

## Evaluating the Performance of ChatGPT in Management Information Systems Exam: A Comparative Analysis of ChatGPT-3.5 and 4.0

Ruchdee Binmad<sup>1\*</sup>

Received: 27-07-2024, Revised: 17-11-2024, Accepted: 02-01-2025

### Abstract

The study aims to assess the accuracy of ChatGPT by evaluating its performance on various types of questions from a Management Information Systems (MIS) exam. The study specifically focuses on comparing the performance of different ChatGPT models, namely ChatGPT-3.5 and ChatGPT-4.0. Researchers retrieved 200 questions from a test bank covering five chapters of an MIS textbook, encompassing multiple-choice, true/false, and essay questions. Forty questions were chosen for each chapter, with each question assigned a level of difficulty. The questions were queried using ChatGPT versions 3.5 and 4.0 via the OpenAI website. The study quantitatively assessed the answers using two query formations: the original question without prompt formation and a modified question with a prompt. Specifically, the ROUGE-L score, a Longest Common Subsequence (LCS) oriented measure, was used to represent the accuracy score for essay answers. The findings of this study revealed that ChatGPT-4.0 outperforms ChatGPT-3.5 in MIS questions with a success rate of 70.58%, compared to 63.42% for ChatGPT-3.5. This aligns with previous studies showing high scores on various exams. However, ChatGPT-4.0 excels in handling higher difficulty levels but performs less effectively at low difficulty levels. The study highlights the potential applications and challenges of using comprehensive language models in educational environments. As Generative AI advances, this study serves as a foundation for future research into the responsible and efficient integration of AI models in various domains. The study also examines issues regarding the potential impact of Generative AI on the credibility of MIS exams and emphasizes the need for further investigation.

**Keyword:** ChatGPT; Management information systems; AI in education

---

<sup>1</sup> Faculty of Management Sciences, Prince of Songkla University

\* Corresponding Author, E-mail: ruchdee.bi@psu.ac.th

## Introduction

Artificial intelligence (AI) has undergone substantial advancements in recent decades, shifting from theoretical exploration to tangible implementations that directly influence everyday existence. Artificial intelligence refers to a wide range of technologies that aim to imitate human intelligence by using algorithms integrated into a dynamic computing system. The core of AI lies in its capacity to analyze data, acquire knowledge from it, and make well-informed choices based on the acquired insights (Russell & Norvig, 2021). AI technologies have infiltrated diverse industries such as education, healthcare, finance, and autonomous driving, showcasing their adaptability and capacity for transformation.

A subset of artificial intelligence models, generative artificial intelligence (GenAI) is capable of generating new content by analyzing and learning patterns and structures from pre-existing data. These models are capable of producing content in a wide range of domains, including text, images, music, and more (Feuerriegel et al., 2024; Lim et al., 2023; Ray, 2023). Generative AI models employ neural networks and deep learning techniques to comprehend, generate, and analyze content that closely resembles human-generated outputs. Among the tools under consideration, ChatGPT, a text tool propelled by artificial intelligence and developed by OpenAI (OpenAI, 2023), has distinguished itself as a powerful instrument that finds utility across diverse domains.

ChatGPT is a sophisticated language model that utilizes the transformer architecture introduced by (Vaswani et al., 2017). It was launched on November 30, 2022 (OpenAI, 2023). This language model has the capability to offer answers and assistance with different tasks (Cribben & Zeinali, 2023; Eulerich et al., 2023; Ray, 2023; Rudolph et al., 2023; Shahriar & Hayawi, 2023). ChatGPT was developed on a large language model (LLM) architecture called the Generative Pre-trained Transformer (GPT). These models are regarded as autoregressive language models that employ deep learning techniques to generate human-like text. ChatGPT has undergone training using a wide-ranging dataset obtained from books, websites, and other textual sources. This training enables it to produce coherent and contextually appropriate responses in many subject areas. The technology showcases notable progress in tasks related to natural language processing and has found applications in various services, highlighting its adaptability and the possibilities for interaction between humans and artificial intelligence (Brown et al., 2020). Presently, the most widely used version of ChatGPT is founded on the

GPT-3.5 architecture, while access to ChatGPT on the GPT-4 model is currently restricted and requires explicit approval from OpenAI (Vaughan-Nichols & Diaz, 2024).

The responses provided by ChatGPT typically address both broad and specific queries, thus meeting the expectations of instructors and educators who assert that the system-generated replies are sufficient and satisfactory for achieving a high score. Additionally, it utilizes a fine-tuning technique that enables the model to adjust to different tasks by adjusting the weights of the pre-trained model using a small amount of task-specific data. Consequently, the model has the potential to generate more precise and pertinent responses tailored to specific use cases. However, there have been instances of mistakes in answering questions on various topics (Borji, 2023; Terwiesch, 2023). Therefore, it is crucial to take into account the possible disadvantages and constraints of the technology to guarantee its responsible and ethical use.

The study utilized Management Information Systems (MIS) questions to assess the accuracy of ChatGPT, given its dual function in business education. MIS can serve as a fundamental course, imparting essential corporate expertise on utilizing information technology for making informed decisions. Alternatively, MIS can be a distinct field of study within business/management schools, providing a focused curriculum for individuals who wish to gain comprehensive knowledge of information systems and their practical use in different business settings (AACSB, 2024). The significance of MIS goes beyond the simple act of gathering and storing data. It involves the analysis, manipulation, and distribution of data to generate meaningful insights, integrating advanced technologies like artificial intelligence, big data analytics, and cloud computing, with the aim of influencing business decision-making (Laudon & Laudon, 2020).

## Purpose

This study aims to assess the accuracy of ChatGPT by evaluating its performance on various types of questions from the MIS exam. The study specifically focuses on comparing the performance of different ChatGPT models, namely ChatGPT-3.5 and ChatGPT-4.0. Additionally, the study emphasizes the possible uses and challenges of implementing extensive language models in educational settings. As the field of Generative AI advances, our findings provide a basis for future studies on the responsible and efficient incorporation of AI models in other

fields. The study also addresses concerns regarding the potential influence of Generative AI on the credibility of MIS exams and highlights the necessity for additional research.

## Related Work

OpenAI's ChatGPT, an advanced artificial intelligence large language model, has undergone extensive testing in a variety of academic and professional disciplines, revealing both its vast potential and some of its limitations. (Mooney et al., 2023) conducted an assessment of ChatGPT's performance in the domain of geographic information systems (GIS). Their findings revealed that GPT-3.5 and GPT-4 achieved scores of 63.3% and 88.3%, respectively, on an introductory GIS exam. These results indicate a substantial disparity in performance across different versions of ChatGPT. In line with the findings of (AL-Qadri and Ahmed, 2023). ChatGPT demonstrated an accuracy rate of 93% when responding to statistical questions, although it sometimes needed numerous attempts to reach the accurate response. This indicates a considerable comprehension of statistical principles, albeit with occasional challenges in achieving correct responses on the first try.

In the field of computer science, (Bordt and Luxburg, 2023) evaluated the performance of ChatGPT in an undergraduate algorithms and data structures exam. ChatGPT successfully completed the exam, scoring 20.5 out of 40 points, demonstrating its capability to handle challenging tasks such as university exams. Nevertheless, the authors acknowledged that the training data for ChatGPT may have contained questions that were structurally similar, which poses a challenge in determining whether the model genuinely understands computer science concepts. (Hajj and Sah, 2023) investigated the impact of ChatGPT in a PHP programming course by conducting user studies at two distinct universities in North Cyprus. The study revealed that ChatGPT had a notable beneficial influence on coding questions, a moderately favorable impact on classical and true/false questions, but no discernible difference on multiple-choice questions. On the whole, user attitudes and post-questionnaire responses validated these findings, demonstrating ChatGPT's capability as an assistive tool in educational settings, specifically for coding assignments.

In their study, (Frieder et al., 2023) evaluated the mathematical proficiency of ChatGPT and determined that it excelled as a mathematical assistant, particularly in the areas of fact-based queries, functioning as a mathematical search engine, and serving as an interface for

accessing a knowledge base. Nevertheless, the model's performance was deemed inferior to that of a graduate student, highlighting its limits in tackling mathematics at the graduate level. de (Winter, 2023) examined the performance of ChatGPT-3.5 and 4.0 on Dutch national high school exams, specifically in the area of English reading comprehension. GPT-3.5 achieved an average score of 7.3 on these exams, while GPT-4.0 increased this score to 8.3. The study underscored the capacity of ChatGPT in the educational domain, accentuating its advantages while also acknowledging limitations and challenges pertaining to its application in educational assessments.

Medical education research conducted by (Kung et al., 2023 and Gilson et al., 2023) assessed the efficacy of ChatGPT in achieving the performance standards for the USMLE Steps 1, 2, and 3. Their results revealed that ChatGPT had the potential to serve as a training tool for medical students who lacked specialized preparation. (Antaki et al., 2023) proposed that in specialized domains like ophthalmology, the performance of ChatGPT might be improved by incorporating domain-specific pre-training, despite the promising results.

Moreover, in business education, (Terwiesch, 2023 and Eulerich et al., 2023) investigated the potential of ChatGPT in MBA core courses and accounting certification tests. The model showed competence in basic operations and process analysis but encountered difficulties when dealing with intricate computations and advanced topics. (Shahriar & Hayawi, 2023) conducted a scoping review to investigate the development, technology, applications, and constraints of ChatGPT. They analyzed existing literature and its applications in several domains such as healthcare, education, and customer service. The study revealed that ChatGPT demonstrated the potential to produce responses that closely mimicked those of humans, particularly in the fields of medical diagnostics and educational support. However, it also showed deficiencies in handling complex reasoning, lacked emotional understanding, and was limited by the data it was trained on. These issues raised concerns related to privacy, dissemination of false information, and the need for human supervision in the use of AI.

Engineering domains frequently require the utilization of intricate mathematical principles and problem-solving abilities, which can provide distinctive challenges for AI models like ChatGPT. (Ogundare et al., 2023) conducted a case study in the field of industrial engineering to evaluate ChatGPT's ability to solve complicated problems in areas like oil and gas engineering, factory automation, and PLC programming. Their discoveries emphasized the limitations of

ChatGPT in automatically discerning robust and feeble solutions to fundamental physics equations that govern various industrial processes. (Pursnani et al., 2023) evaluated the practicability and effectiveness of employing ChatGPT, a model based on GPT-4, for the Fundamentals of Engineering (FE) Environmental Exam. The study revealed that ChatGPT exhibited a noteworthy enhancement in performance through noninvasive prompt modifications, resulting in a remarkable accuracy rate. This indicates the potential for AI to aid in learning and examination processes. In addition, (Elder et al., 2023) examined whether ChatGPT has the ability to successfully fulfill the requirements of a second-year digital design laboratory course. The study found that although ChatGPT lacked the capability to physically wire the circuits, it was able to provide a detailed description of the wiring, generate the necessary programming code, and respond to laboratory queries, achieving a score of 73%.

The study conducted by (Shen et al., 2023) aimed to assess the reliability of ChatGPT on an extensive scale. They utilized a dataset consisting of 5,695 questions from ten different datasets and eight domains. The findings revealed significant variability in ChatGPT's reliability across several areas, with noticeable underperformance observed in the fields of law and science. In their study, (Fergus et al., 2023) thoroughly assessed the academic responses of ChatGPT, examining its accuracy, reliability, and depth in comparison to conventional approaches. The study recognized the capabilities of ChatGPT in generating coherent and contextually relevant responses quickly, while also noting its weaknesses in handling complicated and detailed queries. A combination of qualitative and quantitative assessments was employed to propose improvements for AI in the field of education. Findings from the study indicated that ChatGPT had potential as an educational instrument; however, it required continuous improvement and oversight to achieve its full potential in academic settings.

Chauncey & McKenna (2023) examined the ethical and practical consequences of AI in education, specifically focusing on how ChatGPT can enhance critical thinking and cognitive flexibility in learning environments supplemented with AI. Nevertheless, studies conducted by (Shahriar & Hayawi, 2023 and Borj, 2023) highlighted that although ChatGPT showed exceptional performance in producing responses that resemble human-like behavior, it encountered challenges with handling complex reasoning and understanding emotions. These impediments were further complicated by concerns related to privacy and the spread of misinformation.

## Methodology

In this section we describe in detail the methodology followed in this study to evaluate the capabilities of ChatGPT with respect to MIS exam.

### Test Questions

A total of 200 questions was retrieved from the test bank comprising five chapters of the renowned textbook on Management Information Systems (Laudon & Laudon, 2020). Chapter 1: Information Systems in Global Business Today; Chapter 2: Global E-Business and Collaboration; Chapter 3: Information Systems, Organizations, and Strategy; Chapter 4: Ethical and Social Issues in Information Systems; and Chapter 5: IT Infrastructure and Emerging Technologies. By using random sampling without replacement technique, 40 questions were randomly chosen for each chapter that covered all three different question types. Specifically, each chapter consists of 20 multiple-choice questions (one answer and four distractors), 13 true/false questions, and 7 essay questions. In addition to quality criteria, each individual question was also assigned a level of difficulty, i.e., easy, moderate, and challenging.

### Setup

In this study, each question was queried using ChatGPT version 3.5 and version 4.0 via OpenAI website (OpenAI, 2023). In order to quantitatively assess answers from ChatGPT, we instructed ChatGPT and compared answers with two query formations: original question from test bank with no prompt formation and modified question using prompt adopted from (HuggingFace, 2024; Kocoń et al., 2023; Shen et al., 2023). More specifically, the latter encapsulated prompt with question and additional information, e.g., context and options, for different question types as presented in Table 1.

**Table 1** Modified questions using query prompt

| Question Type   | Prompt Format                                                                                                                                                                                                                      |
|-----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Multiple-Choice | I will provide a question with {option_number} answers to you. The answer is marked as (A), (B), (C), (D), (E). You need to return the answer ID to me, i.e., (A), and write explanations. My question is: {question}<br>{options} |
| True/False      | I will provide a passage and a yes-no question to you. The answer is 'yes' or 'no'. You need to return me your answer, i.e., 'yes', and write                                                                                      |

| Question Type | Prompt Format                                                                                                                                                 |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
|               | explanations. The passage is {context} and the question is {question}.<br>Now, please answer the question.                                                    |
| Essay         | I will provide a context and a question to you. You need to answer me the question based on the context. The context is {context}. The question is {question} |

As a result, there were 800 queries sending to ChatGPT in the aggregate to process its output.

### Score Conversion

For multiple-choice and true/false questions, if ChatGPT is able to answer correctly, score of 1 is given, otherwise it is 0 (zero). However, answers of essay questions are a long text, in this case, ROUGE-L (Lin, 2004) which is one of the ROUGE metrics is utilized to represent an accuracy score. ROUGE-L is the Longest Common Subsequence (LCS) oriented. LCS is the longest sequence of words that appear in both the candidates and reference summaries, while keeping the order of the words intact. The ROUGE-L score lies between 0 and 1, with higher scores (close to 1) indicating higher similarity between the generated summary and the reference. The computation of ROUGE-L score based on an answer to essay question and a result provided by ChatGPT using Python programming language is shown in Algorithm 1.

| Algorithm 1 Computation of ROUGE-L Score |                                                             |
|------------------------------------------|-------------------------------------------------------------|
|                                          | <b>Input:</b> reference-summary $R$ , generated-summary $G$ |
|                                          | <b>Output:</b> ROUGE-L-score $S$                            |
| 1                                        | install ROUGE package                                       |
| 2                                        | initialize the ROUGE object                                 |
| 3                                        | $R \leftarrow$ reference summary or answer from test bank   |
| 4                                        | $G \leftarrow$ generated summary or output from ChatGPT     |
| 5                                        | $S \leftarrow \text{compute\_ROUGE\_scores}(R, G)$          |
| 6                                        | print $S$                                                   |

## Research Results

In this section, the performance of ChatGPT-3.5 and 4.0 were evaluated in generating outputs to address the tasks presented in the previous section.

### Sampling Results

Out of 200 questions from five chapters, they were randomly divided into three levels of difficulty. This resulted in 55 questions (27.5%) classified as Easy, 92 questions (46.0%) classified as Moderate, and 53 questions (26.5%) classified as Challenging as shown in Figure 1(a). The number of questions from the test bank in each difficulty level are varied for each chapter as shown in Figure 1(b). Specifically, most of the questions were classified as Moderate level in chapters 1, 2, 3, and 5 ranging from 18-22 questions. While in chapter 4, the number of Easy and Challenging questions were equal as 14 questions, and 12 questions were of Moderate level. Based on the results shown in Figure 1(b), chapters 3 and 5 appeared to be more difficult than chapters 1, 2, and 4.

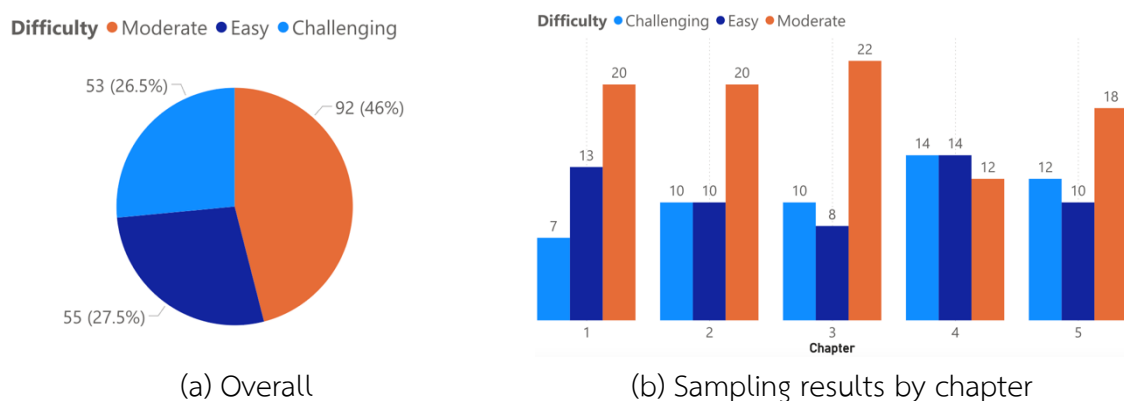


**Figure 1** Results of random sampling without replacement

### Levels of Difficulty

Figure 2 compares the average scores of results between ChatGPT-3.5 and 4.0 based on three levels of difficulty. Each box in the box and whisker plot shows the interquartile range (IQR), which contains the middle 50% of the score for each category. For the result of ChatGPT-3.5 as depicted in Figure 2(a), in the “Challenging” level, the scores were more concentrated around the median (0.19) which indicated a lower average change. For the “Easy” level, the scores were spread out more evenly across the IQR with the lowest median (0.09), suggesting the highest average change in this level of difficulty. The “Moderate” level

had a wider spread of scores when only compared to the challenging level. These characteristics are analogous to the result of ChatGPT-4.0 as shown in Figure 2(b).

The scores in the “Easy” level show a highest positively skewed distribution, with the mean greater than the median. This signifies the spread and variability of higher scores that pull the mean towards a higher value. In other words, both ChatGPT-3.5 and 4.0 performed quite well in answering most of the questions on this level. However, the small IQRs of “Challenging” and “Moderate” levels indicate scores are clustered more tightly in which 50% of the scores from ChatGPT-3.5 are closely around (0.15, 0.21) and (0.11, 0.20) respectively. While the scores from ChatGPT-4.0 result in a similar manner with (0.14, 0.21) and (0.11, 0.21) for “Challenging” and “Moderate” level respectively.

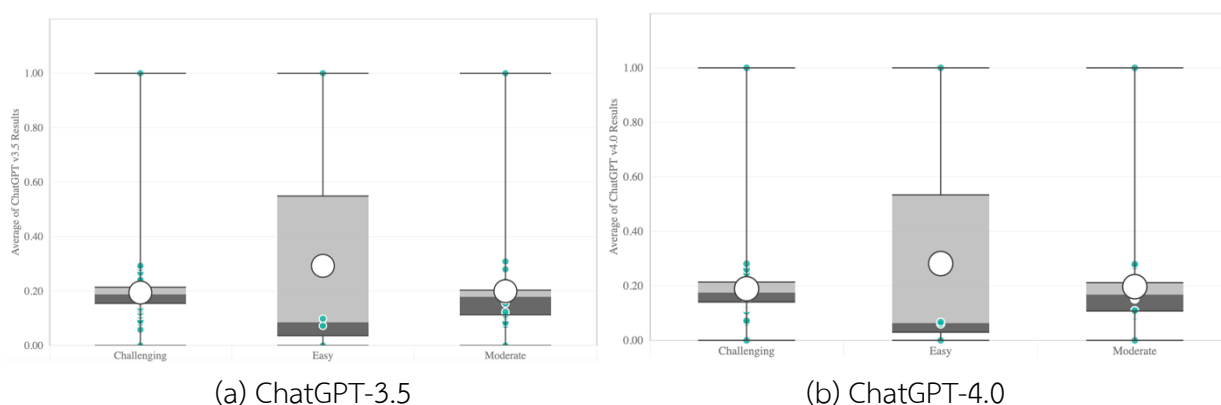


Figure 2 Average of ChatGPT results by difficulty level

### Levels of Difficulty and Chapters

Figure 3 extends the comparison of Figure 2 into chapters. Both ChatGPT versions exhibited the greatest difficulty in answering questions on the “Challenging” level across all five chapters. Especially, Chapter 2 yielded the lowest average score (0.36), followed by Chapter 3 (0.38 and 0.43 for ChatGPT-3.5 and 4.0, respectively). The highest average scores have been found on the “Easy” level where ChatGPT were able to answer correctly and achieved average scores higher than “Moderate” and “Challenging” levels across most chapters with the exception of Chapter 4. For “Moderate” level, questions are viewed as less easy compared to those on the “Easy” level. Consequently, all average scores fell below those on the “Easy” level in most chapters (except Chapter 4) with a particularly notable difference observed in Chapter 5 only for ChatGPT-3.5.

ChatGPT marked the lowest average score of 3.6 for “Challenging” questions in Chapter 2 which focuses on global e-business and collaboration. Conversely, Chapters 1, which is about information systems in global business today, and Chapter 4 that focuses on ethical and social issues in information systems presented the highest average scores of 0.53 and 0.59 for ChatGPT-3.5 and 4.0 respectively. For “Moderate” level, Chapter 4 again achieved the highest average score of 0.88 and 0.96 while Chapter 1 yielded the lowest average score of 0.53 and 0.63 for ChatGPT-3.5 and 4.0 respectively. In case of “Easy” level, Chapter 4 achieved the highest average score of 0.88 and 0.96 while Chapter 1 yielded the lowest average score of 0.53 and 0.63 for ChatGPT-3.5 and 4.0 respectively. Remarkably, ChatGPT-4.0 attained a perfect score on questions in Chapter 3 (Figure 3(b)).

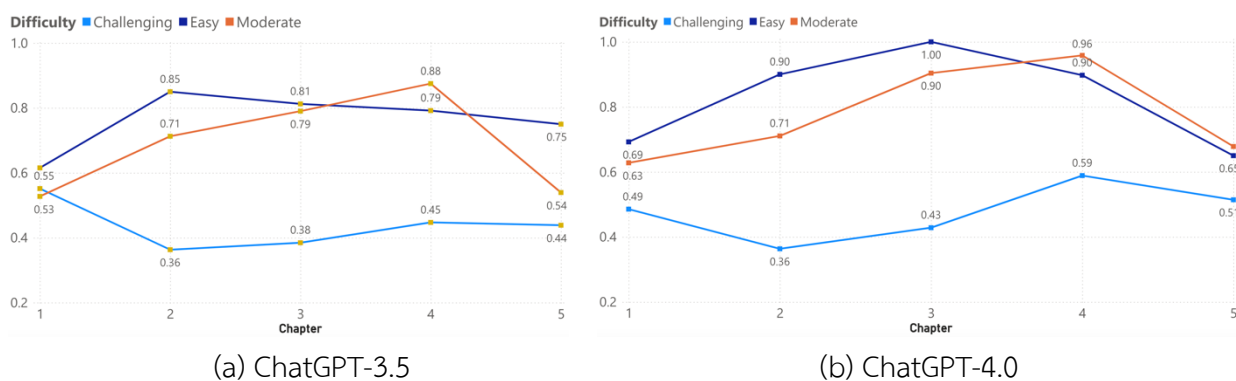


**Figure 3** Average of ChatGPT results by difficulty level and chapter

### Levels of Difficulty and Question Types

As mentioned earlier, the test bank incorporates three distinct question types: multiple-choice (MCQ), true/false (TF), and essay. Each question type encompasses questions across all three difficulty levels. However, the true/false (TF) category deviates slightly, as the test bank offers only “Easy” and “Moderate” difficulty levels for this specific type.

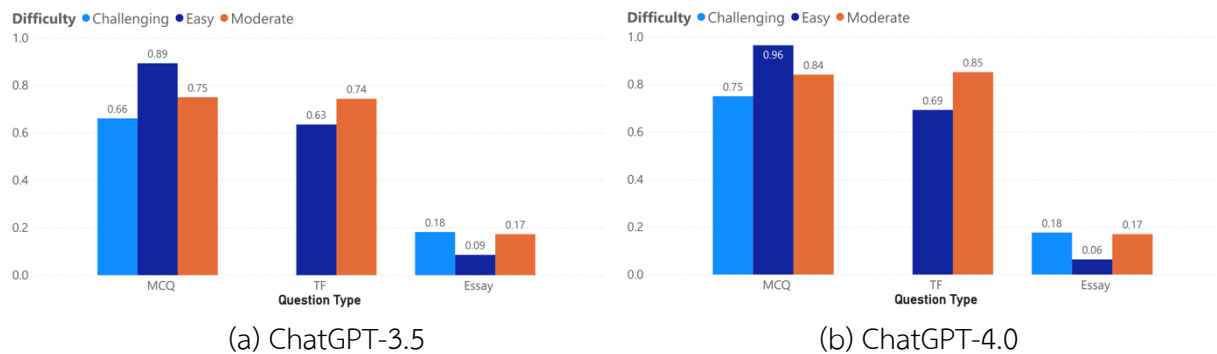


**Figure 4** Average of ChatGPT results by difficulty level and question types

Figure 4 illustrates the performance of ChatGPT-3.5 and 4.0 on various question types and difficulty levels. For multiple-choice questions (MCQ), both versions performed well on “Easy” questions, with ChatGPT-4.0 exhibiting a slight average score advantage across all difficulty levels as shown in Figure 4(b). ChatGPT-4.0 also achieved higher average scores than ChatGPT-3.5 for true/false (TF) questions. However, Essay questions resulted in the lowest average scores for both versions due to the utilization of the ROUGE-L metric, which may not fully capture the nuances of essay generation. Interestingly, ChatGPT-3.5 and 4.0 displayed similar average scores for Essay questions across all difficulty levels.

### Query Prompts

The comparison of average scores between ChatGPT-3.5 and 4.0 on results obtained from query prompts is presented in Figure 5. The results are segregated into responses with and without prompts for both versions. Clearly, modified questions with additional information such as context and options as shown in Table 1 can make ChatGPT to generate answers more accurately. It can be seen that the average scores from ChatGPT-3.5 (Figure 5(a)) are almost consistent to those from ChatGPT-4.0 (Figure 5(b)) regardless of prompt usage. In other words, ChatGPT-4.0 yielded the same meaning, median, 1<sup>st</sup> quartile, 3<sup>rd</sup> quartile, and IQR as 0.19, 0.17, 0.13, 0.21, and 0.08 respectively. While there is a slight difference in median values (from 0.18 to 0.19) for ChatGPT-3.5. The IQR, which measures the spread of the middle 50% of average scores, shows a marginal decrease across both versions when prompts are employed, indicating a tighter clustering of results around the median.

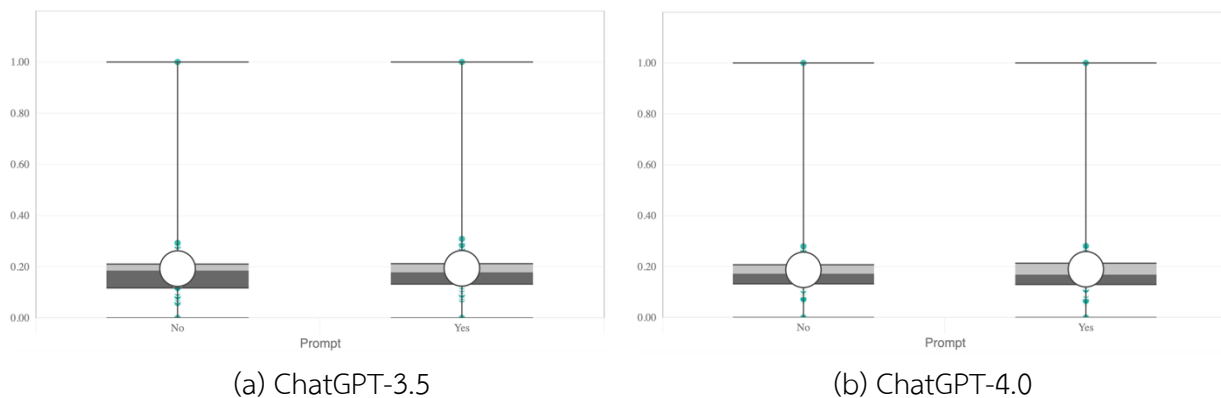


**Figure 5** Average of ChatGPT results by query prompts

For the IQR slightly smaller for prompted responses compared to non-prompted ones indicating a potentially more efficient processing of queries.

### Query Prompts and Chapters

Figure 6 depicts the effect of using a prompt versus its absence across five distinct chapters. For ChatGPT-3.5 as shown in Figure 6(a), the absence of prompt achieved marginally average scores higher compared to the use of prompt in Chapter 1 and Chapter 3. Only Chapter 4 illustrated the impact of prompt usage (with average score of 0.73) was more effective than non-prompted ones (with average score of 0.66). While in Chapter 2 and Chapter 5, the average scores level out to identical values of 0.66 and 0.56 respectively, regardless of prompt usage. For ChatGPT-4.0 as shown in Figure 6(b), the presence of prompts outperformed its absence in terms of average scores in Chapter 1 and Chapter 4. Conversely, Chapter 2, Chapter 3, and Chapter 5 exhibited slightly better average scores when without the prompt at the outset.

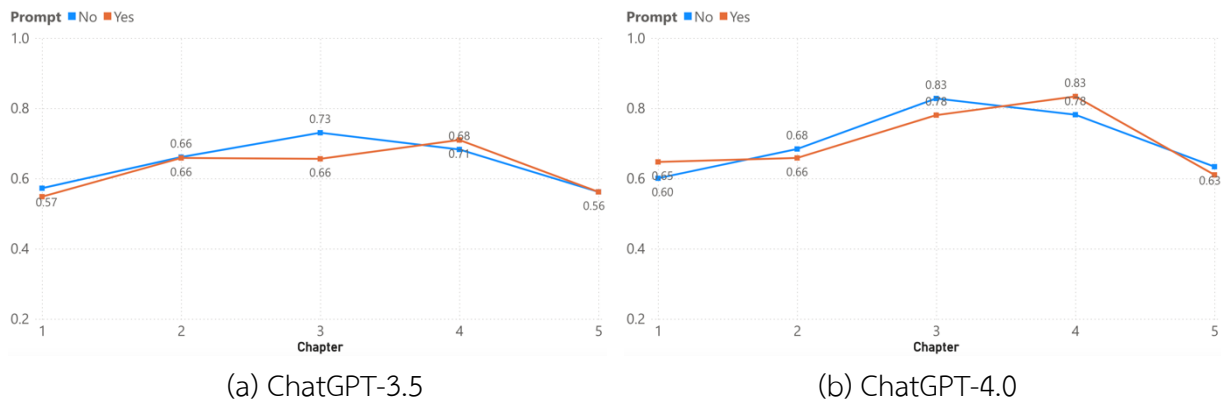


**Figure 6** Average of ChatGPT results by query prompts and chapters

### Query Prompts and Question Types

Figure 7 compares the performance of two different query formats when interacting with ChatGPT across three types of questions. For ChatGPT-3.5 as presented in Figure 7(a), queries with a prompt format have a slightly lower average score at 0.76 compared to those without a prompt format at 0.77 for multiple-choice questions (MCQ). A similar trend is observed for true/false (TF) questions, with prompted queries performing at 0.67 and non-prompted at 0.70. Notably, for essay questions, both prompted and non-prompted queries result in an equal average score of 0.18. Whereas in case of ChatGPT-4.0 as depicted in Figure 7(b), in all question types, the usage of a prompt format did not significantly differ from not using a prompt, as both query formations yielded an identical score of 0.85 for multiple-choice (MCQ), 0.78 for true/false (TF), and 0.17 for essay questions.

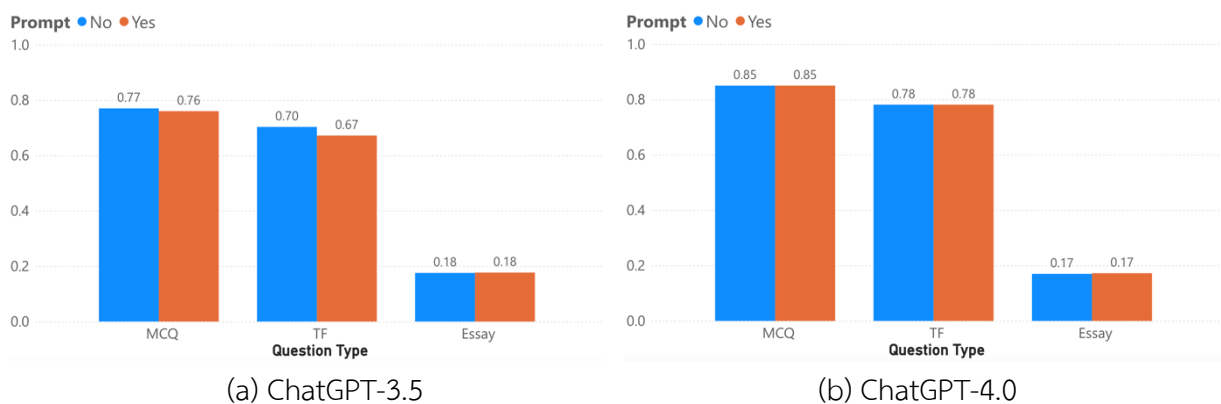


**Figure 7** Average of ChatGPT results by query prompts and question types

## Discussion

This study demonstrates the capabilities of ChatGPT as a very promising tool for enhancing communication and collaboration between humans and machines. It achieves this by offering a conversational interface specifically tailored to assist in addressing queries related to Management Information Systems (MIS). The results presented in this study reveal that the performance of ChatGPT is influenced by two main factors: 1) difficulty levels, and 2) the utilization of query prompts.

Our findings regarding difficulty levels revealed a consistent trend across all ChatGPT versions. Unsurprisingly, ChatGPT excelled at answering easy-level questions. Interestingly, its performance on challenging questions surpassed that on moderate questions. Delving deeper into chapter content, we observed superior performance by ChatGPT, particularly ChatGPT-4.0, when responding to moderate and challenging questions related to information systems and ethics/social issues. For easy questions, ChatGPT prevailed when applied to the contents of global e-business (ChatGPT-3.5) and information systems related to organizations and strategy (ChatGPT-4.0). Notably, ChatGPT-4.0 consistently outperformed ChatGPT-3.5 across all difficulty levels and question types. An interesting pattern emerged for question types: Both versions achieved peak performance on easy-level multiple-choice questions (MCQs) only. Conversely, for true/false (TF) and essay questions, both versions performed better on the moderate level compared to the easy level.

A comparative analysis of the performance metrics for ChatGPT-3.5 and 4.0, distinguished by whether a prompt was used ("Yes" or "No"), revealed that there was no significant impact, as the presence of prompts appeared to slightly improve the efficiency of both versions. Our findings suggested that prompts significantly impacted the content generated by both versions, particularly regarding ethical and social issues in information systems. Conversely, prompts demonstrated a diminished effect on content related to e-business, information systems regarding organizations, strategy, IT infrastructure, and emerging technologies. While the content focused on information systems in global business, it appeared to benefit most from prompts and user queries, particularly when employing ChatGPT-4.0.

Testing ChatGPT-3.5 for multiple-choice (MCQ) and true/false (TF) questions without using a prompt format marginally improved the query performance, while for essay questions, the prompt format made no difference. However, ChatGPT-4.0 exhibited consistent performance

across all question types, irrespective of prompt structure. This could indicate the efficiency of ChatGPT's ability to comprehend and respond to different types of queries in the context of MIS, regardless of its version.

Overall, in this study, we showed that ChatGPT-4.0 achieves a higher level of accuracy in responding to MIS questions, with a success rate of 70.58%, compared to ChatGPT-3.5's success rate of 63.42%. Our findings appeared to be in line with prior studies that have also found high scores on several examinations, for example, medical licensing exams (Gilson et al., 2023), GIS exams (Mooney et al., 2023), national high school exams (de Winter, 2023), and engineering exams (Pursnani et al., 2023). Our analysis revealed that ChatGPT-4.0 demonstrated superior performance in handling higher difficulty levels but performed less effectively at a low difficulty level for a certain context. This phenomenon also confirms results from the study by Hermann-Werner et al. (2024).

## Limitations

This study acknowledges limitations in its scope. Firstly, the study focused on a specific subset of MIS questions. While the field of MIS in the digital age encompasses a broad spectrum, ranging from streamlined business processes to AI-powered decision support (Laudon & Laudon, 2020), this study did not cover that entire range. Secondly, the well-documented issue of information inaccuracy in large language models (LLMs) by Borji (2023) was not addressed here. Analyzing how ChatGPT responds to incorrect questions could provide valuable insights into its limitations.

## References

- AACSB. (2024). Undergraduate. *AACSB International*. Retrieved April 15, 2024, from <https://www.aacsb.edu/learners/journey/undergraduate>
- AL-Qadri, F. M. I., & Ahmed, A. R. (2023). Artificial intelligence in education: A literature review and research agenda. *Research Square*. <https://doi.org/10.21203/rs.3.rs-2610442/v1>
- Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmology Science*, 3(4), 100324. <https://doi.org/10.1016/j.xops.2023.100324>

- Bordt, M., & von Luxburg, U. (2023). Evaluating capabilities of foundation models for algorithmic fairness. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.2309.16115>
- Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv*. <https://doi.org/10.48550/arXiv.2302.03494>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.
- Chauncey, A., & McKenna, L. (2023). ChatGPT and writing pedagogy: Is this the end of writing instruction? *College English*, 86(3), 187–195.
- Cribben, I., & Zeinali, Y. (2023). Artificial intelligence in auditing: The state of the art and future research agenda. *SSRN*. <https://doi.org/10.2139/ssrn.4445017>
- de Winter, J. C. F. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*, 33(3), 675–705. <https://doi.org/10.1007/s40593-023-00372-z>
- Elder, J., Knapp, K., Rose, M., & Belcher, C. (2023, April 1–16). Artificial intelligence in undergraduate education: A review. In *IEEE SoutheastCon 2023* (pp. 1–5). Orlando, FL. <https://doi.org/10.1109/SoutheastCon48659.2023.10114312>
- Eulerich, M., Weitz, S., & Werner, N. (2023). Artificial intelligence in internal auditing: A systematic literature review. *SSRN*. <https://doi.org/10.2139/ssrn.4506303>
- Fergus, T. A., McGrath, P. B., & Malgaroli, M. (2023). An initial evaluation of ChatGPT in generating a treatment plan for social anxiety disorder. *Psychology and Psychotherapy: Theory, Research and Practice*, 96(4), 1125–1133. <https://doi.org/10.1111/papt.12483>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. In *Advances in Neural Information Processing Systems* (NeurIPS 2023).
- Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How well does ChatGPT do when taking the medical licensing exams? The implications

- of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9, e45312. <https://doi.org/10.2196/45312>
- Hajj, M., & Sah, A. (2023, November 23–25). Can ChatGPT pass accounting exams? In *2023 7th International Conference on Information System and Data Mining (ICISDM)* (pp. 25–30). <https://doi.org/10.1145/3594612.3594625>
- Herrmann-Werner, A., Nikendei, C., Greif, R., Roth, S., Hoffmann, G. F., & Zipfel, S. (2024). Artificial intelligence and ChatGPT in medical education: A critical review. *GMS Journal for Medical Education*, 41(1), Doc03. <https://doi.org/10.3205/zma001642>
- HuggingFace. (2024). Awesome ChatGPT prompts. Retrieved February 20, 2024, from <https://huggingface.co/datasets/fka/awesome-chatgpt-prompts>
- Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99, 101861. <https://doi.org/10.1016/j.inffus.2023.101861>
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
- Laudon, K. C., & Laudon, J. P. (2020). *Management information systems: Managing the digital firm* (16th ed.). Pearson Education Ltd.
- Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2), 100790. <https://doi.org/10.1016/j.ijme.2023.100790>
- Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, Barcelona, Spain.
- Mooney, P., Cui, W., Guan, B., & Juhász, L. (2023). Towards understanding the geospatial skills of ChatGPT: Taking a Geographic Information Systems (GIS) exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (AIGKD '23)*, Hamburg, Germany. <https://doi.org/10.1145/3615886.3627745>

- Ogundare, O., Madasu, S., & Wiggins, N. (2023, 1–3 Nov.). Industrial engineering with large language models: A case study of ChatGPT's performance on oil & gas problems. In *2023 11th International Conference on Control, Mechatronics and Automation (ICCMA)*.
- OpenAI. (2023). *ChatGPT* (March 14 version) [Large language model]. Retrieved January 10, 2025, from <https://chat.openai.com/chat>
- Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of ChatGPT on the US Fundamentals of Engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*, 5, 100183.  
<https://doi.org/10.1016/j.caeai.2023.100183>
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit Spewer or The End of Traditional Assessments in Higher Education? *Journal of applied learning and teaching*, 6(1), 342–363. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson.
- Shahriar, S., & Hayawi, K. (2023). Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *Artificial Intelligence and Applications*, 2(1), 11–20. <https://doi.org/10.47852/bonviewAIA3202939>
- Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). In ChatGPT we trust? Measuring and characterizing the reliability of ChatGPT. *arXiv preprint arXiv:2304.08979*.  
<https://arxiv.org/abs/2304.08979>
- Terwiesch, C. (2023). *Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course*. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania.  
<https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf>

Vaughan-Nichols, S., & Diaz, M. (2024, April 20). *ChatGPT vs. ChatGPT Plus: Is it worth the subscription fee?* ZDNet. <https://www.zdnet.com/article/chatgpt-vs-chatgpt-plus-is-it-worth-the-subscription-fee/>