Teaming up with AI:

Assessing Performance Impact of AI Integration in Super Mario Party

Soham Sachdev

January 11, 2024

Table Of Content

Introduction

This project aims to study and analyse the effects of Artificial Intelligence being a part of a team. The question is: how well can AI collaborate with humans to work towards and achieve a goal in a team environment and how does it affect the performance as compared to a team of humans? With the help of this study, one can put up an argument that there is no significant difference in the performance of the teams involving an AI agent when compared to a team of humans. This project involved players playing a mini-game called 'Dash and Dine' of the video game 'Super Mario Party' which requires them to work together around a small kitchen to nab requested ingredients (Lagioia Stephen, 2021). In the end, it was clear that substituting a player with an AI agent or a new player (newhire) or even playing with the same teams did not account for any significant difference. This can prove to be a very important argument for those who believe that AI can thrive and replace humans for the better in all the jobs that humans are currently doing. The report is divided into four main parts, defining the problem statement, methods used to solve the task, evaluating the results, and summary of the study.

Problem Statement

The objective of this study is to know if the involvement of AI in team spaces improves the throughput of the teams or not. In this study, the dataset consists of 55 unique teams playing the mini-game Dash and Dine. Each team played 12 rounds which resulted in 660 records in the dataset. The experiment was divided into two phases: Phase 1 consisted of teams playing the mini-game six times. After these six games, the teams were changed. In 15 teams, one of the members was substituted by a new member, 'newhire', and in 20 teams by an artificial intelligence agent, 'ai'. The other 20 teams remained unchanged,' control'. In Phase 2, the teams played another six rounds of the game. At the end of each round in both phases, the total ingredients collected by each team were recorded. Along with the number of ingredients collected, the dataset also includes variables like 'team id', 'phase', 'group', and 'round'. The questions that I am trying to answer in this study are:

1. Is there a significant difference between the overall performances (Phase 1 and 2) of the three groups (newhire, ai, control)?

2. Is there a significant difference between the performances of the groups after the changes i.e. Phase 2?

Additionally, the performance improvements of the groups from Phase 1 to Phase 2 will be observed to know which groups improved and which one improved the most.

Methodology

The methods followed for this study are divided into two parts: Descriptive Statistics and Hypothesis Testing. By analysing the descriptive statistics of the dataset, it will give an idea about all the variables and overall performance can be examined. After that, a hypothesis test would be ideal to find out if there is a significant difference between the groups.

1. Descriptive Statistics:

To compare the overall performances of the groups, comparing their descriptive statistical measures would be a good choice. Descriptive statistics summarise and organise characteristics of a data set(Bhandari Pritha. 2020). The descriptive statistics can be divided into different types:

1. **Measures of Central Tendency:**

   Measures of central tendency estimate the centre, or average, of a data set(Bhandari Pritha. 2020).

   a. Mean: Calculated by dividing the sum of values by the number of values, also known as average.

   b. Median: It is the value that's exactly in the middle of a data set(Bhandari Pritha. 2020).

2. **Measures of Variability:**

   Measures of variability give you a sense of how spread out the response values are(Bhandari Pritha. 2020).

   a. Standard Deviation: The standard deviation (*s* or *SD*) is the average amount of variability in your dataset. The larger the standard deviation, the more variable the data set is(Bhandari Pritha. 2020).

   b. Range: The range gives an idea of how far apart the most extreme response scores are(Bhandari Pritha. 2020).

To summarise the findings from the above statistics, plotting different statistical graphics would help in understanding them better.

1. Box plot: A boxplot is a standardised way of displaying the distribution of data based on its five-number summary ("minimum", first quartile [Q1], median, third quartile [Q3] and "maximum"). Boxplots can tell you about your

outliers and their values, if your data is symmetrical, how tightly your data is grouped and if and how your data is skewed (Michael et al., 2023). The reason behind choosing the box plot is also because we only have one numeric variable i.e, total ingredients and two categorical variables, group and phase.

2. Bar chart: To compare different group's performance in different phases, bar charts can help to find the differences in their performances visually.

2. Hypothesis Testing:

Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample data to draw conclusions about a population. It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (Ha), and then collecting data to assess the evidence(Biswal Avijeet, 2023). Hypothesis testing is generally used to get an idea about the population by analysing a sample from that population. But it can prove useful in this case by making an assumption and then that assumption can be accepted or rejected based on the analysis of the data.

1. **ANOVA(Analysis of Variance test, also called one-way ANOVA):**

   The purpose of a one-way analysis of variance (one-way ANOVA) is to compare the means of two or more groups (the independent variable) on one dependent variable to see if the group means are significantly different from each other(Timothy C. Urdan, 2010).

   ANOVA suits this use case perfectly as we are trying to find out the difference between the number of ingredients collected by the three groups in Phase 2 of the study.

   ANOVA is a statistical test that assumes that the mean across 2 or more groups are equal. If the evidence suggests that this is not the case, the null hypothesis is rejected and at least one data sample has a different distribution(Brownlee Jason, 2019).

   Fail to Reject H0: All sample distributions are equal. Reject H0: One or more sample distributions are not equal. Importantly, the test can only comment on whether all samples are the same or not; it cannot quantify which samples differ or by how much(Brownlee Jason, 2019).

2. **Tukey’s HSD Test:**

   Tukey's HSD (Honestly Significant Difference) is a statistical method that helps in identifying which pairs of means in a set of groups are significantly different from each other(Faster Capital, 2023). Tukey's HSD is a post-hoc test commonly used after conducting an analysis of variance (ANOVA) to determine if there is a significant difference between the means of three or more groups(Faster Capital, 2023). Tukey's HSD method calculates the minimum difference between two means that is required to be considered statistically significant(Faster Capital, 2023). This minimum difference is called the HSD. If the difference between two means is greater than the HSD, then the two means are considered significantly different(Faster Capital, 2023). After arriving at an initial conclusion after the ANOVA test, it is necessary to confirm it by performing Tukey’s HSD test. The advantage of Tukey’s HSD over ANOVA is that it checks for the difference between the groups in a pair-wise manner.

Evaluation

In this section, assessment of all the statistics and the tests that were performed will be done.

- **Measures of Central Tendency:**
  1. Mean:

Table 1: Mean value of total ingredients collected by the groups in Phase 1 and 2

Groups	Mean
newhire	25.289
ai	24.679
control	24.596

Table 1 presents the mean values of total ingredients collected by each group in both Phase 1 and 2. The 'newhire' group has the highest mean value.

  2. Median:

Table 2: Median value of total ingredients collected by the groups in Phase 1 and 2

Groups	Median
newhire	26.0
ai	25.0
control	25.0

Table 2 displays the median values, with the 'newhire' group again having the highest median total ingredients.

- **Measures of Variability:**
  1. Standard Deviation:

Table 3: Standard Deviation of the total ingredients collected by the groups in Phase 1 and 2

Groups	Standard Deviation
newhire	4.546
ai	4.298
control	4.354

Table 3 illustrates the standard deviation of total ingredients for each group. While there are marginal differences, the 'newhire' group exhibits a slightly higher standard deviation.

  2. Range:

Table 4: Minimum, Maximum and Range of the total ingredients collected by the groups in Phase 1 and 2

Groups	Maximum	Minimum	Range
newhire	36.0	13.0	23.0
ai	36.0	9.0	27.0
control	34.0	11.0	23.0

Table 4 outlines the minimum, maximum, and range of total ingredients. Both 'newhire' and 'ai' groups share the maximum value, but 'ai' has a slightly larger range compared to 'newhire' and 'control.'

- **Box Plot:**

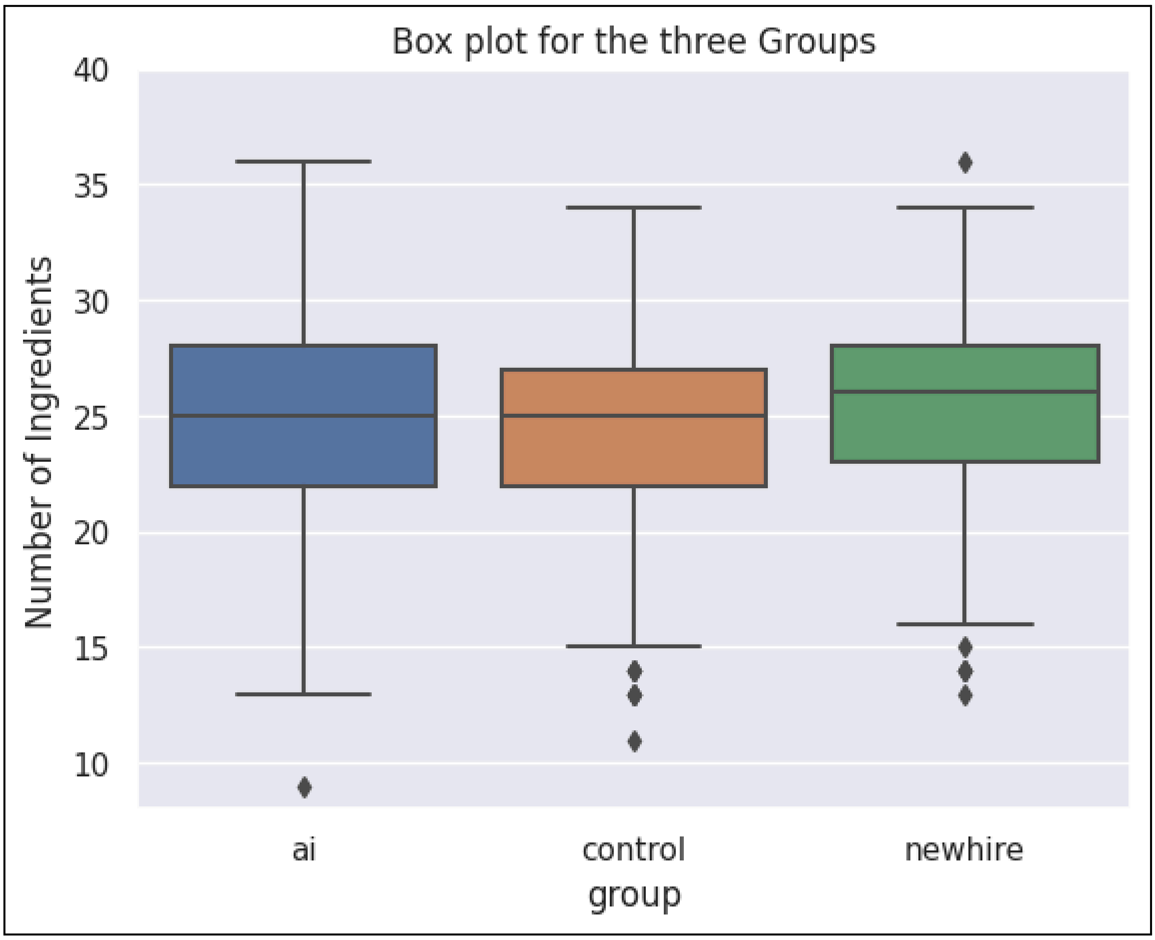

Figure 1: Box plot, Group Vs Number of Ingredients

Figure 1 consists of a box plot depicting the distribution of total ingredients across the three groups in the 12 rounds. The plot allows for visual comparison of the groups.

- **ANOVA:**

  First of all, the null hypothesis (H0) presumes that there is no significant difference between the means of the three groups in Phase 2. The level of significance (α) is set at 0.05, a standard assumption for hypothesis testing.

  After performing the ANOVA test, the calculated p-value is 0.523. According to the rule, since the p-value is greater than the significance level, we do not reject the null hypothesis. This suggests that there is not enough evidence to conclude a significant difference in the means of the three groups in Phase 2.

  ANOVA is employed in this analysis as it allows for the comparison of means across multiple groups. This method is suitable for examining differences in the total ingredients collected by the 'newhire,' 'ai,' and 'control' groups.

Table 5: Multiple Comparison of the Means - Tukey's HSD

Group 1	**Group 2**	**Mean Difference**	**p-adjusted**	**reject**
ai	control	0.517	0.485	False
ai	newhire	0.900	0.155	False
control	newhire	0.383	0.710	False

- **Tukey's HSD:**

  The Tukey HSD test is used for multiple pairwise comparisons after ANOVA. Table 5 shows the mean differences between groups, along with adjusted p-values. The 'reject' column indicates whether the null hypothesis of equal means is rejected for each comparison. In this case, all pairwise comparisons have p-values greater than 0.05. None of the group comparisons (ai vs control, ai vs newhire, control vs newhire) have a significant difference in means.

- **Bar Chart** :

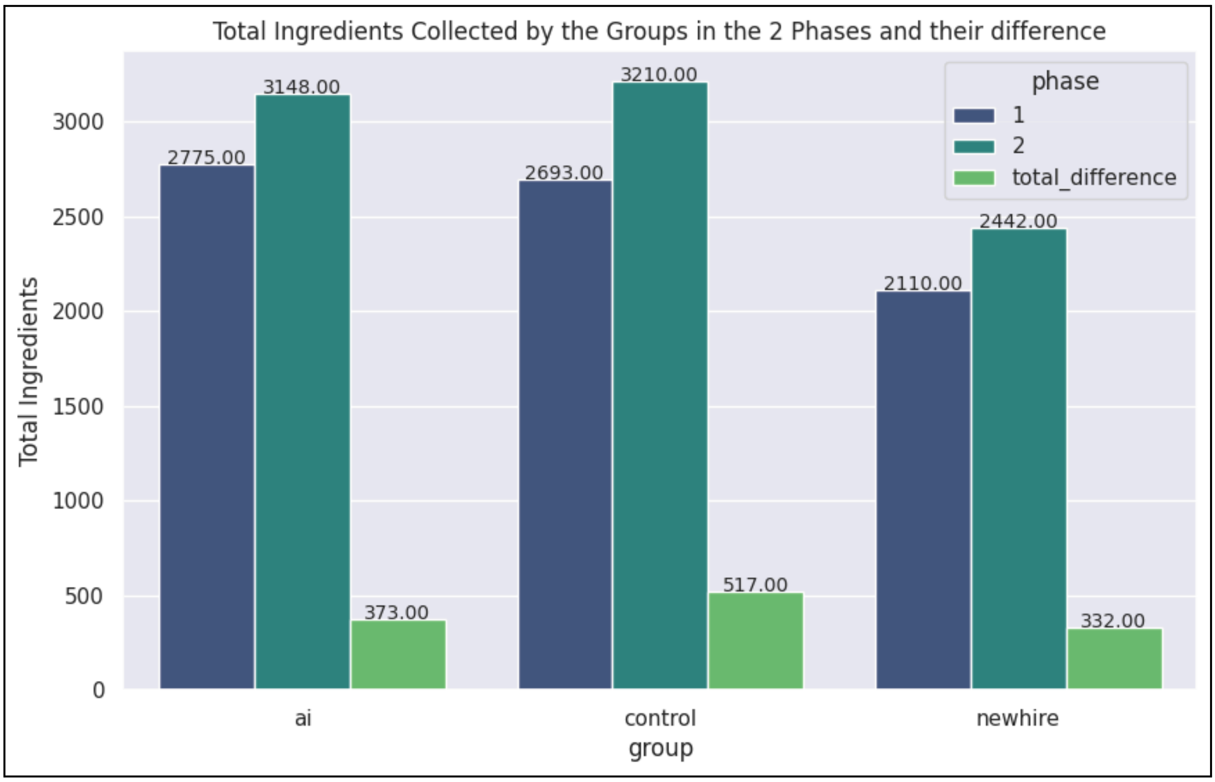

Figure 2: Total Ingredients collected by the Groups in the 2 Phases and their difference

Figure 2 consists of a bar chart that displays the total number of ingredients collected by each group of teams ('ai', 'control', 'newhire') in the 2 Phases and the difference between their phase 1 and 2 collections. This graphic helps in understanding the growth of each group on their own and comparing them with other groups as well.

Summary

There have been lots of advancements in the field of Artificial Intelligence in the last few years. Naturally, because of that, there have been discussions that as AI can individually perform all the tasks effectively as compared to humans, soon AI will be replacing employees in various organisations thus helping companies in reducing the company cost. This experiment was performed to understand the effects of the involvement of AI in a team environment with the help of a mini-game 'Dash and Dine'.

How significant is the effect on the performance of the teams that have AI as a part of their team? How does that compare to replacing a member of the team with a new human member? Also, what if no changes are made to the roster? Based on the results, there is no significant difference between the performances of the teams that remained unchanged(control), teams that played with a new player, and a team(newhire) that played with an AI agent(ai) after the first six rounds of the experiment. Figure 1 shows that there is no visual difference between the three groups which is also backed by statistics calculated such as mean, median, and more. The statistical tests such as ANOVA and Tukey's HSD confirmed that there were no significant differences between the three groups in Phase 2 after the changes were made. Additionally, the unchanged group (control) experienced the most growth (which is depicted in Figure 2) which can be a result of increased coordination and communication between the team members as they played 12 rounds of the game with the same people.

The analytical approach followed in this experiment based on the dataset has enough evidence to suggest that there is no significant effect of replacing AI with a human for the better in a team environment at least as of now. Conversely, it can have negative effects on the synergy of the team while building a new team. Further studies such as human response to working on a team that consists of an AI agent, comparing the effects on performances of humans on an individual basis when being part of a team which consists of an AI agent and when not and more should be carried out before instilling AI as a part of a team environment which completely consists of human beings as of now.

Bibliography

Reference List
Lagioia, Stephen. 2021. “Mario Party: 10 Best Mini Games Throughout The Series, Ranked.” Game Rant. https://gamerant.com/mario-party-best-mini-games-in-the-series/. Accessed January 11, 2024.
Bhandari, Pritha. 2020. “Descriptive Statistics | Definitions, Types, Examples.” Scribbr. https://www.scribbr.com/statistics/descriptive-statistics/. Accessed January 8, 2024.
Galarnyk, Michael, and Brennan Whitfield. 2023. “Understanding Boxplots: How to Read and Interpret a Boxplot.” BuiltIn. https://builtin.com/data-science/boxplot. Accessed January 11, 2024.
Biswal, Avijeet. 2023. “What is Hypothesis Testing in Statistics? Types and Examples.” Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics. Accessed January 9, 2024.
Page 105, Urdan, Timothy C. 2010. Statistics in Plain English. N.p.: Routledge.
Brownlee, Jason. 2019. “How to Calculate Parametric Statistical Hypothesis Tests in Python - MachineLearningMastery.com.” Machine Learning Mastery. https://machinelearningmastery.com/parametric-statistical-significance-tests-in-python/. Accessed January 10, 2024.

“Tukey's HSD: Post Hoc Analysis after ANOVA.” 2023. FasterCapital. https://fastercapital.com/content/Tukey-s-HSD--Post-Hoc-Analysis-after-ANOVA.html.
Accessed January 12, 2024.

Software Used:
Google Colaboratory (https://colab.google/)

Common Libraries:

1. Pandas (https://pandas.pydata.org/)
2. Seaborn (https://seaborn.pydata.org/)
3. Matplotlib (https://matplotlib.org/)

Additional libraries:

1. scipy.stats.f_oneway
   One way ANOVA test
   (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html)

2. statsmodels.stats.multicomp.pairwise_tukeyhsd
   Tukey’s HSD test
   (https://www.statsmodels.org/devel/generated/statsmodels.stats.multicomp.pairwise_tukeyhsd.html)

Appendix

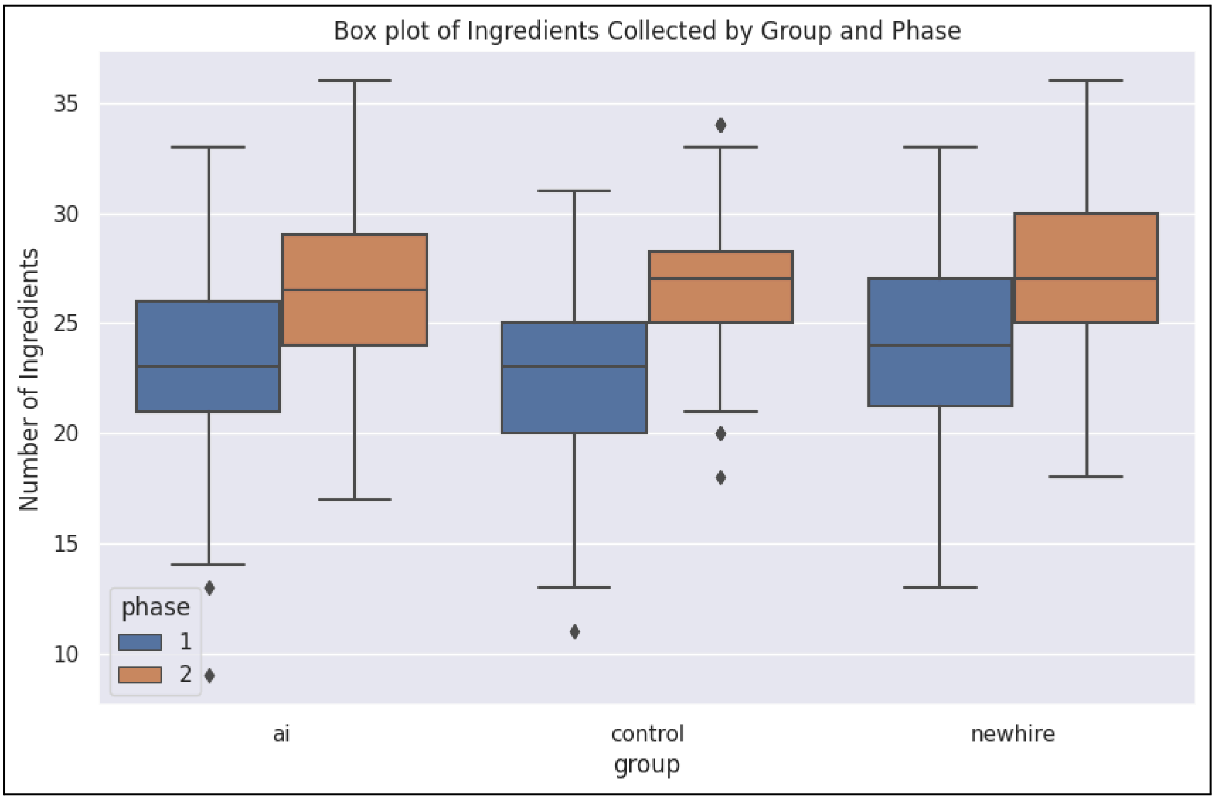

Figure 3: Comparison of box plots of the three groups in the 2 Phases