MSc Computer Science & Big Data Analytics

ChatGPT: The advancement of knowledge and incorporation for its users - Methodology

24/05/2023

Word Count – 2,016

# Methodology

The goal of this chapter is to explain how the research was performed, research questions and hypotheses used, dataset preparation and collection, pre-processing steps, and analysis performed; all to establish whether ChatGPT – if used – has enhanced a user’s knowledge; and if so, has ChatGPT then since been incorporated into a user’s routine for knowledge acquisition.

## Research Hypotheses and Questions

The following information below represents a set of two alternative research hypotheses **(RH)** to be explored as part of the study, where the first hypothesis has impact on the second hypothesis.

### 3.1. Research Hypotheses

1. The utilization of ChatGPT increases the likelihood of enhancing a user's knowledge on a specific topic.”
2. Users who perceive ChatGPT as enhancing their knowledge are more likely to incorporate ChatGPT into their routine for knowledge acquisition compared to those who do not perceive ChatGPT as enhancing their knowledge.

Based on these hypotheses proposed, there are potential leading questions that could be answered to re-confirm or extend the response further.

### 3.2. Research Questions

As well as the hypotheses to explore, there is potential to ask additional research question to provide further context in this area.

1. Is ChatGPT being used by the participants?
2. If ChatGPT is being used by the participant, did the use of it enhance a user’s knowledge based on what was submitted?
3. If ChatGPT did enhance a user's knowledge, has ChatGPT been adopted for routine knowledge acquisition by the user?

## Research Design

This dissertation uses a quantitative study for its speed and ability to address specific questions, allowing for future repeatability and measurement of attitude changes [10-11]. The quantitative method offers objectivity, generalisability to a broader context – enabling a foundation to build further research upon, and the ability to measure the potential cause and effect between users' knowledge enhancement of ChatGPT and its incorporation into daily use. Statistical inferences can also be used to test the hypotheses set [11-12]. Qualitative research, although valuable for building themes, providing context, and gaining insights from participants, will be a good candidate for future studies once this preceding research is complete. Due to time and resource constraints, a mixed method approach is not feasible, making the quantitative study the preferred approach.

### Quantitative Research Method

Quantitative research utilises a survey study methodology to collect primary research information via a questionnaire, asking specific, mainly closed-ended questions and analysing the results [14]. The purpose of the survey is to gather participant feedback on ChatGPT's usage, knowledge enhancement, and adoption. Conducted online, it aimed to access to a large participant group to help ensure an optimal response rate compared to physical submissions [14]. The survey is cross-sectional as the research is novel, to establish responses from one point in time [15] over a period of approximately one month, that could be potentially compared against for future research.

## Preparation

Before research began, the survey was created ready for distribution, aimed to be done within the research project timeline. A copy of the survey used can be found in **Appendix A.**

## Population and Participants

The sample size used in this research consists of all individual ?? responses who interacted with the survey providing input regarding the topic of ChatGPT. However, some user characteristics was collected to allow for analysis and aid future research.

The respondents who participated in this study consisted of individuals who completed a survey titled “ChatGPT: The advancement of knowledge and incorporation for its users”. The survey was distributed to the public - by way of social media platforms as well as a post on the Wrexham Glyndwr University’s Canvas discussion board – to access a link to the Google Forms questionnaire that was created.

The survey was made available to all potential respondents, and their participation was entirely voluntary. By including all respondents who participated, this study aimed to capture a comprehensive understanding of user perspectives regarding ChatGPT. Furthermore, this approach allowed for a broader range of insights to be uncovered, enhanced generalisability of findings from a diverse range of individuals, and reduced potential for sampling bias.

## Data Collection

As this research utilised an online survey, an online questionnaire instrument was deployed. This instrument was created via a software tool known as Google Forms. With this method, a link was distributed via social media posts or direct communication on multiple platforms such as LinkedIn, Discord, and Facebook. Additionally, A link was also provided on the Wrexham Glyndwr University’s Canvas discussion board. The steps followed in the data collection process were the following:

1. Shared a link to individuals across multiple platforms to access the Google Form survey created. Also, a link to the Google form was distributed direct communication on social media where requested.
2. The individual answered questions and made their submission.
3. At the end of the time allotted to receive responses, data was exported from the Google Form tool to a downloadable csv file.
4. From there, the csv file was imported into a python jupyter notebook script producing the relevant statistics and analysis.

### Data Collection Suitability

The decision to collect the required data from the created survey – as described in the section above – was to allow for the speed and quality of responses that were received. Moreover, given the constraints of time and resource available for this research, this was the most optimal method to collect data, as well as allowed for ease and approachability for the participant.

## Data Analysis and Results

### Management and storage

As described in the previous section, data was collected via a Google Forms questionnaire, then exported into a CSV (Comma-Separated Values) file, which then was imported for analysis and statistical generation.

### Software / tools used for Data Analysis

The CSV file was imported into Python, a user-friendly and interpretable high-level programming language [16]. Python supports data analysis through libraries like pandas for data analysis and manipulation [21], SciPy for statistical tests [22], matplotlib for standard data visualisation [23], and seaborn for statistical visualisation [24], all facilitating statistical analysis and data visualization for the collected data. The justification of using python, compared with other data analysis methods such as MATLAB, R or SPSS for example were due to a few factors to contributed to the decision of using python. Mainly, its versatility, ease of use and readability. Python is framed as a general-purpose programming language, with extensive libraries being able to cover various research domains, and beyond, as well as being able to integrate into other programming languages, frameworks, and practices better, allowing for potential future development, regardless of avenue that is taken; where languages such as MATLAB, R, or SPSS remain more domain specific [34].

### Data Analysis Steps

The following list of tasks were executed to present findings and derive insight from the data collection:

#### 1 – Pre-processing

Most participant data collected will consist primarily of categorical responses, including yes or no choices and some characteristic information selected from multiple choice selection (e.g., occupational background, highest level of education). Consequently, data conversion also transformed categorical values into numerical representations, utilising transformative logic. Additionally, free-text information was collected pertaining to a recent prompt used in ChatGPT as well as limited user characteristics. at a high level, the following steps taken regarding the pre-processing of data are as follows:

1. Removal of identifiable information and renaming of questions used in the survey, to represent the research question they relate to.
2. Conversion of values held in the research questions into numeric format.
3. Renaming of the remaining questions used in the survey for conciseness. Table 1 below shows the questions or information - in numerical order - used and collected in the survey, with the variable name they were renamed to ready for analysis in the python code.
4. Conversion of remaining survey values into respective data types. This information can also be seen in Table 1.
5. For the values relating to the prompt entered by the user, the industry sector the user is related to, and the course the student is studying, punctuation and stop words were removed, with the aim to simplify text only keeping the relevant wording [37]. This step was perceived to be useful for additional insight generation.

Table 1 Information collected from survey, renamed to variables for python analysis.

|  |  |  |
| --- | --- | --- |
| Question in survey | Variable in python code | Data type |
| N/A (Captured by survey submission) | Timestamp | datetime |
| Question one | consent\_obtained | category |
| Question two (Related to **RQ1)** | rq1\_question | integer |
| Question three (Related to **RQ2)** | rq2\_question | integer |
| Question four | rating\_response | integer |
| Question five | prompt\_entered | String |
| Question six | prompt\_obtainable\_elsewhere | category |
| Question seven (Related to **RQ3)** | rq3\_question | Integer |
| Question eight | num\_chatgpt\_uses | category |
| Question nine | rating\_comparison\_other | Integer |
| Question ten | occupation\_status | Category |
| Question eleven | education\_level | Category |
| Question twelve | professional\_sector | String |
| Question thirteen | student\_course | String |

#### 2 – Descriptive Analysis

Initial analysis summarised data points using tables and graphs. Examples included total participant count, percentages of population with specific attributes. Additionally, analysis addressed the main research questions:

1. How many participants have used ChatGPT previously? This piece of analysis was related to **RQ1**.
2. How many participants found ChatGPT enhanced their knowledge? What percentage of the population reported knowledge enhancement, and the percentage where it did not. This piece of analysis was related to **RQ2.**
3. Among participants who reported knowledge enhancement, how many now use ChatGPT for knowledge acquisition? This analysis will provide percentages for both the subset and the entire sample. This piece of analysis was related to **RQ3.**

The analysis also explored limited participant characteristics and additional questions asked for potential insights. These characteristics included – at a high-level – the participant’s occupational status, highest level of education, and either the industry the participants works within if a working professional, or the course currently studied if the participant is a student. Moreover, participants were asked additional questions such as rating the response provided form their prompt given, if the information could have been obtained elsewhere, how many times they have previously used ChatGPT, and a rating to compare the response given on their prompt, when compared with other tools to acquire the information they queried.

#### 3 – Pearson’s Correlation Coefficient

Pearson correlation assesses the strength and direction of a linear relationship between two variables, indicating their closeness [19-20]. This study examined the association between ChatGPT users who have experienced knowledge enhancement and the integration of ChatGPT into their daily knowledge acquisition routine. The `dataframe.corr('pearson')` function in pandas [21] will be employed to generate a correlation table. Additionally, the pearsonr(x,y) function from the scipy.stats package [38], was also used to confirm this, along with providing the p-value showing the statistical significance. A positive correlation was anticipated, signifying that users who enhance their knowledge with ChatGPT are inclined to incorporate it into their regular knowledge-seeking practices, supporting RH2. Furthermore, the p-value provided from the pearsonr function was expected to show a value of less than 0.05, indicating statistical significance Table 1 shows the variables used for the Pearson correlation coefficient analysis, which related to the RQ variable referred to, and supported confirmation of claims made in RH2.

Table 2 Variables used for Pearson's Correlation Coefficient Analysis.

|  |  |  |
| --- | --- | --- |
| Question in survey | Variable defined in python after pre-processing steps | RQ related to variable |
| When you used ChatGPT - based on the prompt you entered - did the response help enhance your knowledge in that area? | rq2\_question | **RQ2** |
| If ChatGPT did enhance your knowledge, have you used it more into your day-to-day routine since? | rq3\_question | **RQ3** |

#### 4 - Hypotheses Testing

To assess the statistical significance of survey results and investigate the two hypotheses, an employment of the binomial test in python was used using the scipy.stats library package [39,40]. The null hypothesis, in both cases, will assume that only 50% of the participants reported an improvement in their knowledge and the inclusion of ChatGPT as a tool for regular knowledge acquisition. Should more than half of the responses demonstrate this trend, the null hypothesis will be rejected, and the alternative hypothesis, which aligns with the research theories, would be accepted [35,36].

### Results from Analysis

The results are then displayed in both tabular, graphical – in terms of data visualisations and textual formats for the reader dependant on the context of what is being shown.

## Ethical Considerations

As research involved the perception of participants view in form of a quantitative survey study, there remained potential items that were considered. Participants' rights had been prioritised throughout the study, and all procedures will adhere to ethical guidelines outlined by The Institute of Electrical and Electronics Engineers (IEEE) [25] within their code of ethics [26].

Participants were provided a clear explanation of the study's purpose, benefits this research aims to bring and procedures involved. Informed consent was obtained from each participant before participation ensuring they can withdraw from the questionnaire at any time, confirming they can make informed decisions. Consent was documented through the online survey platform, with participants required to indicate their voluntary participation by selecting the option “I agree to these terms and wish to participate”, before the survey begun.

Participants' confidentiality were maintained. Data collected is stored and anonymised. Only the researcher and the dissertation supervisor will have access to the data, which will be used solely for research purposes. No personal identifying information was expected in the survey results but was separated from survey responses during analysis.

The study poses minimal risks to participants. However, measures were taken to mitigate potential risks, such as providing clear instructions on the survey process to avoid any confusion. No conflicts of interest that could compromise the objectivity or integrity of the research were detected.

## Action to obtain ethical approval.

To get ethical approval for this research project, a proposal was sent to the dissertation supervisor, due to the ethical consideration outlined. Whilst this study involves the use of humans to get their views on the use of ChatGPT, there is minimal to no harm or misconduct expected towards participants who engage with the survey. A signed copy of the ethical approval of research projects in online programmes form can be found in **Appendix B.**

## Limitations

Particularly as this piece delves into a novel area concerning ChatGPT limitations were uncovered and need to be addressed to provide as much transparency as possible. Due to limited time and resources, only one survey will be conducted with a limited series of mostly closed questions to gather prompt and enriching feedback. Additionally, the analysis will be limited to the collected data without comparison to other datasets, due to the lack of research in this area. Furthermore, It is important to note that bias may be introduced in this study as participants interested in the topic may be the only ones who choose to participate.