. SUS

The System Usability Scale questionnaire is a reliable tool for taking a quick assessment of how people perceived the usability of computer systems they interact with.

The final score is derived from scores of the ten questions based on their question number. For odd-numbered questions, subtract 1 from user’s rating scale. For even-numbered questions, subtract user’s rating scale from 5. Then cumulate the 10 scores together and multiply the result by 2.5 to transit it from a 0-40 scale to a 0-100 scale.

* + 1. NASA TLX

The NASA Task Load Index (TLX) is a popular method to measure subjective mental workload. The final score is derived and processed based on 6 dimensions: mental demand, physical demand, temporal demand, performance, effort and frustration. Because we assume the 6 areas of workload are of similar significance to the individual, we chose the unweighted NASA-TLX method, which involves directly averaging the scores of the six dimensions without considering their relative importance.

* + 1. Long-Term Usage and Satisfaction Survey

This questionnaire is designed to track changes in user satisfaction over time and evaluate long-term usage pattern among main sections of the website like AR tree, chatbot and routine. Participants are asked to report daily whether they have used each section and rate their satisfaction on a scale from 0 (very unsatisfied) to 4 (very satisfied).

1. Results and Analysis
   1. Basic information of participants

We have 11 participants who tried our website for one time as our short term evaluators and 5 participants who are family and friends to the 5 team member as our long term evaluators. It's a lot to ask someone to use our website every day, and since we have limited time and energy to dedicate to the evaluation, we've had to rely on our family and friends, which might introduce some bias.



* + 1. Age and Gender Distribution

The majority of participants are male. Most participants are within the 18-24 age group.

* + 1. Daily Computer/Mobile Usage

Younger, male participants who spend more time on computers and mobile devices are likely to be more proficient with technology. They might find the website easier to navigate and use.

* + 1. Environmental Enthusiasm and Level of Liking

There is a wide range of enthusiasm for environmental protection among participants, which could influence how much they value the environmental features of the website. But the level of liking chart shows participants are positive about this no matter they care about the environment or not.

* + 1. Summary

From the participant demographic information, it’s evident that the current evaluation data is skewed towards younger, male, and tech-savvy users. This is unavoidable because this is precisely the main demographic profile appearing in the MVB. This could lead to limitations if we want to represent a wider user base. If possible, we intend to include a greater diversity of people into the evaluation and improve the website specifically for specifically targeted group.

* 1. Short Term Evaluation
     1. Think Aloud

Participants gave us many valuable feedbacks in the think aloud part. Something I need to mention is that we prepared heuristic evaluation and think aloud evaluation together in the first place, but we dropped heuristic evaluation in the end and chose to conclude the answers in think aloud by the heuristic aspects for the following two reasons. Firstly, this enable us to systematically organize and present user problems by heuristic principle. Secondly, it saves time and resources by avoiding the repetition of conducting two separate evaluations allows resources to be focused on a more in-depth analysis of the user experience.

In the help and documentation aspect, 3 users complained about not knowing where to turn to when they are assigned the first random task by the system. We decided to send users a message in the tutorial part when they encounter such situation.

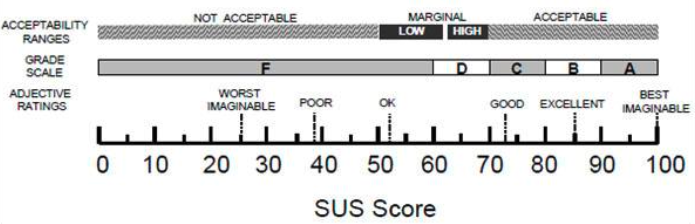
In the visibility of system status aspect, a participant pointed out that we should keep track of the tree planting process, which means that we need to display the tree points collected by users.

In the error prevention aspect, two of our participants reported a UI layout bug on phones. The button extends beyond the screen and cannot be clicked. After testing on phones, tablets, and computers, we found that the tablet version is the most suitable. In a word, the webpage still needs optimization.

In the flexibility and efficiency of use aspect, a participant suggested that the guidance provided by the chatbot in the first chat on how to use each function of our website should be more specific, such as "What can you do?" and "How many tasks need to be completed to plant a tree?" Another user thought it would be better if the admin could choose to push tasks to specific users.

* + 1. SUS

Bangor, Kortum, and Miller found there is a correlation between SUS scores and how people rated the systems and products they were evaluating using adjectives like "good," "poor," or "excellent", as shown in the picture below.



The final sus score of our system is 77.05, which falls into the category of “Good” , , approaching "Excellent". This suggests that most users had a positive experience with the system, though there may still be areas for improvement.

High scores on items such as complexity (Q2), ease of use (Q3), integration of functions (Q5), and required learning (Q10) suggest that the system is intuitive and that users feel comfortable using it.

Lower scores on items like the need for technical support (Q4)，cumbersomeness (Q8) and user confidence (Q9) indicate that some users might find the system slightly complex or feel the need for additional guidance. This indicated us that we need to improve our tutorial system and chatbot first chat guidance in the future.

* + 1. NASA TLX



The average NASA TLX score is 43.64, which suggests a moderate level of workload among the participants. This score indicates that while the tasks were not overly demanding, they were not entirely effortless either. The workload perceived varies in these six dimensions and six subscale scores revealed that the predominant effort required to engage with the website stemmed from the performance of each user and their exertion while using the website.

* 1. Long Term Evaluation

As we had to move to the writing part of the whole project, we don’t have much time for evaluation. Although it’s called long term evaluation, it only lasted for 5 days. This may result in not being able to fully and accurately reveal the correlations between different parts of the system and the trends in satisfaction changes.

* + 1. Correlation Analysis Between Different System Sections



We designed the questionnaire like this. We included questions on whether the user has used the three main sections of our website, trying to find the inner correlation between the usage of different sections of the system. But because we have really limited data, we were not able to find clear positive or negative correlation between these sections.

* + 1. User Satisfaction Trends Over Time

Most participants started with relatively high satisfaction scores (mostly 4s) and a slight satisfaction drop was seen in the middle two days, but it recovered in the last days. The reason might be that the user found the website quite complicated to use in the first few days so that they don’t like it. But when they completed several tasks and planted the first virtual tree, they started to find it interesting, inspiring and useful, which led to an increase in satisfaction in the following days.

The satisfaction scores given by most of the evaluators are quite high. Here I have to mention a possibility of bias. The long term users of our project are close friends and family to the team members. Although we have claimed them to be fair on our project and speak up their mind, there is still a minor uncertainty that they overrated our website.