**Robot Replaceability Radar Accuracy Validation**

**Xiaoyun Bian**

1. **Compare Chatbot automation level to O\*Net automation level (Accuracy: 64%):**

**A screenshot of a document

Description automatically generated**

To investigate the correlation between "Chatbot Automation Level" and "O\*NET Automation Level", we can use several methods to assess how well the chatbot automation predictions align with the established O\*NET levels. Here are the approaches we can take:

1. **Pearson Correlation Coefficient:** Measures the linear correlation between two datasets. It assumes that both datasets are normally distributed.
2. **Spearman's Rank Correlation Coefficient:** A non-parametric measure of rank correlation, which assesses how well the relationship between two variables can be described using a monotonic function.
3. **Kendall’s Tau:** Another non-parametric measure that assesses the ordinal association between two variables.

Here are the correlation coefficients between "Chatbot Automation Level" and "O\*NET Automation Level":

* **Pearson Correlation Coefficient:** 0.532
* **Spearman's Rank Correlation Coefficient:** 0.521
* **Kendall’s Tau:** 0.492

All three coefficients suggest a moderate positive correlation between the two sets of data. This indicates that as the O\*NET Automation Level increases, the Chatbot Automation Level also tends to increase, suggesting that the chatbot's assessments are somewhat aligned with the established O\*NET levels. However, the correlation is not very strong, indicating there is room for improvement in the chatbot's accuracy or that other factors might be influencing the levels set by the chatbot.

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) between the "Chatbot Automation Level" and "O\*NET Automation Level" are as follows:

**Mean Absolute Error (MAE): 0.48**

**Root Mean Square Error (RMSE): 0.85**

Let's visualize the agreement between these two measurements using a Bland-Altman plot, which will help identify any bias or trends in the discrepancies as the automation levels change. ​

A graph with red and green lines

Description automatically generated

The Bland-Altman Plot visualizes the differences between the "Chatbot Automation Level" and the "O\*NET Automation Level" against the average of these two measures. Here's what we can interpret from this plot:

* The **red dashed line** represents the mean difference, which is close to zero, indicating that there is no significant bias over the entire range of automation levels.
* The **green dashed lines** represent the limits of agreement (mean difference ± 1.96 times the standard deviation). Most of the points lie within these bounds, suggesting that the differences between the chatbot and O\*NET levels are generally consistent across the range of values, without extreme outliers.

This plot suggests that while there is a moderate average error (as indicated by the MAE and RMSE), the chatbot's assessments are not systematically biased in overestimating or underestimating the automation levels compared to O\*NET. These findings could be useful for further tuning the chatbot's algorithm to reduce discrepancies or understanding specific cases where the differences are larger.

1. **Compare Human Beings automation level to O\*Net automation level (Accuracy: 52%):**

A screenshot of a document

Description automatically generated

The correlation coefficients between the "Human Automation Level" and the "O\*NET Automation Level" are as follows:

**Pearson Correlation Coefficient:** 0.654

**Spearman's Rank Correlation Coefficient:** 0.634

**Kendall’s Tau:** 0.572

These coefficients indicate a moderate to strong positive correlation, suggesting that the human assessments are more closely aligned with the O\*NET levels compared to the chatbot's assessments. This could imply that humans are potentially better at estimating these levels, or they interpret the tasks and roles in a manner more aligned with the established standards.

For the "Human Automation Level" compared to the "O\*NET Automation Level," the statistical error measurements are:

**Mean Absolute Error (MAE):** 0.48

**Root Mean Square Error (RMSE):** 0.69

These results show a similar MAE to the chatbot's assessment but a slightly lower RMSE. This indicates that while the average errors are similar, the human assessments have fewer large errors compared to the chatbot when measured against the O\*NET levels.

Next, let's create a Bland-Altman plot to visualize the agreement between the human and O\*NET automation levels.

A graph with red and blue dots

Description automatically generated

In conclusion, while the accuracy rate of the chatbot is 12% higher than that of humans, the human assessments may be more consistently close to the mark, showing less variation from the O\*NET levels even when they are incorrect. This could make human assessments potentially more reliable in a practical, real-world application where exact matches are less critical than being generally correct. These insights highlight the importance of choosing the right metrics based on the specific needs and contexts of the assessment.

1. **Compare Chatbot Score to Human Score**

For the comparison between "Chatbot Automation Level" and "Human Automation Level," the calculated error metrics are:

**Mean Absolute Error (MAE):** 0.72

**Root Mean Square Error (RMSE):** 0.94

These values suggest that there is, on average, a discrepancy of about 0.72 levels between the chatbot and human assessments, with a higher RMSE indicating that larger discrepancies occur as well.

The correlation coefficients between "Chatbot Automation Level" and "Human Automation Level" are as follows:

**Pearson Correlation Coefficient:** 0.567

**Spearman's Rank Correlation Coefficient:** 0.552

**Kendall’s Tau:** 0.526

These values indicate a moderate positive correlation, suggesting that while the chatbot and human assessments generally agree to some extent, there are noticeable discrepancies. Although there are some differences between chatbot and human evaluations, they are not excessively varied, implying a general agreement with some level of consistent error.

In conclusion, while they generally agree, the differences in MAE and RMSE, along with the moderate correlation, suggest room for improvement in aligning the chatbot's evaluations more closely with human judgments.