**Research questions for evaluation to answer:**

1. Does JSONTalk accurately represent the contents of a JSON file?
2. Does JSONTalk allow users to quicker understand the structure and contents of a JSON file when it is represented by a screen reader?
3. Is the JSONTalk tool usable?

**Limitations of evaluation:**

The evaluation was conducted with the involvement of sighted programmers, rather than visually impaired programmers. Prior to the evaluation, we conducted a feasibility study to determine the viability of engaging visually impaired programmers in the evaluation. Based on the project timeline, we arrived at the conclusion that an evaluation with sighted programmers would be more appropriate.

As the primary target audience of the JSONTalk tool comprises visually impaired programmers who utilize screen-readers and other assistive technologies to perform programming tasks, the sighted users were presented with screen-reader transcripts of the JSON files instead of the JSON files themselves. This approach ensured that the use case of the tool was represented more realistically.

**Evaluation overview:**

The evaluation was carried out in the form of a between-subjects A/B testing, where participants were required to answer a series of questions related to three different JSON file screen-reader transcripts. The A group answered the questions using the JSONTalk tool, while the B group answered the questions without the tool. The evaluation was conducted online through a survey, which consisted of the following sections.

To ensure that all participants were on a level playing field when it came to prior JSON familiarity, both groups were primed on JSON syntax. Although all participants who were recruited had a strong programming background, this primer was included to ensure that the potential differences in JSON knowledge did not confound the results.

Next, the A group was shown how to download and use the JSONTalk tool, after which both groups were presented with three JSON file screen-reader transcripts and asked to answer three specific questions. The A group was allowed to use the JSONTalk tool to assist them with deciphering the contents of the JSON file from the screen reader transcript, while the B group was not provided with any additional tools. The tasks that participants were required to complete were as follows:

Task 1: Determine the level of nesting of a particular object.

Task 2: Determine whether there are any objects with duplicate structures within the JSON file.

Task 3: Rewrite the JSON file into proper syntax.

Following the completion of the tasks, the A group was asked to fill out a System Usability Scale matrix based on their experience with the JSONTalk tool.

The decision to use a between-subjects A/B testing design for this evaluation was made to enable a direct comparison between the A and B groups, while reducing the potential impact of individual differences in participant characteristics. By priming all participants on JSON syntax, the potential differences in prior JSON knowledge between the groups were minimized. Additionally, by limiting the use of the JSONTalk tool to the A group, we were able to assess the effectiveness of the tool in facilitating the completion of the tasks, while also providing a fair comparison of the performance between the two groups. The use of an online survey format for the evaluation allowed for efficient data collection from a geographically diverse set of participants.

**Data collected**

The evaluation collected various types of data to assess the effectiveness of the JSONTalk tool. The error rate for each question was collected, allowing for an assessment of the accuracy of the responses provided by the participants. Additionally, the similarity between the JSON files rewritten by participants and the actual JSON file used was measured. This provided insight into the ability of participants to accurately understand and replicate the syntax of the JSON file.

The time taken for users to complete each task was also recorded, providing an indication of the relative efficiency of using the JSONTalk tool in completing the tasks. Finally, the answers to the System Usability Survey were collected, which allowed participants to provide their feedback on their experience using the JSONTalk tool. This data was then used to determine the level of user satisfaction with the tool, as well as any areas for improvement.

**Participant overview**

* Include info of sample size

**Statistical methods**

* Include calculations and why they are being used

The data collected in the evaluation was analyzed using various statistical methods to determine the effectiveness of the JSONTalk tool. Here are some of the calculations and methods used:

Error rate analysis: The error rate for each question was calculated by dividing the number of incorrect responses by the total number of responses. The error rates for the A and B groups were then compared to determine the impact of using the JSONTalk tool on the accuracy of responses.

T-test analysis: A t-test was performed to determine the statistical significance of the differences in error rates between the A and B groups. This helped to determine whether the use of the JSONTalk tool resulted in a significant improvement in the accuracy of responses.

Levene's test: Levene's test was used to determine the equality of variances between the A and B groups. This helped to determine whether the assumption of equal variances could be made for the t-test analysis.

Cohen's d analysis: Cohen's d was calculated to determine the effect size of using the JSONTalk tool on the error rates. This helped to determine the practical significance of the differences in error rates between the A and B groups.

Similarity analysis: The similarity between the JSON files rewritten by participants and the actual JSON file used was measured using a similarity metric. The average similarity score for each group was then calculated and compared to determine the impact of using the JSONTalk tool on the ability of participants to accurately understand and replicate the syntax of the JSON file.

Mann-Whitney U test: A Mann-Whitney U test was performed to determine the statistical significance of the differences in similarity scores between the A and B groups. This helped to determine whether the use of the JSONTalk tool resulted in a significant improvement in the ability of participants to accurately understand and replicate the syntax of the JSON file.

Time analysis: The time taken for users to complete each task was compared between the A and B groups to determine the relative efficiency of using the JSONTalk tool in completing the tasks.

System Usability Scale (SUS) analysis: The responses to the SUS questions were analyzed to determine the level of user satisfaction with the JSONTalk tool, as well as any areas for improvement. The average SUS score for the A group was compared to the average SUS score for a normative sample to determine the relative usability of the JSONTalk tool.

The statistical methods used in the evaluation helped to determine the effectiveness of the JSONTalk tool in facilitating the completion of the tasks and improving the accuracy of responses.