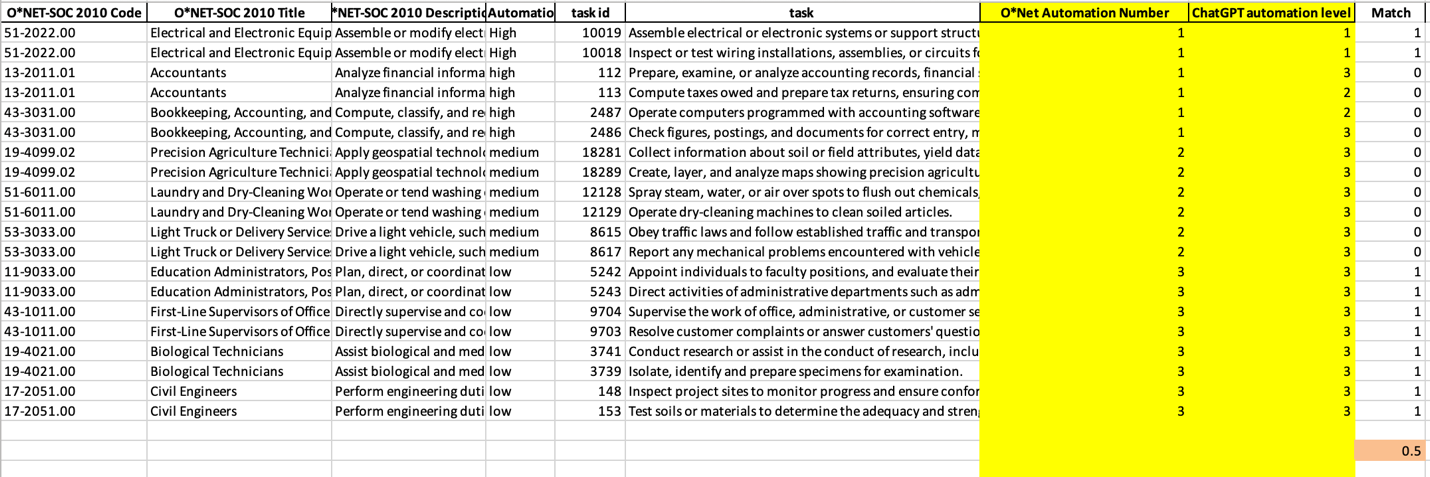
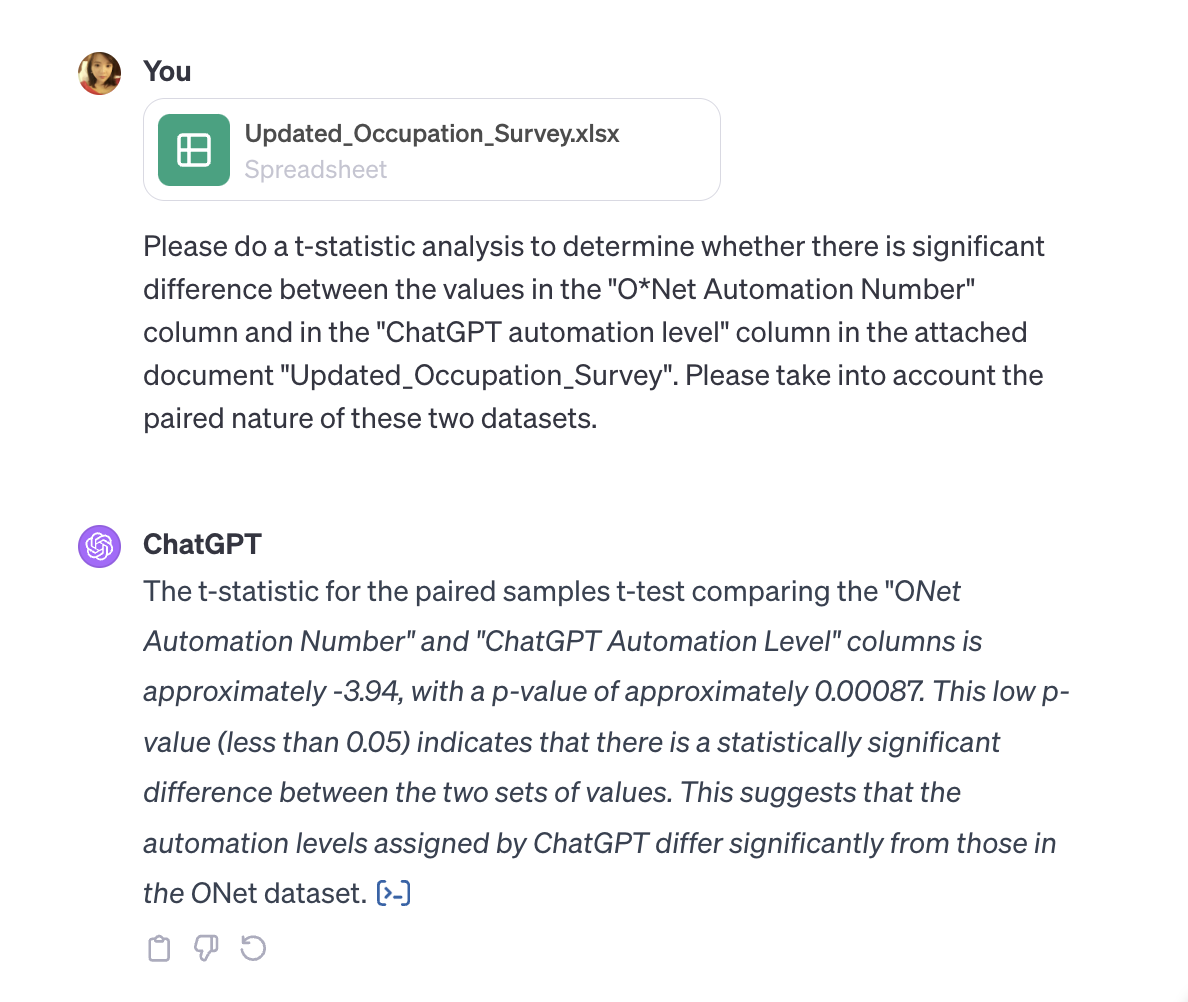
**Compare ChatGPT score to O\*Net automation level:**

Data:



T-statistic analysis:



Cohen’s kappa test:

A screenshot of a chat

Description automatically generated

Rate of match:

A screenshot of a computer

Description automatically generated

Analyzing the reasons for the differences between the O\*Net evaluation and the ChatGPT automation level requires a multifaceted approach, considering both the nature of the data and the methodologies used for evaluation. Here are some key aspects to consider:

1. **Evaluation Criteria Differences**: O\**Net and ChatGPT may use different criteria or frameworks for evaluating automation levels. O\**Net's evaluations are likely based on standardized industry research and expert opinions, whereas ChatGPT's assessments might be based on its training data, which includes a wide range of texts from the internet, books, papers, etc. The difference in sources and evaluation criteria can lead to variations in the automation level assessments.
2. **Subjectivity in Interpretation**: ChatGPT's responses are generated based on the patterns it has learned from its training data. This can introduce a degree of subjectivity, especially in areas where there is no clear consensus or where the information is evolving. In contrast, O\*Net's data may be more consistent and standardized.
3. **Differences in Update Frequency**: The O\*Net database is periodically updated with input from a broad range of professionals and reflects current industry standards. ChatGPT's knowledge, on the other hand, is dependent on its training data, which might not include the most recent industry changes or trends.
4. **Scope of Automation Assessment**: O\*Net might focus on specific aspects of a job's potential for automation, such as the technical feasibility, economic viability, or the current state of technology. ChatGPT's assessment might be more influenced by general discussions about job automation, which can sometimes be speculative or based on broader trends.
5. **Potential for Ambiguity in Data Interpretation**: The way each dataset interprets and categorizes automation levels might differ. For example, what one dataset considers a 'high' level of automation might be considered 'medium' by another, depending on their respective scales and benchmarks.
6. **Human Error in Data Entry**: There's also a possibility of human error in entering data, either in the original sources or in the process of compiling these datasets.
7. **Lack of Contextual Understanding by ChatGPT**: While ChatGPT can process and generate text based on patterns, it may not fully understand the context or nuanced aspects of certain jobs, which can affect its ability to accurately assess automation levels.

To get a more detailed understanding, one would need to conduct an in-depth analysis of the methodologies used by both O\**Net and ChatGPT for their assessments, along with a comparative study of specific instances where the evaluations diverge significantly. This might involve a qualitative analysis of the reasoning behind ChatGPT's evaluations and a comparison with the criteria used by O\**Net.

For the**O\*Net evaluation**, the T-test analysis indicates significant differences between the ChatGPT score and the O\*Net evaluation. The Cohen's kappa test indicates slight agreement between the two sets of values. The rate of match is 50% means that half of the ChatGPT scores are exactly the same as the O\*Net evaluations.

**Compare ChatGPT score to human score for Job search process**

T-test:

A screenshot of a chat

Description automatically generated

Rate of match for each student:A screenshot of a chat

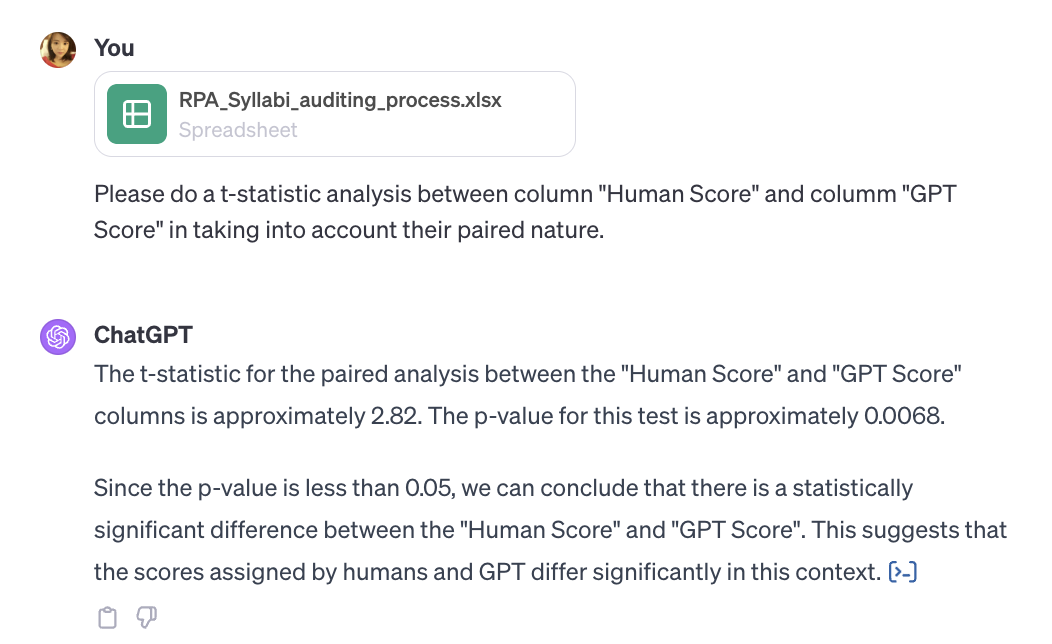
Description automatically generated

Rate of match for the entire group:A screenshot of a chat

Description automatically generated

For the **Job search Process RPA evaluation**, the results are very encouraging. The T-test analysis indicates no statistically significant difference between the 2 groups. And the rate of match for the entire group is 60.76% means more than a half evaluation results are exactly the same between the 2 groups.

**Compare ChatGPT score to human score for Syllabi auditing process**

T-test entire group: 

T-test without team 1:A screenshot of a computer

Description automatically generated

T-test mean score by team:

A screenshot of a text

Description automatically generated

Cohen’s kappa test:

A screenshot of a computer

Description automatically generated

The rate of match for each team:

A screenshot of a white background

Description automatically generated

The rate of match for the entire group:

A screenshot of a white background

Description automatically generated

For the **Syllabi auditing RPA evaluation**, the T-test results is not good enough to prove the tool's accuracy. But the Cohen's Kappa test result indicates slight agreement beyond chance between the human scores and GPT scores. And The rate of match of the entire group is 56% means more than half evaluation scores are exactly the same between the two groups.