Vol.3 No.2 October 2026: 49-56
International Journal of Computer Science and Humanitarian AI (IJCSHAI)
e-ISSN: 3064-4372
DOI: 110.21512/ijcshaijournal.v3i2.16335

Comparative Evaluation of ChatGPT, Gemini, Claude, DeepSeek, and Grok for Descriptive Statistical Analysis of Likert-Scale UAT Data

Yoseph Satria Praka

Digital Multimedia Engineering Programme,
Jakarta State Polytechnic,
Depok, Indonesia 16425
yoseph.satria.praka.tik23@stu.pnj.ac.id

Abstract*— Large Language Models (LLMs) may be employed as an aid for data analysis such as calculating descriptive statistics from questionnaire responses. Nevertheless, how well the calculations of UAT data are accurate is in need of additional evaluation. The purpose of this study was to assess ChatGPT, Gemini, Claude, DeepSeek, and Grok when analyzing five-point Likert-scale UAT data; the reference software was IBM SPSS Statistics 26. The study utilized questionnaire responses from ten participants to five items. All models were provided with the same data set and prompt to calculate respondent-level scores and descriptive measures including minimum, maximum, mean, standard error, standard deviation, and variance. Then, the calculated results were compared to the corresponding SPSS results. Overall, the majority of the results obtained by the models agreed with the SPSS results; however, there were some discrepancies in the variance results for Q1, Q3, and Q4. Additionally, DeepSeek had a slightly different result for the standard deviation of the total scores. There was no additional calculation or controlled test conducted to identify why the models differed. The study generally demonstrated that the results of the five models were consistent with the SPSS generated results for the data set analyzed. However, this does not suggest that the models will yield similar results when applied to other data sets, or when utilizing alternative statistical methods. Researchers conducting future studies may want to apply larger data sets, conduct multiple replications of experiments, and apply different forms of statistical analysis to evaluate how effectively LLMs operate under varying experimental conditions. Should a study require high degrees of numeric precision, researchers may wish to compare the results produced by an LLM to those produced by a widely accepted statistical software package.

Keywords—Large Language Models; User Acceptance Testing; Descriptive Statistics; Comparative Analysis

Received: Jul. 31, 2026; received in revised form: Oct 12, 2026; accepted: Oct. 12, 2026; available online: Oct. 28, 2026.

Corresponding: yoseph.satria.praka.tik23@stu.pnj.ac.id

I. INTRODUCTION

The use of Large Language Models (LLMs) has begun to expand beyond simple tasks such as data entry and processing; they now are being used to help researchers with data processing, result interpretation, and even scientific writing [1], [2]. Models like ChatGPT [3], Gemini [4], Claude [5], DeepSeek [6], and Grok [7] are all examples of LLMs that have the ability to take in structured data, then create natural language explanations based on what it found. Because of this capability, researchers have started exploring whether or not LLMs could be used for quantitative analysis [8], [9], [10], [11]. It is important to note, however, that the numerical results created by an LLM should always be verified using standard statistical software before publishing.

The purpose of user acceptance testing (UAT) is to evaluate if a software application satisfies the needs of users and if it will work as intended [12]. Many times, in research, UAT uses Likert scale questionnaires to obtain feedback from users; this feedback can then be summarized using descriptive statistics [13]. While there are many ways to calculate descriptive statistics, it is important to double-check the calculations for correctness so that each respondent's answers, the total number of respondents, and the summary measures are reported accurately. It is possible that large language models (LLMs) could help process questionnaire data and explain the results, however, the numerical output generated by LLMs must be verified against an accurate source prior to being incorporated into additional analyses or reporting [14].

Previous research has investigated the role of LLMs in software testing and usability evaluation. One study examined the use of LLMs to generate Behavior-Driven Development (BDD) acceptance tests and reported that the quality of the generated test artifacts varied with the prompting strategy. Few-shot prompting, in particular, helped improve the accuracy of the generated outputs [15]. Another study developed an LLM-based tool to assess user experience from interaction records using Nielsen's usability heuristics. Its findings indicated that the assessments were influenced by the prompts provided to the model [16]. A preliminary study also explored the generation of usability testing tasks with LLMs and highlighted the need for expert supervision [17].

Copyright © 2026

Research has also examined LLM performance in statistical reasoning and subject-specific assessments. A comparative study evaluated six LLMs on their ability to select statistical tests for hypothesis-testing scenarios. While the models could identify appropriate tests in many cases, the quality of their explanations and statistical reasoning varied [9]. In another study, six LLMs were assessed using multiple-choice questions on blood physiology that had undergone item analysis. The results showed differences in answer accuracy and consistency when the models were used repeatedly by different users [18]. Together, these studies indicate that LLM performance can vary according to the task, the assessment criteria, and the way the models are used.

Despite these investigations, the calculation of descriptive statistics from Likert-scale questionnaire data remains a distinct task that warrants further examination. Previous studies have focused on creating automated acceptance tests, evaluating usability, statistically selecting tests for use in a given situation, and answering questions that are specific to a particular subject. None of these topics directly address the issue of whether or not LLMs generate accurate descriptive statistics based on UAT responses. By comparing several different models while keeping the same data set and instructions constant, we will be able to compare differences in the numerical output of each model and see if the numerical outputs of each model align with the descriptive statistics produced by standard statistical software. For that reason, this paper will compare five LLMs when evaluating Likert-scale UAT data and will use IBM SPSS Statistics as a control. This paper evaluates the descriptive statistic outputs generated by each model under the same experimental conditions; therefore, it does not suggest a new way to conduct analysis. Therefore, the results of this study are limited to the data set used in this study and the descriptive statistics evaluated in this study should not be used as evidence that LLMs produce consistent results across all datasets and types of statistical analyses.

The purpose of this study is to compare how well ChatGPT, Gemini, Claude, DeepSeek, and Grok calculate descriptive statistics based upon user acceptance testing (UAT) data collected through Likert-scale responses. To determine which model performs better than others in generating the same numerical output as the corresponding result produced by IBM SPSS Statistics, the performance of each model was evaluated using two research questions:

RQ1: How much of an agreement is there between the descriptive statistics produced by the five LLMs and those generated through the use of IBM SPSS Statistics for the same UAT dataset?

RQ2: What is the variation of numerical differences observed between the output generated by the five LLMs and the SPSS benchmark and how does this variation differ among the various descriptive measures being studied?

In providing a direct comparison of five different Large Language Models (LLMs) performing the exact same descriptive statistical task, this study will provide clarity on whether the results produced by the LLMs match the results produced by the use of statistical software and also identify any discrepancies in numbers that would warrant additional verification.

II. Proposed Methods

This study utilized comparative experimental research design [19], [20], [21] to analyze how five of the largest language models (LLMs) handled User Acceptance Testing (UAT) data. Rather than creating a new analytical process or software application, this research compared the performance of several LLMs based upon identical input data and instructional sets. Results produced by each of the LLMs were then assessed against reference values derived through the use of IBM SPSS Statistics 26.

The first stage of the research was a review of the existing literature about the use of LLMs in software testing, usability evaluations and the collection of quantitative data. A UAT dataset was then created by compiling questionnaire answers from 10 participants. The survey had 5 questions that could be answered using a 5-point Likert Scale; each question could have answers varying from 1 (Strongly Disagree) to 5 (Strongly Agree) [22], [23]. Data were collected solely for the purpose of this research project. Prior to evaluating the LLMs, the dataset was analyzed using IBM SPSS Statistics 26 to determine reference values such as total participant scores and descriptive statistics. These reference values provided a base line for comparing the numerical accuracy of the models [24], [25]. In addition to evaluating the numerical accuracy of the models, the researchers also evaluated if the acceptance categories created by the models were consistent with the interpretation criteria that had been defined in advance for this research project.

The experiment took place on July 25, 2026, with five models; ChatGPT (GPT-5.5), Gemini 3.6 Flash, Claude Sonnet 5, DeepSeek-V3, and Grok Fast. All five models were accessed via their respective official websites, with the publicly available free versions of each being utilized: ChatGPT (chatgpt.com), Gemini (gemini.google.com), Claude (claude.ai), DeepSeek (chat.deepseek.com) and Grok (grok.com). Each model was tested once using two separate chat sessions. The two chat sessions were not independently assessed as individual test trials; they were part of the evaluation procedure. During the initial chat session, each model performed analysis on the UAT dataset. During the second chat session, each model organized the findings from the first chat session into a tabular representation. All models had the exact same dataset and directions for conducting analysis to create comparable conditions in the evaluation setting for all five models. The experimental prompt used in this study may be found in the Data Availability section. No additional evaluation trials or independent trials were performed.

The assessment of the accuracy of both the numerical calculations performed by the models and the interpretations of those calculations was used to determine how accurate the models were. For example, to assess the accuracy of the numerical calculations, the output produced by each model was compared with the outputs produced when analyzing the same data set using SPSS. To assess the accuracy of the interpretations, the acceptance categories identified by each model were evaluated based on the predefined UAT interpretation criteria. After comparing the outputs descriptively, similarities and differences among the five models were identified. Since the study used only one data set containing data from ten participants and analyzed each model only once, the results of this study are specific to the conditions under which the study was conducted. Thus, this

study cannot be generalized to assess consistency across multiple trials or indicate general reliability of the models. Figure 1 shows the entire research process.

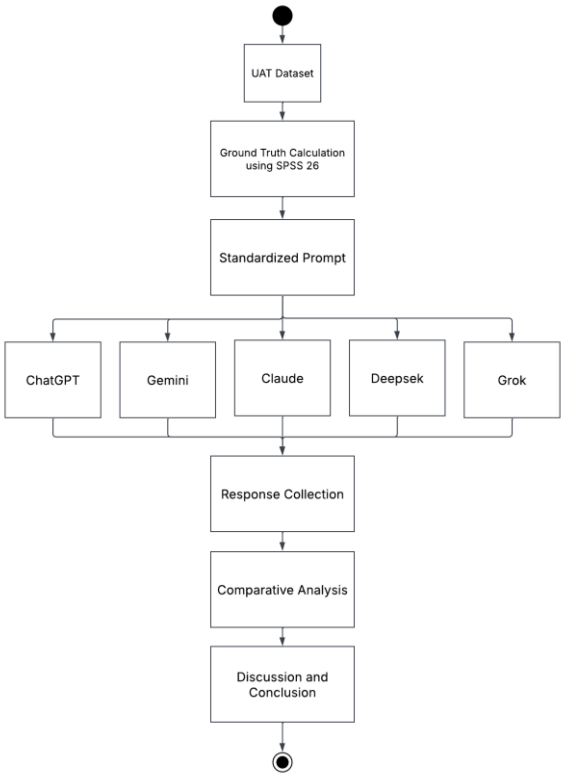


Fig. 1. Research Workflow

III. Experimental Result

A. Results of the UAT Questionnaire Using SPSS

To describe the respondents' evaluations, the UAT questionnaire data was analyzed using descriptive statistics in IBM SPSS Statistics. The analysis summarized the responses to each item by calculating the minimum and maximum values, range, total score, mean, standard deviation, and variance. These measures provide an initial picture of how respondents rated the system before the acceptance results are interpreted further. The results of this analysis are shown in Table 1.

TABLE I. Descriptive Statistical Results of the UAT Questionnaire Using IBM SPSS Statistics

Resp.	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Total*
R1	5.00	4.00	4.00	5.00	3.00	21.00
R2	5.00	5.00	5.00	5.00	4.00	24.00
R3	5.00	4.00	4.00	5.00	4.00	22.00
R4	5.00	5.00	4.00	5.00	5.00	24.00
R5	5.00	5.00	5.00	4.00	4.00	23.00
R6	5.00	5.00	5.00	4.00	4.00	23.00
R7	5.00	5.00	5.00	4.00	3.00	22.00
R8	4.00	4.00	4.00	5.00	5.00	22.00
R9	5.00	5.00	5.00	5.00	3.00	23.00
R10	4.00	5.00	5.00	4.00	4.00	22.00
Resp.	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Total*
Descriptive Statistic						
Min.	4.00	4.00	4.00	4.00	3.00	21.00
Max.	5.00	5.00	5.00	5.00	5.00	24.00
Mean	4.80	4.70	4.60	4.60	3.90	22.60
Std. Error	.133	.152	.163	.163	.233	.305
Std. Deviation	.421	.483	.516	.516	.737	.966
Variance	.178	.233	.267	.267	.544	.933

As illustrated in Table 1, the respondents' average responses to each question were on a 5-point Likert Scale; the averages ranged between 3.90 and 4.80. The highest average score was obtained for Q1 (M = 4.80, SD = 0.422) followed closely by Q2 (M = 4.70, SD = 0.483). Q3 and Q4 obtained equal average scores of 4.60 and similar standard deviations of 0.516. Conversely, Q5 had the lowest average score (M = 3.90, SD = 0.738) and also the largest standard deviation indicating that there were greater variations among respondents' answers to Q5 as compared to other questions. Total scores ranged from 21 to 24 with an average score of 22.60 and a standard deviation of 0.966. Overall, these statistics represent the response patterns observed in the data set and will serve as baseline metrics against which the results produced by the five LLMs will be compared.

B. Results of the UAT Questionnaire Using ChatGPT

The descriptive statistical results generated by ChatGPT are presented in Table 2.

TABLE II. Descriptive Statistical Results of the UAT Questionnaire Using ChatGPT

Resp.	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Total*
R1	5	4	4	5	3	21
R2	5	5	5	5	4	24
R3	5	4	4	5	4	22
R4	5	5	4	5	5	24
R5	5	5	5	4	4	23
R6	5	5	5	4	4	23
R7	5	5	5	4	3	22
R8	4	4	4	5	5	22
R9	5	5	5	5	3	23
R10	4	5	5	4	4	22
Descriptive Statistic						
Min.	4	4	4	4	3	21
Max.	5	5	5	5	5	24
Mean	4.80	4.70	4.60	4.60	3.90	22.60
Std. Error	0.133	0.152	0.163	0.163	0.233	0.305
Std. Deviation	0.421	0.483	0.516	0.516	0.737	0.966
Variance	0.177	0.233	0.266	0.266	0.544	0.933

Table 2 shows the descriptive statistics calculated by ChatGPT from the UAT questionnaire data. The mean scores for the five items ranged from 3.90 to 4.80. Q1 had the highest mean at 4.80, followed by Q2 at 4.70. Both Q3 and Q4 recorded a mean of 4.60, while Q5 had the lowest mean at 3.90. The standard deviation values ranged from 0.421 to 0.737, indicating that responses to Q5 varied the most among the five items. For the total scores, the minimum was 21 and the maximum was 24, with a mean of 22.60 and a standard deviation of 0.966. Overall, these results describe the distribution of responses and the variation in scores calculated by ChatGPT.

C. Results of the UAT Questionnaire Using Gemini

The descriptive statistical results generated by Gemini are presented in Table 3.

TABLE III. DESCRIPTIVE STATISTICAL RESULTS OF THE UAT QUESTIONNAIRE USING GEMINI

Resp.	***Q1***	***Q2***	***Q3***	***Q4***	***Q5***	***Total***
R1	5	4	4	5	3	21
R2	5	5	5	5	4	24
R3	5	4	4	5	4	22
R4	5	5	4	5	5	24
R5	5	5	5	4	4	23
R6	5	5	5	4	4	23
R7	5	5	5	4	3	22
R8	4	4	4	5	5	22
R9	5	5	5	5	3	23
R10	4	5	5	4	4	22
Descriptive Statistic						
Min.	4	4	4	4	3	21
Max.	5	5	5	5	5	24
Mean	4.80	4.70	4.60	4.60	3.90	22.60
Std. Error	0.133	0.152	0.163	0.163	0.233	0.305
Std. Deviation	0.421	0.483	0.516	0.516	0.737	0.966
Variance	0.177	0.233	0.266	0.266	0.544	0.933

Table 3 summarizes the descriptive statistics produced by Gemini using the UAT questionnaire data. The mean scores for the five items were in the range of 3.90 - 4.80. Q1 had the highest mean score (4.80) followed by Q2 (4.70). Both Q3 and Q4 had a mean score of 4.60. Q5 had the lowest mean score (3.90). The standard deviation for all five items was between 0.421 and 0.737. Q5 had the largest variance in response. Total scores ranged from 21 - 24, with a mean score of 22.60 and a standard deviation of 0.966. These data provide an indication as to how the responses were distributed and the types of differences that can be anticipated from these results.

D. Results of the UAT Questionnaire Using Claude

The descriptive statistical results generated by Claude are presented in Table 4.

TABLE IV. DESCRIPTIVE STATISTICAL RESULTS OF THE UAT QUESTIONNAIRE USING CLAUDE

Resp.	***Q1***	***Q2***	***Q3***	***Q4***	***Q5***	***Total***
R1	5	4	4	5	3	21
R2	5	5	5	5	4	24
R3	5	4	4	5	4	22
R4	5	5	4	5	5	24
R5	5	5	5	4	4	23
R6	5	5	5	4	4	23
R7	5	5	5	4	3	22
R8	4	4	4	5	5	22
R9	5	5	5	5	3	23
R10	4	5	5	4	4	22
Descriptive Statistic						
Min.	4	4	4	4	3	21
Max.	5	5	5	5	5	24
Mean	4.80	4.70	4.60	4.60	3.90	22.60
Std. Error	0.133	0.152	0.163	0.163	0.233	0.305
Std. Deviation	0.421	0.483	0.516	0.516	0.737	0.966
Variance	0.177	0.233	0.266	0.266	0.544	0.933

Table 4 provides a summary of the descriptive data obtained from Claude's analysis of the UAT questionnaire results. All five items had average scores ranging from 3.90 to 4.80. Item Q1 had the highest average score (4.80), followed by Q2 (4.70). Items Q3 and Q4 both had the same average score (4.60). Item Q5 had the lowest average score (3.90). The standard deviations of the items were between 0.421 and 0.737; item Q5 showed the greatest variability in how participants responded. The total score range was 21 to 24 with a mean score of 22.60 and a standard deviation of 0.966. These statistics illustrate the response distribution and variation in scores reported by Claude.

E. Results of the UAT Questionnaire Using Deepsek

The descriptive statistical results generated by Deepsek are presented in Table 5.

TABLE V. DESCRIPTIVE STATISTICAL RESULTS OF THE UAT QUESTIONNAIRE USING DEEPSEK

Resp.	***Q1***	***Q2***	***Q3***	***Q4***	***Q5***	***Total***
R1	5	4	4	5	3	21
R2	5	5	5	5	4	24
R3	5	4	4	5	4	22
R4	5	5	4	5	5	24
R5	5	5	5	4	4	23
R6	5	5	5	4	4	23
R7	5	5	5	4	3	22
R8	4	4	4	5	5	22
R9	5	5	5	5	3	23
R10	4	5	5	4	4	22

Resp.	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Total*
Descriptive Statistic						
Min.	4	4	4	4	3	21
Max.	5	5	5	5	5	24
Mean	4.80	4.70	4.60	4.60	3.90	22.60
Std. Error	0.133	0.152	0.163	0.163	0.233	0.305
Std. Deviation	0.421	0.483	0.516	0.516	0.737	0.964
Variance	0.177	0.233	0.266	0.266	0.544	0.933

Table 5 shows the descriptive statistics calculated by DeepSeek from the UAT questionnaire dataset. The mean scores across the five items ranged from 3.90 to 4.80. Q1 recorded the highest mean of 4.80, followed by Q2 at 4.70. Q3 and Q4 shared the same mean score of 4.60, while Q5 had the lowest mean at 3.90. The standard deviation values for the questionnaire items ranged from 0.421 to 0.737, with Q5 showing the greatest variation in responses. For the total scores, DeepSeek reported a minimum of 21, a maximum of 24, and a mean of 22.60, with a standard deviation of 0.964. These results summarize the distribution of responses and score variation calculated by DeepSeek.

F. Results of the UAT Questionnaire Using Grok

The descriptive statistical results generated by Grok are presented in Table 6.

TABLE VI. DESCRIPTIVE STATISTICAL RESULTS OF THE UAT QUESTIONNAIRE USING GROK

Resp.	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Total*
R1	5	4	4	5	3	21
R2	5	5	5	5	4	24
R3	5	4	4	5	4	22
R4	5	5	4	5	5	24
R5	5	5	5	4	4	23
R6	5	5	5	4	4	23
R7	5	5	5	4	3	22
R8	4	4	4	5	5	22
R9	5	5	5	5	3	23
R10	4	5	5	4	4	22
Descriptive Statistic						
Min.	4	4	4	4	3	21
Max.	5	5	5	5	5	24
Mean	4.80	4.70	4.60	4.60	3.90	22.60
Std. Error	0.133	0.152	0.163	0.163	0.233	0.305
Std. Deviation	0.421	0.483	0.516	0.516	0.737	0.966
Variance	0.177	0.233	0.266	0.266	0.544	0.933

Table 6 presents the descriptive statistics generated by Grok using the UAT questionnaire data. The mean scores ranged from 3.90 to 4.80 across the five items. Q1 had the highest mean score of 4.80, followed by Q2 at 4.70. Q3 and Q4 both recorded a mean of 4.60, while Q5 had the lowest mean at 3.90. The standard deviation values ranged from 0.421 to 0.737, with Q5 showing the largest variation in responses among the items. The total scores ranged from 21 to 24, with a mean of 22.60 and a standard deviation of 0.966. Overall, these statistics describe the response distribution and score variation reported by Grok.

G. Comparative Analysis of Descriptive Statistical Results

To compare the descriptive statistical results produced by the five evaluated large language models (LLMs), the outputs from ChatGPT, Gemini, Claude, DeepSeek, and Grok were checked against the reference results from IBM SPSS Statistics. The comparison examined whether the models produced matching respondent-level scores and descriptive measures, including the minimum, maximum, mean, standard error, standard deviation, and variance. The results of this comparison are presented in Table 7.

TABLE VII. COMPARISON OF DESCRIPTIVE STATISTICAL RESULTS GENERATED BY FIVE LLMS AND IBM SPSS STATISTICS

Stat. Measure	*SPSS*	*ChatGPT*	*Gemini*	*Claude*	*Deep Seek*	*Grok*
Resp. Scores	✓	✓	✓	✓	✓	✓
Min.	✓	✓	✓	✓	✓	✓
Max.	✓	✓	✓	✓	✓	✓
Mean	✓	✓	✓	✓	✓	✓
Std.Error	✓	✓	✓	✓	✓	✓
SD (Q1 - Q5)	✓	✓	✓	✓	✓	✓
SD (Total)	0.966	✓	✓	✓	0.964	✓
Variance (Q1)	0.178	0.177	0.177	0.177	0.177	0.177
Variance (Q2)	✓	✓	✓	✓	✓	✓
Variance (Q3)	0.267	0.266	0.266	0.266	0.266	0.266
Variance (Q4)	0.267	0.266	0.266	0.266	0.266	0.266
Variance (Q5)	✓	✓	✓	✓	✓	✓
Variance (Total)	✓	✓	✓	✓	✓	✓

Table 7 compares the results generated by the five LLMs with the benchmark obtained using IBM SPSS Statistics 26. All models were given the same User Acceptance Testing (UAT) dataset, allowing their output to be examined under the same conditions. The comparison included respondent-level scores and all selected descriptive statistical measures.

The results indicate that all five models produced respondent-level scores and minimum, maximum, and mean values that matched the SPSS results. Their standard error values were also consistent with the benchmark for all five questionnaire items and the total scores. The standard deviation values for Q1 to Q5 matched the SPSS results. However, a difference appeared in the standard deviation of the total scores calculated by DeepSeek. While SPSS reported a value of 0.966, DeepSeek returned 0.964. The other four models produced the same value as the benchmark, making this a small but identifiable numerical difference.

The comparison also revealed differences in the variance values for Q1, Q3, and Q4. SPSS reported values of 0.178, 0.267, and 0.267, respectively, whereas all five models returned 0.177, 0.266, and 0.266. The variance values for Q2,

Q5, and the total scores matched the benchmark when reported to three decimal places. Each of the differing values had an absolute difference of 0.001.

Taken together, the results show that the five LLMs produced descriptive statistics that largely agreed with the SPSS benchmark for this dataset. ChatGPT, Gemini, Claude, and Grok differed from SPSS only in the variance values for Q1, Q3, and Q4. DeepSeek showed similar differences; however, it also differed from the other model in terms of the standard deviation of its total score. Since the current data does not permit us to know why the results vary, we must examine the methods of determining the total score to determine what accounts for the variation. Additionally, please be aware that the scope of this research project was restricted to evaluating a single model on a single dataset composed of responses from 10 participants, and each model was evaluated only once. As such, although our evaluations indicate how the two models fared under the unique circumstances of this investigation, we cannot state whether each model would fare better or worse than demonstrated herein if evaluated on other datasets or if other types of statistical analyses were conducted on them.

H. Discussion

In this study, the descriptive statistical outputs produced by five large language models (LLMs), including ChatGPT, Gemini, Claude, DeepSeek, and Grok, were compared to those produced by IBM SPSS Statistics 26, which served as the gold standard for assessing the numerical outputs of each model. A User Acceptance Testing (UAT) dataset, containing five-point Likert scale questions, was used to generate these descriptive statistics. While previous studies have examined the use of LLMs for generating software tests, evaluating usability, and providing domain-specific answers to questions, the focus of this study was on assessing the accuracy with which each model calculated descriptive statistics from questionnaire data.

The study found that the five LLMs produced respondent-level scores and minimum, maximum, and mean values that were very similar to the SPSS output. Additionally, the five LLMs produced similar standard errors and standard deviations for Q1-Q5 as well. One area where the two programs differed was with respect to the standard deviation of the overall score. DeepSeek's overall score had a standard deviation of.964 whereas SPSS' overall score had a standard deviation of.966. In addition to differences in the standard deviation of overall scores, there were also differences in the variances of Q1, Q3, and Q4. At the three-decimal place level of analysis that was performed in this study, the difference between the LLMs and SPSS for each variable analyzed was.001. Therefore, it appears that the LLMs are producing results that are similar to the benchmark but still contain small differences in terms of numeric values.

The results can be considered alongside earlier research on LLM applications in software engineering and usability evaluation Study [15] investigated the generation of Behavior-Driven Development (BDD) acceptance tests and found that GPT-based models produced the highest-quality outputs. Study [16] examined automated User Experience (UX) testing by analyzing user interaction logs and interface screenshots through Nielsen's usability heuristics. Similarly, study [17] reported that LLMs can generate test tasks that are similar to those created by human experts, and therefore, they can be used to help with usability testing. While the previous studies evaluated how LLMs could assist in the areas of software testing and usability engineering, the current study examined another area of the evaluation process -- analysis of quantitative UAT questionnaire responses collected after testing. Therefore, this study expands on the existing research in this area by determining if the descriptive statistics generated using an LLM are consistent with the analysis performed using traditional statistical analysis software.

The results of the current study are related to a previous study [9] in which researchers evaluated how well different large language models (LLMs) performed at identifying appropriate statistical tests to use for hypothesis testing, as well as providing explanations as to why they selected those particular tests. This is different from the focus of the present study, which was to assess the accuracy of the descriptive statistical calculations generated by the LLMs. While both studies evaluate the use of LLMs for conducting statistical analyses, they used different evaluation tasks. Therefore, it is important to interpret the results of this study in light of what it set out to accomplish. For example, simply because an LLM agreed with SPSS on one UAT data set does not mean that it will consistently perform as well in all other statistical applications.

In terms of how these results impact practice, the results show that LLMs can provide an early assessment of questionnaire data. With natural-language prompts, the models can produce summaries of respondent scores and generate descriptive statistics commonly used. In addition, the results of this research show that the accuracy of quantitative results produced using LLMs must be verified. Therefore, if you require an accurate quantitative analysis, you must continue to use traditional statistical software, such as IBM SPSS Statistics. While LLMs may help with the analysis process, their results do not support using them in place of traditional statistical software.

The findings of the present study should be considered with caution because of several limitations in interpreting the results. First, the present study used a single UAT data set (comprising 10 participants' responses to 5 survey questions). Therefore, it is possible that the results do not accurately reflect the way models perform when evaluated on larger data sets or on data sets with differing response distributions. Second, each model was evaluated only once. Although the two chats utilized during the procedure were used for different purposes (i.e., Chat 1 was used to analyze the data and Chat 2 was used to organize the results into a table), they were not independent trials or repeated evaluations. Consequently, this research does not evaluate whether the output generated by each model remains consistent over time, nor does it evaluate whether the output generated by each model is reproducible. Third, the present study used only descriptive statistics and did not use any inferential statistical techniques (such as hypothesis testing, correlation analysis, or regression analysis). Fourth, the present study evaluated only five publicly available LLMs. Therefore, it is possible that the findings of the present study do not apply to other models or other versions of the same models. Fifth, although there were some numerical differences identified among the models, there was no additional analysis performed to determine what factors contributed to those differences. Additionally, there were no controlled experiments conducted to determine whether the performance of each model was affected by

calculation and/or rounding settings. Therefore, it is unknown why those numerical differences existed.

Future studies could help resolve some of the issues that were identified here by examining much larger samples of data from a wider variety of distribution types and running multiple independent trials for every type of model. Quantitative methods to assess the level of numerical agreement between LLM generated outputs and reference calculations would allow for a more structured comparison of the models. Additional experiments could determine if there are any differences in the results due to changes in prompt language, model version, or how outputs are formatted. The study can also be expanded to evaluate other quantitative analysis tasks such as inferential statistics.

IV. Conclusion

This study evaluated five large language models (LLMs), ChatGPT, Gemini, Claude, DeepSeek, and Grok, by comparing their descriptive statistical calculations for five-point Likert-scale User Acceptance Testing (UAT) data with the results obtained using IBM SPSS Statistics 26. The comparison found that all five models generated respondent-level scores and descriptive statistics that matched the SPSS results for minimum, maximum, mean, standard error, and item-level standard deviation. However, differences were observed in the variance values for Q1, Q3, and Q4, as well as in DeepSeek's standard deviation for the total scores. Since the types of errors identified in the first paragraph have not yet been evaluated through the use of multiple calculations or controlled experiments regarding the accuracy of a computer's arithmetic and its rounding of numbers, the cause of these types of errors is still unknown. Both of the programs that we evaluated yielded similar results for the data sets that we used; therefore, we can state that both programs will likely yield similar results. While this consistency alone does not show that either program has overall reliability in terms of conducting statistical analyses.

In addition to exploring the potential application of LLMs for acceptance testing, usability evaluation, and statistical interpretation as previously discussed; this study explores the potential use of LLMs for descriptive statistical calculations with data obtained through completed UAT questionnaires. Previous studies have shown that there is a large variability in the performance of LLMs depending on the type of task they are asked to complete, the prompts provided to direct their responses, and how well the LLM performs compared to a benchmark. Results from the five different models tested in this study showed that the results obtained using the five models were generally consistent with those obtained using SPSS. This suggests that LLMs may be a reasonable alternative for conducting preliminary descriptive analysis when constrained by the limitations of this study. However, before presenting findings requiring a high degree of precision, it would be necessary to compare the numerical output generated by the LLM with a known statistical software program (such as SPSS) to determine if the numerical output is accurate. There were also limitations associated with the findings of this study. The limitations we considered were as follows: we only analyzed one dataset with responses of ten participants regarding five questionnaire items; Participants completed the questionnaire only once; Only descriptive statistics were computed; Thus, it was unclear whether these results could be replicated by analyzing larger datasets, different response distributions, by having multiple attempts per participant or with other statistical analyses.

Larger data sets with more heterogeneous responses as well as the experimental design (repeating independent trials) could be an essential module of future research based on the current one. Other experimental designs may also examine whether the phrasing of the prompts; which version of the model is used; or model instructions for numeric formatting influences how accurately it can perform calculations. Along with evaluating whether the model generates similar outputs using different types of inferential statistics (e.g., hypothesis testing, correlation, regression, non-parametric tests), future studies will help establish the extent to which the model is able to conduct various types of analyses. Therefore, collectively, these areas of investigation should contribute toward a clearer understanding of when LLM-generated statistics yield comparable results to those produced by traditional statistical software, as well as how LLMs can facilitate the quantitative research process while acknowledging the limitations inherent in the models.

Author's contribution

The author created all parts of the work including developing the experimental methodology, collecting and organizing the data for the experiment, conducting the experiment itself, analyzing the results of the experiment, and ultimately writing the final version of the manuscript.

Availability data and materials

The data used in this study to support the results found through this research can be accessed through *s.pnj.ac.id/LLMUAT*. Within this repository you will find the original UAT data set, IBM SPSS Statistics results, descriptive statistics created using ChatGPT, Gemini, Claude, DeepSeek, and Grok, images for each analysis, and the web links to each experimental conversation. Additionally, some additional data supporting the research can be requested from the author who conducted the study.

References

[1] R. Watkins, "Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows," *AI Ethics*, vol. 4, no. 4, pp. 969–974, Nov. 2024, doi: 10.1007/s43681-023-00294-5.

[2] N. L. Rane, A. Tawde, S. P. Choudhary, and J. Rane, "Contribution and performance of ChatGPT and other Large Language Models (LLM) for scientific and research advancements: a double-edged sword," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 05, no. 10, pp. 875–899, 2023, doi: 10.56726/irjmets45213.

[3] D. Prandner, D. Wetzelhütter, and S. Hese, "ChatGPT as a data analyst: an exploratory study on AI-supported quantitative data analysis in empirical research," *Front. Educ.*, vol. 9, Jan. 2025, doi: 10.3389/feduc.2024.1417900.

[4] E. Supriyadi and S. Pahmi, "Exploring the Potential of AI Tools in Education: A Thematic Analysis of Gemini.ai," *Gema Wiralodra*, vol. 15, no. 3, Dec. 2024, doi: 10.31943/gw.v15i3.771.

[5] G. Ling, S. Zhong, and R. Huang, "Agent Skills: A

Data-Driven Analysis of Claude Skills for Extending Large Language Model Functionality," *arXiv*, Feb. 2026.

[6] Z. Deng *et al.*, "Exploring DeepSeek: A Survey on Advances, Applications, Challenges and Future Directions," *IEEE/CAA J. Autom. Sin.*, vol. 12, no. 5, pp. 872–893, May 2025, doi: 10.1109/JAS.2025.125498.

[7] E.R. Hurtado-Terranova, "IMPACT OF GROK AI ON THE DEVELOPMENT OF SCIENTIFIC WRITING SKILLS IN HIGHER EDUCATION STUDENTS," *Genet. Mol. Res.*, vol. 25, no. 7S, pp. 1–5, 2026, doi: https://doi.org/10.4238/ddfv0549.

[8] L. Xie, C. Zheng, H. Xia, H. Qu, and C. Zhu-Tian, "WaitGPT: Monitoring and Steering Conversational LLM Agent in Data Analysis with On-the-Fly Code Visualization," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA: ACM, Oct. 2024, pp. 1–14. doi: 10.1145/3654777.3676374.

[9] M. Shukla, D. Pandey, S. Kaur, M. Agarwal, A. Goyal, and H. Sharma, "Evaluating the Accuracy and Explanatory Quality of Large Language Models ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat in Statistical Test Selection for Hypothesis Testing Decisions," *Cureus*, Oct. 2025, doi: 10.7759/cureus.94949.

[10] M. E. D. C. Souza and L. Weigang, "Grok, Gemini, ChatGPT and DeepSeek: Comparison and Applications in Conversational Artificial Intelligence," *Intell. Artif.*, pp. 1–7, 2025, doi: 10.5281/zenodo.14885243.

[11] S. Altmäe, A. Sola-Leyva, and A. Salumets, "Artificial intelligence in scientific writing: a friend or a foe?," *Reprod. Biomed. Online*, vol. 47, no. 1, pp. 3–9, Jul. 2023, doi: 10.1016/j.rbmo.2023.04.009.

[12] R. T. Y. Tong, Y. K. Yuan, N. W. Dong, and R. K. Ramasamy, "A Review: Methods of Acceptance Testing," 2022, pp. 76–86. doi: 10.2991/978-94-6463-080-0_7.

[13] W. Wulandari, N. Nofiyani, and H. Hasugian, "USER ACCEPTANCE TESTING (UAT) PADA ELECTRONIC DATA PREPROCESSING GUNA MENGETAHUI KUALITAS SISTEM," *J. Mhs. Ilmu Komput.*, vol. 4, no. 1, pp. 20–27, Mar. 2023, doi: 10.24127/ilmukomputer.v4i1.3383.

[14] D. Meva and H. Kukadiya, "Performance Evaluation of Large Language Models: A Comprehensive Review," *Int. Res. J. Comput. Sci.*, vol. 12, no. 03, p. 109, Mar. 2025, doi: 10.26562/irjcs.2025.v1203.03.

[15] S. Karpurapu *et al.*, "Comprehensive Evaluation and Insights Into the Use of Large Language Models in the Automation of Behavior-Driven Development Acceptance Test Formulation," *IEEE Access*, vol. 12, pp. 58715–58721, 2024, doi: 10.1109/ACCESS.2024.3391815.

[16] N.-L. Hsueh, H.-J. Lin, and L.-C. Lai, "Applying Large Language Model to User Experience Testing," *Electronics*, vol. 13, no. 23, p. 4633, Nov. 2024, doi: 10.3390/electronics13234633.

[17] M. Calvano, A. Curci, R. Lanzilotti, A. Piccinno, and A. Ragone, "Leveraging Large Language Models for Usability Testing: a Preliminary Study," in *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, New York, NY, USA: ACM, Mar. 2025, pp. 78–81. doi: 10.1145/3708557.3716341.

[18] M. Agarwal, P. Sharma, and P. Wani, "Evaluating the Accuracy and Reliability of Large Language Models (ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat) in Answering Item-Analyzed Multiple-Choice Questions on Blood Physiology," *Cureus*, Apr. 2025, doi: 10.7759/cureus.81871.

[19] A. Alsajri, H. A. Salman, and A. Steiti, "Generative Models in Natural Language Processing: A Comparative Study of ChatGPT and Gemini," *Babylonian J. Artif. Intell.*, vol. 2024, pp. 134–145, Nov. 2024, doi: 10.58496/BJAI/2024/015.

[20] K. Shahzad and S. Iqbal, "Comparative Analysis of ChatGPT, DeepSeek, and Gemini for Automated Code Generation," in *2025 18th International Conference on Engineering of Modern Electric Systems (EMES)*, IEEE, May 2025, pp. 1–4. doi: 10.1109/EMES65692.2025.11045587.

[21] D. M. Colaco and D. Antao, "Exploring self-learning efficacy: a comparative analysis of ChatGPT and Gemini," *Interact. Technol. Smart Educ.*, vol. 22, no. 4, pp. 688–698, Oct. 2025, doi: 10.1108/ITSE-12-2023-0251.

[22] A. P. Kusuma and A. Yufron, "ANALYSIS OF USER ACCEPTANCE TESTING ON A SHIPPING APPLICATION TO DETERMINE THE QUALITY OF THE SYSTEM," *Antivirus J. Ilm. Tek. Inform.*, vol. 18, no. 2, pp. 234–243, Nov. 2024, doi: 10.35457/antivirus.v18i2.4002.

[23] N. Dufey-Portilla *et al.*, "Assessing the validity of ChatGPT-4o and Google Gemini Advanced when responding to frequently asked questions in endodontics," *J. Appl. Oral Sci.*, vol. 33, 2025, doi: 10.1590/1678-7757-2025-0321.

[24] D. Zhang *et al.*, "DataSciBench: An LLM Agent Benchmark for Data Science," in *Findings of the Association for Computational Linguistics: ACL 2026*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2026, pp. 3685–3728. doi: 10.18653/v1/2026.findings-acl.181.

[25] V. Najahah and U. Pujianto, "Automatic Classification of Artificial Intelligence Generated Question Difficulty Levels," *Indones. J. Innov. Stud.*, vol. 27, no. 1, Jan. 2026, doi: 10.21070/ijins.v27i1.1880.