International Journal of Economics, Commerce and Management
United Kingdom ISSN 2348 0386 Vol. 11, Issue 11, Nov 2023

https://ijecm.co.uk/

THE IMPACT OF LARGE LANGUAGE MODEL-ASSISTED LEARNING VERSUS TRADITIONAL LEARNING METHODS ON UNIVERSITY STUDENTS LEARNING OUTCOMES AND KNOWLEDGE RETENTION

Siniša Milošević
Department of management International Burch University, Bosnia and Herzegovina
sinisa.milosevic@stu.ibu.edu.ba

Adisa Omerbegović Arapović, PhD
Professor, Department of management International Burch University, Bosnia and Herzegovina
adisa.omerbegovic@ibu.edu.ba

Malcolm Duerod, PhD
Assist. Professor, Department of management International Burch University, Bosnia and Herzegovina
malcolm.duerod@ibu.edu.ba

Abstract

The purpose of this work is to investigate the impact of large language models, such as GPT-3.5 (ChatGPT) on the academic performance and knowledge retention of university students. The experiment took place at International Burch University and involved first-year students from various non-economic faculties. In light of the increasing importance of artificial intelligence (AI) in the field of education, this study seeks to explore the efficacy of LLMs to transfer knowledge as well as explore their potential impact in education. The participants were divided into two groups using a random selection technique. The experimental group made use of GPT-3.5 (ChatGPT) chat-bot, whereas the control group relied on conventional research strategies. The quizzes aimed to assess the students' understanding and capability to apply the knowledge they acquired during the experiment. Surprisingly, the experimental group performed *significantly better in comparison to the control group in the second quiz where the participants had to rely on the knowledge acquired. These findings suggest that AI-driven learning tools have the potential to complement traditional classroom instruction and promote deeper comprehension and retention of the subject matter. This research adds to the existing body of literature by illustrating the effectiveness of large language model-based instructional approaches in enhancing university students' learning outcomes and knowledge retention.*

Keywords: AI, Education, LLMs, ChatGPT, Learning Outcomes, Knowledge Retention

INTRODUCTION

The rapid advancements in technology have significantly transformed the way we learn and process information. One of the most notable breakthroughs in recent years is the development of large language models, such as GPT-3, GPT-4, BigScience Bloom, OpenAssistant, and GPT-J 6B which employs deep learning techniques to generate human-like text (Brown et al., 2020; OpenAI, R. 2023; ElutherAI, 2020; Teven Le Scao et al. 2023; Köpf et al. 2023). Artificial Intelligence (AI) has made remarkable strides, particularly in the area of Natural Language Processing (NLP), with large language models like OpenAI's GPT-4 leading the charge. These models utilize advanced neural network architectures to process and generate contextually and grammatically coherent text (Vaswani et al., 2017). GPT-3, a transformer-based model, is trained on an extensive amount of text data, enabling it to perform a wide range of NLP tasks, including text generation, translation, summarization, and question answering, among others (Brown et al., 2020).

As interest escalates concerning innovative pedagogical approaches bolstered via AI, research delves deeply into the associated implications. Present research strives to bridge the void prevailing within the existing literature implementing experimentation assessing the efficacies encompassing large language model facilitated learning opposite established educational methodologies. An important challenge facing the use of AI-powered educational tools in schools is making sure these tools help and strengthen the way people learn instead of taking away the jobs of human teachers completely (Chan & Tsi, 2023).

The objectives of this paper are specifically to:

(1) Compare and contrast the effectiveness of large language model-assisted learning and traditional learning methods on university students learning outcomes and knowledge retention,

(2) provide recommendations in regards to the integration of LLMs into university education.

This study is based on an experiment where freshman students were segmented into two groups, a control group and an experiment group. Both groups were asked to complete two quizzes, with the experiment group being instructed to use a chat-bot powered by a large language model notably ChatGPT (GPT 3.5-Turbo) for the first quiz, while the control group was restricted from using such technology and was asked to use traditional methods of research, such as relying on the internet and other sources. Following the quizzes, the students' performance was analysed to determine which group has learned more effectively.

The integration of AI into education has been motivated by various elements, including the increasing availability of large datasets, advances in machine learning algorithms, and the growing demand for personalized learning experiences (Baker & Inventado, 2014). Education can be transformed by leveraging artificial intelligence to offer customized feedback and guidance to learners, identifying gaps in their knowledge, and recommending targeted learning resources (Woolf, 2010). Moreover, AI can assist educators in identifying students who may be struggling and provide them with additional support, ultimately improving learning outcomes for all students (Baker & Yacef, 2009).

This study is important because gives valuable insights towards the potential impact of large language models on learning and education. The findings of this study could inform educational policies and practices, as well as guide subsequent studies in the realm of AI and education. The methodology used in this study follows a qualitative stud*y design, using a before-and-after test experimental arrangement. to compare the learning outcomes and knowledge retention of both groups.

The rationale behind this research is rooted in the transformative potential of AI and LLMs in education. The advent of AI has brought forth an array of problems and possibilities for enhancing learning experiences, personalizing education, and improving learning outcomes. AI-driven learning tools, such as large language models, have shown promise in generating coherent, contextually relevant text, and executing a diverse array of natural language processing (NLP) assignments, which can be leveraged to augment traditional learning methods (Vaswani et al., 2017; Brown et al., 2020).

LITERATURE REVIEW

The transformative potential exploration of AI's influence on reconfiguring education has been a topic of fascination for many years, with initial research centred on the creation of intelligent tutoring systems (ITS) and computer-assisted instruction (CAI) (Sleeman & Brown, 1982). These pioneering systems paved the way for AI's integration into education, offering personalized feedback and support to learners based on their unique learning needs (Graesser et al., 2014; VanLehn, 2011). However, these early systems' capabilities were constrained by the technology available at the time and the lack of large-scale data for training AI models (Baker & Inventado, 2014).

Large language models have potential applications that go beyond education and extend into various commercial sectors. For instance, large language models have been used to create high-quality content for marketing campaigns, draft legal documents, and assist customer support teams in handling user inquiries (Dwivedi et al., 2023; Haleem, Javaid, & Singh, 2022). Furthermore, these models have shown promise in the development of conversational agents, which can be used in industries like healthcare, finance, and retail to provide personalized and contextually relevant information to users (Rana, 2023).

Hoffman, Owen, and Calvert (2021) studied the experiences of parents with their children's parasocial relationships with conversational agents. They found that these relationships can be seen as trusted voices in the lives of children, providing comfort and support. In a similar vein, Ramadan, Farah, and El Essrawi (2021) found that conversational agents, such as Amazon's Alexa, are redefining companionship and interdependence for people with special needs. Lee, Kavya, and Lasser (2021) explored social interactions and relationships with an intelligent virtual agent. The use of chatbots and AI in education has also been the focus of several studies. Hiremath et al. (2018) developed a chatbot for an education system, finding that it improved student engagement and performance. Pfeffer et al. (2021) also explored the potential of large language models, such as OpenAI's ChatGPT, for education, highlighting both opportunities and challenges. Tamkin, Brundage, Clark, and Ganguli (2021) conducted a comprehensive study on the capabilities, limitations, and societal impact of large language models. They found that while these models have the potential to revolutionize various industries, there are also important ethical and social considerations to keep in mind.

Learning presents a complex phenomenon subjected to extensive study within the field of educational psychology. It entails acquiring, assimilating new knowledge, skills, attitudes, and values influenced by learners' cognitive abilities, motivation, prior knowledge, and the learning environment (Schunk, 2012; Driscoll, 2000). Significant theories proposed to explain the learning process include the constructivist and cognitive load theory. The constructivist theory, proposed by Piaget (1970), further developed by Vygotsky & Cole (1978), proposes that learners construct their knowledge and understanding via environment interaction. The theory emphasizes learner's active role in learning, suggesting that the most effective learning occurs when learners actively engage in knowledge construction (Vygotsky, 1978; Piaget, 1970).

Research in education context reaches the consensus that constructivist learning environments providing learners with real-world, contextualized learning experiences boost problem-solving skills and critical thinking abilities. Large language models may support this active learning process by offering personalized feedback and support, allowing them to engage in meaningful dialogues with the AI system and construct their understanding of the subject matter (Basham et al., 2016).

However, these models can potentially present challenges to the learning process. Despite their impressive capabilities, they can sometimes generate grammatically correct yet semantically nonsensical text or provide factually incorrect information (Brown et al., 2020). Misconceptions and misunderstandings may arise, hindering the learning process. Using large language models may also lead to an over-reliance on AI, potentially minimizing the role of active engagement and critical thinking in the learning process.

The cognitive load theory by Sweller (1988) suggests learning effectiveness is contingent upon how instructional methods are designed to minimize extraneous cognitive load and optimize working memory resource use. Effective use of large language models might reduce cognitive load, offering learners personalized, contextually relevant information, thereby freeing cognitive resources for higher-order thinking and learning (Kirschner, Sweller, & Clark, 2006).

Students traditionally engage in methods that mirror the principles of the constructivist theory. For instance, taking notes, answering textbook chapter problems and questions, developing hypothetical essay questions, creating vocabulary flash cards, reviewing key points from professor-provided Power Point slides, and reviewing case studies and application examples to understand theories taught in the course. Studies such as those conducted by Ausubel (1960) and Novak (1998) confirm that these traditional methods often lead to enhanced knowledge understanding and retention.

Moreover, these traditional methods have often been evaluated and compared to practice tests and other simulated exam questions often available in core courses such as physics, calculus, macro, and microeconomics. Findings from studies such as Roediger and Karpicke (2006) and Butler (2010) suggest that students gain from both types of instructional methods, but the scale does tilt in favor of practice tests and simulated exam questions when it comes to enhanced understanding and long-term information retention.

Practice and Memorable Learning - How traditional methods compare

Notably, traditional methods like note-taking, flashcards, can be helpful for information recall in the short term, yet they fall short in terms of promoting deep understanding and long-term information retention (Karpicke, Butler, & Roediger III, 2009). On the other hand, a learning method that perfectly aligns with the constructivist learning theory, like project-based learning, can stimulate meaningful learning and enhance long-term knowledge retention (Blumenfeld et al., 1991).

Moreover, studies reveal advantages in using Large Language Models to improve learning outcomes. For example, Shermis & Burstein (2013) found that such models can assist educators in grading essays -- a repetitive task that's often criticized for its subjective nature. Furthermore, a study by Hew & Cheung (2014) identified AI teaching assistants as effective tools for answering student queries. This capability reduces the response time for learners' queries and provides a customized learning experience (Miao et al., 2021).

The cognitive load theory, proposed by Sweller (1988), posits that the amount of information that a learner can process at any given time constrained at any particular moment by their working memory capacity. According to this theory, learning is most effective when instructional methods are designed to minimize additional cognitive burden and optimize the utilization of working memory assets.

Several studies have examined the ramifications of cognitive load theory on the creation of instructional designs. For instance, a study by Paas, Renkl, & Sweller (2003) found that instructional methods that reduce extraneous cognitive load, such as worked examples and problem-solving practice, can enhance learning outcomes. Similarly, a study by Kalyuga, Ayres, Chandler, & Sweller (2003) found that adaptive instructional methods, which adjust the level of guidance provided to learners based on their prior knowledge, can effectively manage cognitive load and improve learning efficiency.

On the other hand, if used effectively, large language models can potentially reduce cognitive load by providing learners with personalized and contextually relevant information, thereby freeing up cognitive resources for higher-order thinking and learning (Kirschner, Sweller, & Clark, 2006). For instance, a large language model can be employed to produce concise synopses of intricate texts, diminishing the quantity of information that learners need to process and making it easier for them to understand and assimilate the information (Brown et al., 2020).

A variety of research works have delved into the implementation of artificial intelligence-powered learning instruments within educational settings, focusing primarily on customizable learning environments and adaptable learning systems (Troussas et al., 2022; Cui, W., Xue, Z., & Thai, K. P 2018). The overall consensus derived from these investigations has been largely affirmative regarding the favourable influence exerted by AI-augmented learning resources on students' academic achievements, indicating that AI-driven educational implements possess the capability to upgrade the overall learning process (Woolf, 2010). In particular, an extensive meta-analysis undertaken by Kulik & Fletcher (2016) revealed that participants who have utilized intelligent tutoring systems obtained higher grades in standardized assessments when juxtaposed against individuals who resorted to conventional instructional methodologies. Various analyses have ventured to explore the prospect of AI-motivated learning appliances to cultivate discerning logical analysis and involvement amidst the learning procedure. By way of instance, AI-driven tutoring platforms have been observed to bolster knowledge acquisition by means of feedback that incites learner's critical thinking (Roll et al., 2012). Correspondingly, conversation-oriented tutoring systems have displayed the capability to invigorate critical thinking by getting involved in Socratic dialogue, propounding research questions, and contesting learners' suppositions (Graesser et al., 2014). Large language models, due to their adeptness at generating coherent and situational pertinent text, hold the latent capacity to underpin analogous sorts of interacting pedagogical encounters. Another important consideration in the incorporation of AI-driven learning tools into education is the potential impact on students' motivation and engagement. The present line of inquiry draws its roots from the conceptual structure of constructivism, which advocates for the proposition that learners assume an energetic function in erecting their personal awareness and discernment via communication with their surrounding milieu (Piaget, 1970; Vygotsky & Cole, 1978). This hypothesis accentuates the significance of the learner assuming the character of an immersed participator within the didactic sequence, standing in stark contradistinction to traditional pedagogical paradigms that perceive learners as passive beneficiaries of wisdom. Within a constructivist educational setting, learners involve themselves in discovery, problem solving, and also intellectual perception, fabricating their comprehension through these active procedures. AI-driven learning tools, such as LLMs, have the potential to facilitate this process by affording students personalized feedback and support, allowing them to engage in meaningful dialogue with the AI system and forge their individual grasp of the subject matter (Basham et al 2016). AI-driven learning tools, such as LLMs, have the potential to facilitate this process through furnishing students with customized feedback and support. These models, with their ability to understand and output human-like text, can simulate the role of a tutor, providing explanations, answering queries, and even posing challenging questions. This can enable students to engage in meaningful dialogue with the AI system, thereby actively participating in their learning process. This interactive learning environment can stimulate facilitating the development of critical thinking and problem-solving abilities, thereby enabling students to independently construct their comprehension of the subject matter (Basham et al 2016). The literature review reveals a promising, yet complex, landscape for the integration of LLMs into education. The potential of these AI-driven tools to revolutionize learning outcomes and knowledge retention is evident, mirroring the transformative impact of the internet in its early stages. However, the effectiveness of these AI-driven tools, particularly large language models, is influenced by various factors, including the learning theories and processes they are designed to support. For instance, the constructivist learning theory, as proposed by Piaget (1970) and further developed by Vygotsky & Cole (1978), suggests that learners construct their understanding and knowledge through interaction with their environment. Large language models can potentially support the active learning process by providing personalized feedback and support, allowing learners to engage in meaningful dialogues with the AI system and construct their understanding of the subject matter (Basham et al., 2016), the learning process can also be influenced by the cognitive load theory proposed by Sweller (1988). This theory suggests that the effectiveness of learning is contingent upon how instructional methods are designed to minimize extraneous cognitive load and optimize the use of working memory resources. Therefore, the effective use of large language models might reduce cognitive load by offering learners personalized, contextually relevant information, thereby freeing cognitive resources for higher-order thinking and learning (Kirschner, Sweller, & Clark, 2006). Studies such as those conducted by Ausubel (1960) and Novak (1998) confirm that these traditional methods often lead to enhanced understanding and retention of knowledge. However, these traditional methods can be compared to practice tests and other simulated exam questions, often available in core courses such as physics, calculus, macro, and microeconomics. Findings from studies such as Roediger and Karpicke (2006) and Butler (2010) suggest that students gain from both types of instructional methods, but the scale does tilt in favour of practice tests and simulated exam questions when it comes to enhanced understanding and long-term information retention.

RESEARCH QUESTIONS AND HYPOTHESES

The research questions and hypotheses for this study are designed to delve into the impact of Large Language Model-assisted learning, specifically using ChatGPT, on university students' learning outcomes and knowledge retention. These inquiries are grounded in the existing literature on AI-assisted learning (Chen et al., 2020; Maghsudi et al., 2021) The study aims to answer questions by conducting an experiment involving two groups of students - one group using ChatGPT for learning and the other using conventional research methods. The students' academic performance was assessed through quizzes, and their knowledge retention was evaluated by comparing their performance on a second quiz taken without the use of any technological assistance. The difference in performance between the two groups was used to determine the impact of Large Language Models on learning outcomes and knowledge retention.

Hypothesis 1 (H1): The post-test outcomes of the experimental group (ChatGPT) and the control group (traditional learning methods) are significantly different

Null Hypothesis 1 (H0): There is no significant difference between the post-test results of the treatment group (ChatGPT) and the control group (traditional learning methods).

In cases where the data follows a normal distribution, parametric tests such as Independent Samples t-test or Paired Samples t-test are utilized, the Independent Samples t-test is used to compare the means of two independent groups to determine if there is a statistically significant difference between them (Field, 2013). This test has been used extensively in educational research, including studies investigating the impact of AI-assisted learning on students' learning outcomes (Maghsudi et al., 2021). The Paired Samples t-test, on the other hand, is used to compare the means of the same group at two different times (pre-test and post-test). This test is particularly useful in studies that employ a pre-test-post-test control group design, such as the present study (Chen et al., 2020).

The null hypothesis posits that there is no significant difference between the post-test results of the treatment group (ChatGPT) and the control group (traditional learning methods). This hypothesis will be tested using the same Mann-Whitney U Test.

Research Question 1: What is the effect of using a large language model such as ChatGPT on university students' educational outcomes compared to traditional learning methods?

Research Question 2: How does the use of ChatGPT impact university students' knowledge retention compared to traditional learning methods?

The first research question is designed to compare the learning outcomes of students who employ ChatGPT-assisted learning methods with those who utilize traditional learning methods. This question is motivated by the growing body of research suggesting that AI-assisted learning can enhance students' learning outcomes (Chen et al., 2020; Maghsudi et al., 2021). However, the specific impact of Large Language Model-assisted learning, such as ChatGPT, on learning outcomes remains less explored.

The other research question delves on the impact of ChatGPT on students' knowledge retention. While previous research has suggested that AI-assisted learning can enhance knowledge retention (Chen et al., 2020; Maghsudi et al., 2021), the impact of Large Language Model-assisted learning, such as ChatGPT, on knowledge retention has not been extensively investigated.

METHODOLOGY

The research structure employed in this study followed a pre-test-post-test control group design, a highly effective experimental design that permits a direct evaluation of the impacts of the experimental and control conditions on the dependent variables (Campbell & Stanley, 1963). This design is particularly advantageous as it allows for the control of pre-existing differences between groups, thereby ensuring that any observed differences in post-test scores can be ascribed to the experimental intervention (Shadish, Cook, & Campbell, 2002). The pre-test-post-test control group design has been widely used in educational research to investigate the impact of different instructional methods on student learning outcomes. For example, Chen, L., Chen, P., & Lin, Z. (2020) used this design to investigate the effects of AI-assisted learning on students' academic performance. Similarly, Maghsudi, S., Lan, A., Xu, J., & van Der Schaar, M. (2021) employed a pre-test-post-test control group design to explore the effects of AI-based personalized learning environments on students' learning outcomes. These studies provide valuable benchmarks for regarding the methodology employed in this study and underscore the appropriateness of the research design.

The pre-test was an assignment requiring students to calculate the inflation rate for a specific year and the total inflation from one year to another using Consumer Price Index (CPI) data. This assignment was designed to evaluate the students' comprehension of basic economic principles and their capability to perform calculations related to inflation, while being aided with technology or research material. The reason why only students from non-economic faculties were eligible to participate in this study was to account for the students own knowledge of this subject matter. Thus the pre-test served as the first ever time the students were required to calculate the inflation rate, allowing them to use assistance on the first quiz served them as a learning experience, although with different tools involved, with the control group relying on traditional methods, and the experiment group relying on LLMs

The post-test was similar to the pre-test, where it was an assignment requiring students to calculate the inflation rate for a specific year and the total inflation from one year to another using Consumer Price Index (CPI) but prohibited the use of any tools or assistance. This assignment was designed to assess the students' ability to perform the same calculations without any assistance, thereby providing a measure of their knowledge retention of the knowledge they acquired during the pre-test. This approach to assessment is consistent with the recommendations of educational researchers who argue that the ability to apply knowledge without assistance is a key indicator of learning and understanding (Hattie & Donoghue, 2016).

In both the pre-test and post-test, the students' answers were scored based on the accuracy of their calculations, as the grading criteria table on page 39 further illustrates.

This scoring method is consistent with the approach used in previous research on AI-assisted learning (Maghsudi et al., 2021) and provides a robust metric of the students' educational achievements and the retention of knowledge.

Participants

The individuals involved in this study were freshman university students from International Burch University, a higher education institution renowned for its commitment to innovation and research. The participants were enrolled in non-economical faculties, ensuring a diverse range of academic backgrounds and interests. To ensure a diverse representation, we utilized the university student information system to post general information about the research project, inviting volunteers for an experiment. Individuals who expressed their willingness to participate were then randomly assigned to either the experimental group or the control group. This diversity is important as it allows for the generalizability of the study's findings across different academic disciplines (Creswell & Plano Clark, 2017).

The participants were assigned randomly to either the experimental group or the control group. To achieve random assignment, a segment of this group was selected in a systematic manner. Specifically, individuals were grouped based on their physical location in front of the classrooms. Those standing on the left side were designated as the control group, while those on the right side were assigned to the experimental group. This division was made to ensure a random distribution of participants across the two groups.

It's important to note that the physical placement of individuals was not related to any characteristics or attributes relevant to the study. The assignment was further refined to achieve an equal distribution of participants, with 26 individuals in the control group and 23 in the experimental group. Further refinement was done by moving over individuals who have not yet identified their physical placement. This random assignment process is consistent with established procedures in experimental research

Random assignment represents a vital element of experimental research as it helps to control for confounding variables and ensures that any differences in the students' educational achievements and retention of knowledge can be attributed to the experimental manipulation, rather than pre-existing disparities among the groups (Shadish et al., 2002). This random assignment is consistent with the procedures used in previous research on AI-assisted learning (e.g., Chen et al., 2020; Maghsudi et al., 2021).

The gender distribution of the participants was also considered in this study. The data provided indicates a balanced representation of both genders, which is important in educational research as gender can influence learning outcomes and knowledge retention (Halpern et al. 2011). Past studies have demonstrated that gender can impact the efficacy of different instructional methods, with some studies suggesting that males and females may respond differently to AI-assisted learning (Wang, Jiao, Young, Brooks, & Olson, 2007).

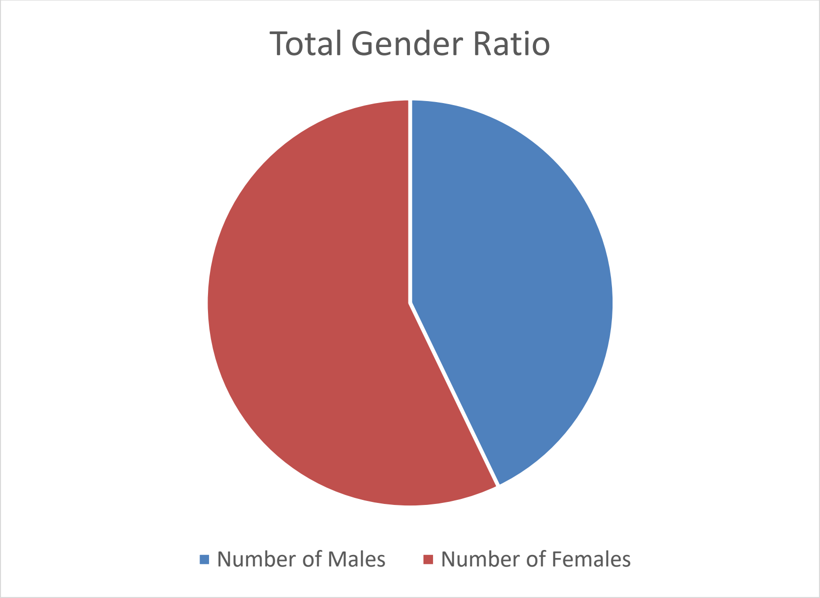


Figure 1 Total Gender ratio in the experiment: both groups

The experimental group consisted of 16 females and 7 males, while the control group consisted of 12 females and 14 males. The slight gender imbalance in the groups is not expected to significantly impact the results of the study, as the main focus of the study is on the comparison of learning outcomes between the experimental and control cohorts, as opposed to between genders. However, the gender distribution of the groups will be considered in the analysis of the data.

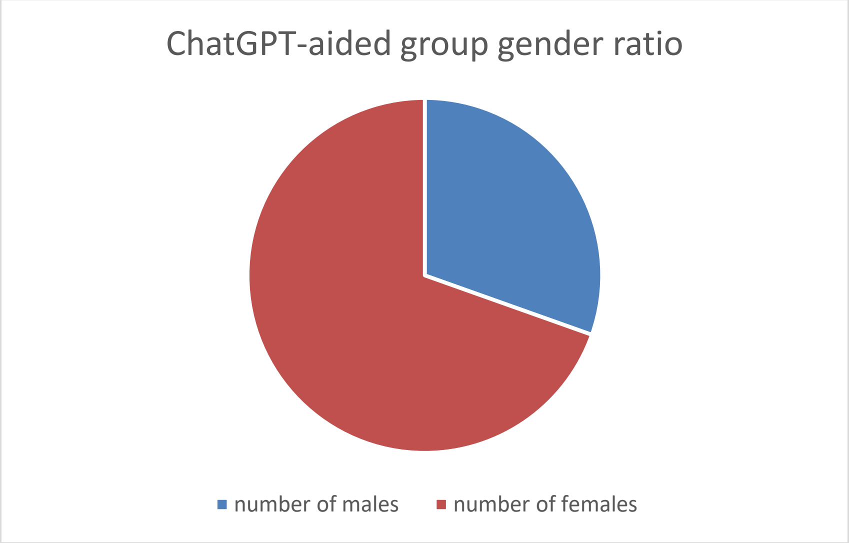


Figure 2 ChatGPT-aided group gender ratio

The participants' grades on two tests were used as the primary data for this study. The grades ranged on a scale from 0 to 10, where higher scores denote superior performance. The first test was administered before the implementation of the experimental treatment (pre-test), while the second test was administered after the treatment (post-test). The use of pre-test and post-test scores is a common practice in educational research as it allows for the comparison of students' performance before and after the implementation of an instructional intervention.

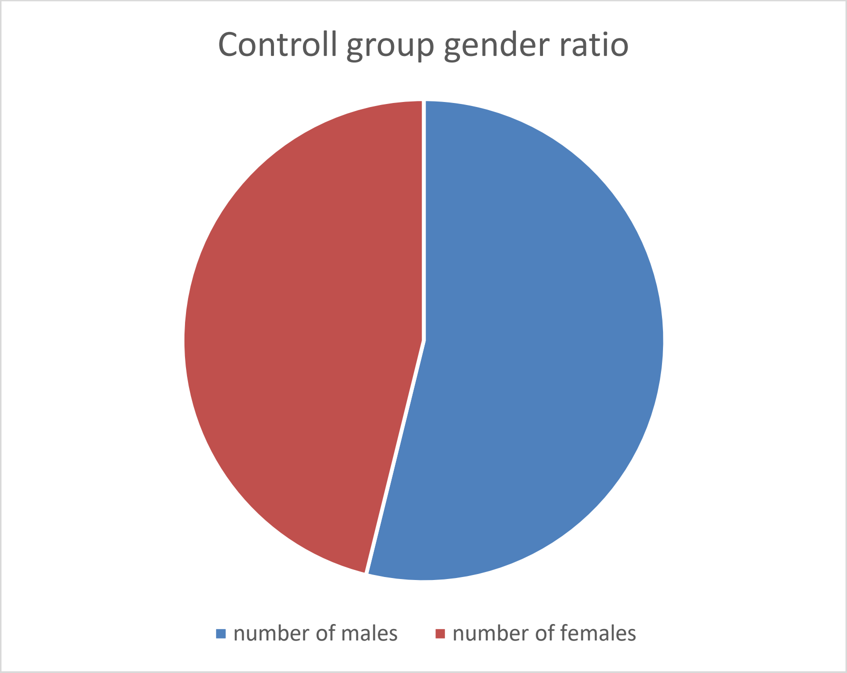
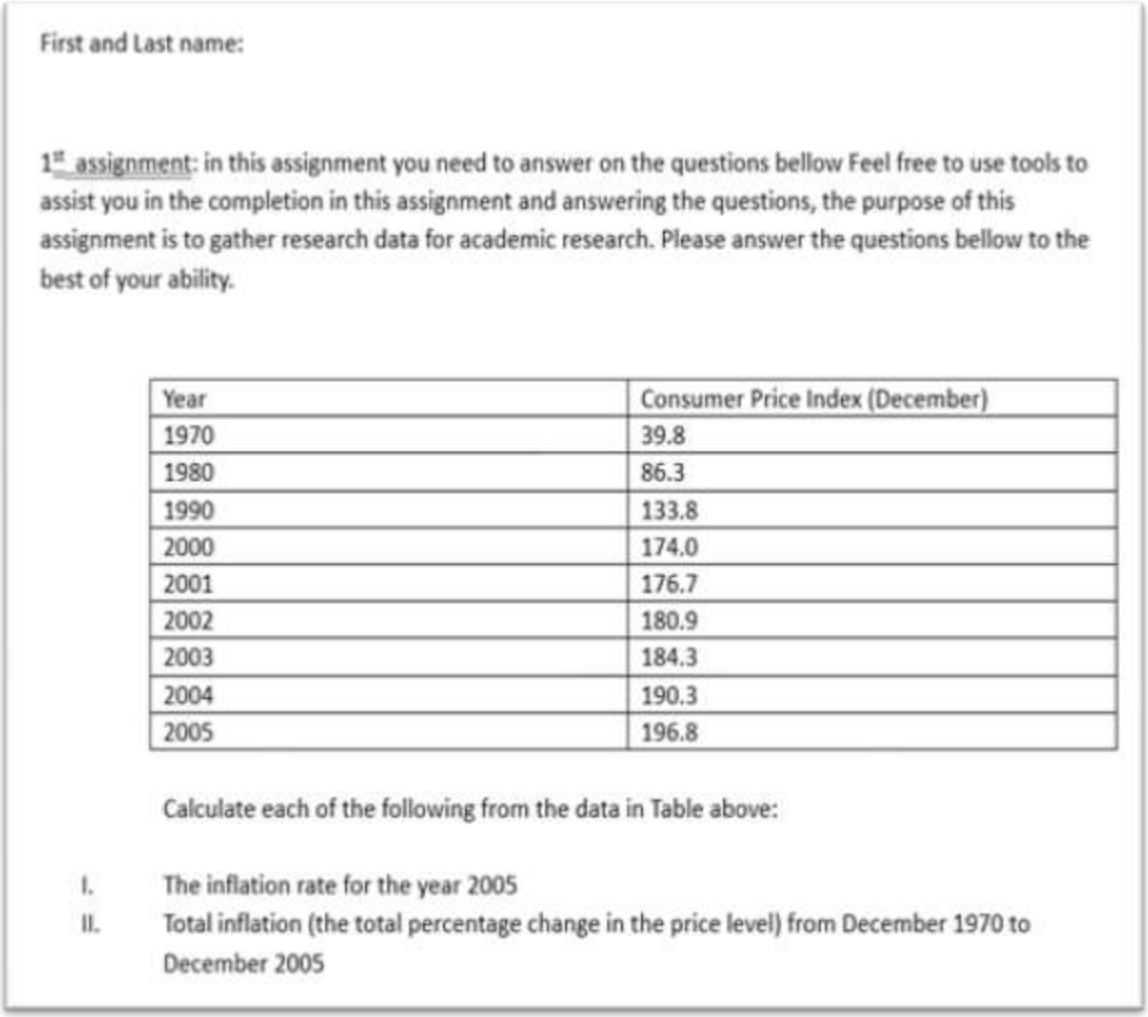


Figure 3 Control group gender ratio

Assignments

The first assignment required students to calculate the inflation rate for a specific year and the total inflation from one year to another using hypothetical Consumer Price Index (CPI) data. This assignment was designed to evaluate the students' comprehension of economic principles and their capability to perform calculations related to inflation. The use of hypothetical data in this assignment, while not real-world, still provides a valuable context for students to apply their understanding, the data within the assignment was modelled for ease of calculations, aligning with the principles of authentic assessment and is consistent with the approach used in previous research on AI-assisted learning (Chen et al., 2020; Maghsudi et al., 2021). The assignment for the first test is as follows:



First and Last name:

1st assignment: in this assignment you need to answer on the questions bellow Feel free to use tools to assist you in the completion in this assignment and answering the questions, the purpose of this assignment is to gather research data for academic research. Please answer the questions bellow to the best of your ability.

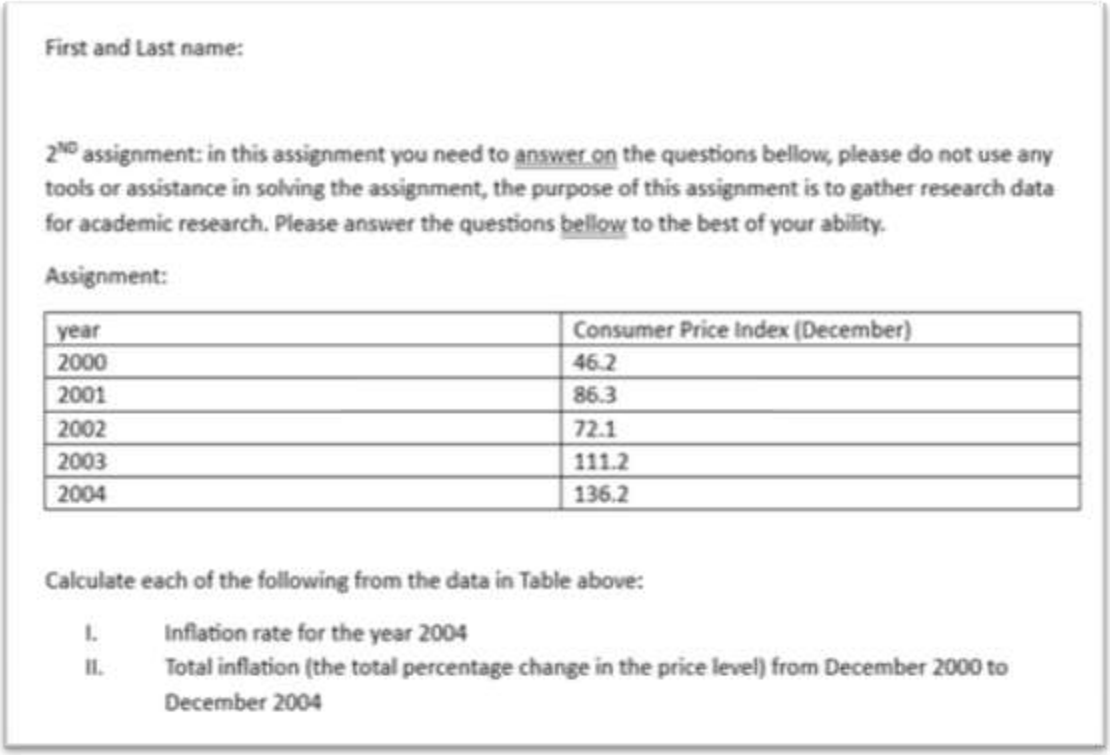
Year	Consumer Price Index (December)
1970	39.8
1980	86.3
1990	133.8
2000	174.0
2001	176.7
2002	180.9
2003	184.3
2004	190.3
2005	196.8

Calculate each of the following from the data in Table above:

I. The inflation rate for the year 2005
II. Total inflation (the total percentage change in the price level) from December 1970 to December 2005

Figure 4 Screenshot of the first quiz

The second assignment was similar to the first assignment but prohibited the use of any tools or assistance. This assignment was designed to assess the students' ability to perform the same calculations without any assistance, thereby providing a measure of their knowledge retention. This approach to assessment is consistent with the recommendations of educational researchers who argue that the ability to apply knowledge without assistance is a key indicator of deep learning and understanding (Hattie & Donoghue, 2016). The assignment used in the second test is as follows:



First and Last name:

2ND assignment: in this assignment you need to answer on the questions bellow, please do not use any tools or assistance in solving the assignment, the purpose of this assignment is to gather research data for academic research. Please answer the questions bellow to the best of your ability.

Assignment:

year	Consumer Price Index (December)
2000	46.2
2001	86.3
2002	72.1
2003	111.2
2004	136.2

Calculate each of the following from the data in Table above:

I. Inflation rate for the year 2004
II. Total inflation (the total percentage change in the price level) from December 2000 to December 2004

Figure 5 Screenshot of the first quiz

In both assignments, the students' answers were scored based on the accuracy of their calculations and their understanding of the economic concepts involved. This scoring method is consistent with the approach used in previous research on AI-assisted learning (Maghsudi et al., 2021) and provides a robust metric of the students' educational achievements and the retention of knowledge.

ANALYSIS AND FINDINGS

The study employed a controlled experimental design, which replicates similar experimental designs used in previous research (Campbell & Stanley, 1963; Maghsudi et al., 2021). The experimental design allowed for a robust comparison of the learning outcomes and knowledge retention between students who used Large Language Model-assisted learning and those who employed traditional learning methods (Chen et al., 2020). When comparing means between two groups, whether they are control or experimental, statistical analyses help determine if a difference exists between them. This process follows well established procedures in econometrics and statistics (Stock & Watson, 2015). This current study involved two quizzes administered to test participant comprehension of course material. During Quiz 1, the experimental group worked directly with ChatGPT, while the control group had access only to general internet resources (Google searches, etc.). For Quiz 2, neither group could seek outside support except for basic calculator functions. We recorded details about participating individuals and their respective scores from these assessments. Using IBM SPSS Statistics, we analyzed the gathered information. Our findings indicated several key observations once we processed the raw data through the software program.

The descriptive statistics calculated in this study include measures of central tendency, encapsulated by the mean, measures of variability typified by the standard deviation, and measures delineating the distribution patterns, collectively contribute to a comprehensive characterization of the data (minimum and maximum scores).

Table 1 Descriptive statistics of the experiment data

Tests:	Number of Grades	Mean	Standard Error of the mean	Standard Deviation
Test 1 Grade	49	5.45	0.419	2.930
Test 2 Grade	49	3.49	0.460	3.222

The mean score for the first quiz was 5.45, indicating that on average, students scored slightly above the midpoint of the grading scale. This implies that the students possessed a moderate level of comprehension regarding the subject matter assessed in the first quiz. The mean, as a measure of central tendency, serves as a representation of the data's central point, encapsulating the arithmetic average of the values therein that provides an average score, giving an overall picture of the students' performance. However, it does not provide information about individual student performance or the range of scores.

The standard deviation for the first quiz was 2.93, which is relatively high, indicating a significant variation in the students' performance. This suggests that while the average score was moderately high, the students' scores varied widely, with some students scoring much higher or lower than the mean. The standard deviation, functioning as a measure of variability, quantifies the extent of dispersion or spread among the data points, thereby offering insights into the data's heterogeneity, provides an indication of the spread of the scores around the mean. A high standard deviation suggests a wide range of scores, In contrast, a low standard deviation signifies that the scores exhibit a tight clustering around the mean, implying a diminished degree of variability or spread within the dataset.

For the second quiz, the mean score was 3.49, which is lower than the mean score for the first quiz. This suggests that the students found the second quiz more challenging than the first one. The lower mean score indicates that on average, students scored below the midpoint of the grading scale, suggesting a lower level of understanding of the topic evaluated in the second quiz.

The standard deviation for the second quiz was 3.22, which is higher than the standard deviation for the first quiz. This indicates a wider range of scores for the second quiz, suggesting that the students' performance varied significantly. The higher standard deviation for the second quiz suggests that the students found the quiz more challenging, resulting in a wider spread of scores.

The minimum and maximum scores provide information about the range of the students' performance. The scoring was based on a scale of 0 to 10, with 0 being the student has not completed any assignments nor has even attempted to do the assignment. For both quizzes, the minimum score was 0 and the maximum score was 10. With both Tests containing two assignments, where each correct assignment yielded 5 points, with the total score being calculated by simply adding the scores achieved on the two assignments into a total grade, this grade being a score ranging from 0 to 10. The table below illustrates the grading criteria for the assignments:

Table 2 Grading criteria for the assignments

Assignment Grading Criteria	
5	The student completed the assignment given with correct result and showed their work
4	Student Completed the assignment given with either a correct result but didn't show their work or showed the work but the result was off by 2 decimal points
3	Student completed the assignment but was off by 2 decimal points and showed their work containing the correct formulas
2	Student completed the assignment with an incorrect result but showed their work
1	Student completed the assignment with an incorrect result and didn't show their work
0	The student submitted a blank paper

The wide range of scores suggests a diverse group of students with varying levels of understanding and ability to apply the knowledge acquired. This information is useful for understanding the overall performance of the students and identifying areas where additional instruction or support may be needed.

Frequency Distribution

The frequency distribution of the scores provides an overview of how the students' scores are distributed across the grading scale. It provides a graphical depiction of the data, allowing for a quick assessment of the students' performance and the identification of patterns and trends.

Table 3 Frequency distribution table for the first test

Test 1 Grade Frequency distribution		
Grade	Frequency	Percent
0	5	10.2%
1	1	2%
2	3	6.1%
3	4	8.2%
4	5	10.2%
5	4	8.2%
6	5	10.2%
7	6	12.2%
8	10	20.4%
9	4	8.2%
10	2	4.1%
Total	49	100%

For the first quiz, the frequency distribution showed that the most common score was 8, obtained by 20.4% of the student population. This signifies that a substantial portion of the students performed well on the first quiz, scoring close to the maximum possible score. The high frequency of scores around 8 suggests that the majority of the students demonstrated a commendable comprehension of the subject matter assessed in the first quiz and were able to apply the knowledge effectively, as shown above in Grade distribution, which also can illustrate the normality or rather absence of the normal distribution in the data.

Test 2 Grade Frequency distribution		
Grade	Frequency	Percent
0	17	34.7%
1	2	4.1%
2	4	8.2%
3	1	2%
4	1	2%
5	7	14.3%
6	12	24.5%
7	0	0%
8	1	2%
9	1	2%
10	3	6.1%
Total	49	100%

Table 4 Frequency distribution for the second test

In contrast, the frequency distribution for the second quiz showed a different pattern. The most common score was 0, obtained by 34.7% of the student body. This implies that a notable segment of the students struggled with the second quiz, or decided to quit, and submitted a blank paper, failing to score any points. The high frequency of scores around 0 indicates that many students either submitted a blank paper for the second quiz or didn’t understand retain any knowledge. The frequency distribution also provides information pertaining to the dispersion of the scores. For the first quiz, the scores were relatively evenly distributed across the grading scale, with a slight skew towards the higher scores. This suggests that the students' performance varied widely, with some students scoring much higher or lower than the mean. The even distribution of scores indicates a diverse group of students with varying levels of comprehension and the proficiency to apply acquired knowledge.

For the second quiz, the scores tended to exhibit a skew toward the lower range of the grading scale. This signifies that a substantial portion of the students struggled with the second quiz, resulting in a high frequency of low scores. The skew towards the lower scores suggests that the second quiz was more challenging than the first one, or that the subject matter assessed in the second quiz was more difficult for the students to understand or apply, moreover the lack of outside resources or help along with the fact that the participants have never before done assignments where they were asked to calculate the inflation rate by their own possibly contributed to the low scores in the second quiz. This information can be useful for identifying areas where additional instruction, support and research may be needed

Normality tests

The normality tests, specifically the Kolmogorov-Smirnov and Shapiro-Wilk tests, are used to assess the distribution of the students' scores. These tests provide information about whether the scores are normally distributed, which is an important assumption for many statistical analyses. A normal distribution is distinguished by a symmetrical, bell-shaped curve, with the majority of scores clustered around the mean and fewer scores at the extremes.

Table 5 Kolmogorov-Smirnov test table

Kolmogorov-Smirnov			
Tests:	Statistic	Degrees of freedom	significance
Test 1 grade	.151	49	0.007
Test 2 grade	.208	49	<.001

Table 6 Shapiro-Wilk test table

Shapiro-Wilk			
Tests:	Statistic	Degrees of freedom	significance
Test 1 grade	.930	49	0.006
Test 2 grade	.854	49	<.001

For the first quiz, the results of the Kolmogorov-Smirnov and Shapiro-Wilk tests ($p < .05$) indicated that scores were not within the levels of a normal distribution of data. This suggests that the distribution of the students' scores deviated from the bell-shaped curve of a normal distribution. The deviation from normality could be due to a variety of factors, such as the presence of outliers, skewness, or kurtosis in the data.

The deviation from normality in the first quiz scores suggests that the students' performance varied widely, with some students scoring much higher or lower than the mean. This could indicate a diverse group of students with varying levels of understanding and ability to apply the knowledge. It could also suggest that the quiz was more challenging for some students than others, resulting in a wider spread of scores.

For the second quiz, the outcomes of the Kolmogorov-Smirnov and Shapiro-Wilk tests ($p < .05$) also signified that the scores were not normally distributed. This suggests that the distribution of the scores for the second quiz also deviated from the bell-shaped curve of a normal distribution. The deviation from normality in the second quiz scores could be due to similar factors as the first quiz, such as the presence of outliers, skewness, or kurtosis within the dataset.

The deviation from normality in the second quiz scores suggests that the students found the quiz more challenging, resulting in a wider spread of scores. This could indicate a lower level of understanding of the subject matter assessed in the second quiz, or difficulty in applying the knowledge effectively. The results of the normality tests provide valuable information for understanding the students' performance and identifying areas where additional instruction or support may be needed.

According to Stock and Watson (2015), to analyse the difference in means between two groups, it is necessary to ascertain whether the data conforms to a normal distribution. If it does, then a t-test for independent samples could be applied. However, In instances where the data deviates from a normal distribution, the adoption of a non-parametric test like the Mann-Whitney U test is deemed more suitable.

We initiated our analysis by employing the independent samples t-test, a statistical technique commonly employed for comparing the means of two autonomous groups, is of particular interest. The test, however, relies on some key assumptions, one of which is the normal distribution of the data, an aspect that's paramount to offering an unbiased, legitimate estimation of the population mean.

In executing this analysis, both Shapiro-Wilk and Kolmogorov-Smirnov tests were conducted to assess normal distribution of data. For both Test_1_grade and Test_2_grade, both tests generated p-values less than .001, resulting in the null hypothesis being rejected, which holds that the data distribution aligns with a normal distribution. Given this factor, the independent samples t-test may not provide accurate results due to this violation of the assumption of normality, which could lead to a skewed representation of the data (Ghasemi & Zahediasl, 2012). Despite these concerns, the independent samples t-test was applied on the data, showing significant disparities between two groups for both Test_1_grade and Test_2_grade, with t(47) values of -2.783 and -2.940 and p-values of .008 and .005 respectively. Keep in mind, however, that the violation of the normality assumption in the dataset makes these results potentially unreliable and might lead to false conclusions about the study's hypotheses.

Table 7 Independent samples T test for the sample

Type	Statistics	Test 1 Grade	Test 2 Grade
		Equal variances	Equal variances
Levente's test for equality variances	F- statistic	0.788	0.480
	Significance	0.379	0.492
	t-score	-2.783	-2.940
	degrees of freedom	47	47

Table 7…

T-Test for equality of means	t-score		-2.783	-2.940
	one-tailed p-value		0.004	0.003
	two-tailed p-value		0.008	0.005
	Mean difference		-2.186	-2.518
	standard error difference		0.785	0.857
	95% confidence interval	Lower	-3.766	-4.242
		Upper	-0.606	-0.795

To address this issue, a viable alternative to the t-test is the nonparametric Mann-Whitney U test, which circumvents the need for normality by comparing medians instead of means. The Mann-Whitney U test carried out on the test scores indicated significant differences between the two groups for both test grades. Test_1_grade resulted in a U-value of 148, a p-value of .002, whereas Test_2_grade resulted in a U-value of 165, a p-value of .006.

Chiefly, these results suggest that even without a normal distribution, conclusions about the group differences can still be drawn. In other words, a great disparity exist between the two groups in both Test_1_grade and Test_2_grade. Notably, these findings are closely aligned with results from the independent samples t-test, underlining the usefulness of the Mann-Whitney U test as a legitimate secondary alternative to the t-test when the assumption of normality of the data is violated.

The choice to employ the Mann-Whitney U test in this study was contingent upon the data's inherent characteristics and the objectives of the research. The Mann-Whitney U test (also known as the Wilcoxon rank-sum test), is a nonparametric test that is used to contrast the distributions of two independent samples (Mann & Whitney, 1947). One of its key advantages is that it does not require the assumption of normality, which makes it suitable for analysing the data in this study, which, as discussed earlier, was not normally distributed. Moreover, the test is particularly effective when comparing scores from different groups (in this study, the experimental and control groups), because it compares their medians rather than their means thereby making it a powerful tool for detecting differences in central tendency for comparing two groups when the data does not adhere to a normal distribution (Stock & Watson, 2015). Within the confines of this research, the Mann-Whitney U test was judiciously applied to conduct a comparative analysis, discerning the performance scores between the experimental group, characterized by their participation in ChatGPT-assisted learning, and the control group, distinguished by their adherence to traditional pedagogical methods in both quizzes. The use of the Mann-Whitney U test was appropriate because the scores were not normally distributed, as indicated by the results of the Kolmogorov-Smirnov and Shapiro-Wilk tests.

In econometrics, the testing for difference in means between two independent groups is usually performed by employing the t-test. This test assumes that data samples are normally distributed, and uses this assumption to ascertain whether substantial disparities exist between the means of the two groups

Specifically, the test assumes that the population from which samples are taken is divided into two independent groups, and that the means of these two groups can be compared without any mutual interference. If we fail to reject the null hypothesis (H0: μ1 = μ2), it means that there are no significant differences between the two groups. If we opt to reject the null hypothesis in favor of the alternative hypothesis (H1: μ1 ≠ μ2), it would mean that there are substantial differences between the groups' means.

However, in our study, our data showed that it was not normally distributed as shown by the outcomes of the normality assessments tests (Kolmogorov-Smirnov and Shapiro-Wilk tests). This finding indicates that the t-test, which assumes normal distribution, wouldn't be valid for the test (Stock & Watson, 2015).

Considering the non-parametric characteristics of our data, we chose to use the Mann-Whitney U test (also known as the Wilcoxon rank-sum test). This test does not require the assumption of normally distributed data and is specifically designed for comparing two independent samples (Mann & Whitney, 1947).

The Mann-Whitney U test determines whether there is a significant difference in the distributions of two independent variables. It takes into account both the ranks and the actual values of the data, which enables it to carry out a more robust comparison of the educational achievements between the two cohorts of students: the ChatGPT-assisted learning group, and the traditional learning methods group.

In the context of this study, the findings derived from the Mann-Whitney U test (p = .002) revealed a substantial distinction in the scores attained by the experimental and control groups across both quizzes. This observation intimates that the incorporation of ChatGPT-assisted learning exerted a noteworthy influence on the students' educational achievements and the retention of acquired knowledge. The Mann-Whitney U test also provides the Z statistic, which measures the quantity of standard deviations the U statistic is from the mean. In this study, the Z statistic was -2.769 for the first quiz and -3.048 for the second quiz. These values indicate that the U statistic was more than three standard deviations distant from the mean, further confirming the substantial disparity in scores between the two cohorts.

Table 8 Mann Whitney rank distribution

Ranks				
	Group	Number	Mean Rank	Sum of ranks
Test 1 Grade	Control	26	19.85	516
	ChatGPT-aided	23	30.83	709
Test 2 Grade	Control	26	19.19	499
	ChatGPT-aided	23	31.57	726
	Total	49		

Table 9 Mann Whitney test statistics

Mann Whitney Test Statistics		
	Test 1 Grade	Test 2 Grade
Wilcoxon rank-sum test	165	148
Wilcoxon signed-rank test	516	499
Z-Score	-2.769	-3.048
Asymptotic Significance (2-tailed)	0.006	0.002
Grouping Variable: Group (Controll/ChatGPT-aided)		

The use of the Mann-Whitney U test in this study furnishes substantial substantiation for the efficacy of ChatGPT-assisted learning. The test's ability to handle non-normally distributed data and its sensitivity to differences in ranks make it a powerful tool for comparing the learning outcomes of different teaching methods.

In the context of this study, the findings from the Mann-Whitney U test corroborated a notable disparity in the scores of the experimental and control groups. As a result of the descriptive statistics, the choice of Mann-Whitney U test was justified due to the non-normal distribution of the scores, thus, its usage was vital to confirm significant differences between the experimental group (ChatGPT-assisted learning) and the control group (traditional learning).

CONCLUSION AND RECOMMENDATIONS

The aim of this study was to investigate the effects of Large Language Model-assisted learning, specifically ChatGPT, versus traditional learning methods on university students' learning outcomes and knowledge retention. The outcomes of the study provide compelling evidence supporting the use of Large Language Model-assisted learning in enhancing students' knowledge of subject matter and their ability to apply knowledge effectively.

The data collected from the quizzes indicated a substantial discrepancy in the performance of students who used ChatGPT-assisted learning and those who relied on traditional learning methods. Students in the experimental group, who used ChatGPT-assisted learning, performed better on the quizzes than those in the control group, who relied on traditional learning methods. This suggests that the use of ChatGPT-assisted learning can enhance students' learning outcomes and knowledge retention.

The results of the Mann-Whitney U test further confirmed the significant difference in the scores of the experimental and control groups. The test revealed that the use of ChatGPT-assisted learning had a significant impact on the students' learning outcomes and knowledge retention. This finding is in line with previous research that has highlighted the potential of AI-assisted learning tools in enhancing students' academic performance (Chen, L., Chen, P., & Lin, Z., 2020; Maghsudi, S., Lan, A., Xu, J., & van Der Schaar, M., 2021).

The results of the study also showed a wide range of scores for both quizzes, indicating a diverse group of students with varying levels of knowledge and understanding and ability to apply the knowledge and understanding acquired. This suggests that while ChatGPT-assisted learning can enhance overall learning outcomes, it may not be equally effective for all students. This finding underscores the importance of personalized learning approaches that accommodate the individual requirements and aptitudes of students..

The study also highlighted the challenges that students faced in the second quiz, where they were not allowed to use any assistance other than calculators. The diminished mean score and elevated standard deviation for the second quiz suggest that the students found the quiz more challenging than the first one. This underscores the significance of furnishing suitable assistance and resources to aid students in the application of knowledge effectively, particularly in challenging learning contexts.

The future of work will increasingly involve complex problem-solving tasks that require the synergy of human and AI. As such, it is critical to provide students with the necessary support and resources to develop their problem-solving skills and their ability to work effectively with AI systems.

The findings of this study have important implications for educators and policymakers. The significant impact of ChatGPT-assisted learning on students' learning outcomes and knowledge retention suggests that Large Language Model-assisted learning can be an effective tool for enhancing university education. Educators can incorporate ChatGPT-assisted learning into their teaching strategies to facilitate students' understanding of complex concepts and improve their ability to apply knowledge effectively.

Given the increasing prevalence of AI in the workforce, educators need to incorporate AI-assisted learning tools into their teaching strategies. This will not only enhance students' learning outcomes but also equip them with the necessary skills to interact effectively with AI systems in the workforce.

The study also highlights the need for further research to explore the potential of Large Language Model-assisted learning in different educational contexts. Future studies could investigate the impact of ChatGPT-assisted learning on students' learning outcomes in different subject areas, or explore its effectiveness in enhancing students' critical thinking and problem-solving skills.

As the integration of AI systems into industries becomes more prevalent, it is critical to explore the potential of Large Language Model-assisted learning in different educational contexts. Future research should investigate the impact of AI-assisted learning on students' critical thinking and problem-solving skills, which are crucial for the future workforce.

Given the rapid integration of AI systems into industries, it is critical that students learn how to use and prompt LLMs effectively. Without these skills, they risk being less productive and less competitive in the future workforce. The education sector needs to respond to this trend by incorporating AI-assisted learning tools into teaching strategies and providing students with the necessary support and resources to develop their problem-solving skills and their ability to work effectively with AI systems.

In conclusion, this study provides robust evidence supporting the use of Large Language Model-assisted learning, specifically ChatGPT, in enhancing university students' learning outcomes and knowledge retention. The findings suggest that ChatGPT-assisted learning can be an effective tool for improving students' understanding of complex concepts and their ability to apply knowledge effectively. However, the study also highlights the need for further research to explore the potential of ChatGPT-assisted learning in different educational contexts and the importance of personalized learning approaches.

Based on the findings of this study, the following recommendations are proposed:

1. Incorporate Large Language Model-assisted learning into university education: Given the significant impact of ChatGPT-assisted learning on students' learning outcomes, it is recommended that educators incorporate this tool into their teaching strategies. This will not only enhance students' understanding of complex concepts but also equip them with the necessary skills to interact effectively with AI systems, a critical skill in the future workforce. It is understandable that some academic institutions may give pushback against implementing or rather requiring the use of Large language models into education, as some may view it as cheating, however the rapid advancement of such a technology and its incorporation into the workforce, especially in the future, there will be a necessity for knowledge on properly prompting large language models to get a proper output. The pushback from academic institutions is

similar to some pushback on any new technology, and merely stems from misunderstanding the potential and the applications of such technologies.

2. Develop training programs for effective use of LLMs: To ensure students can effectively use and prompt LLMs, universities should consider developing training programs. These programs could provide students with the necessary skills and knowledge to utilize these tools to their full potential, thereby enhancing their productivity and competitiveness in the future workforce.

3. Provide appropriate support and resources for students: The challenges faced by students in the second quiz underscore the importance of providing appropriate support and resources. This could include additional instruction or resources, or offering tutoring or mentoring programs. Furthermore, resources should be developed to help students apply knowledge effectively when working synergistically with AI systems.

4. Implement personalized learning approaches: The wide range of scores suggests that students have different learning needs and abilities. Therefore, it is recommended that educators design and implement learning strategies that cater to the individual needs of students. AI-assisted learning tools could be utilized to provide personalized learning experiences.

5. Conduct further research on the effectiveness of Large Language Model-assisted learning: While this study provides compelling evidence supporting the use of ChatGPT-assisted learning, further research is needed to confirm these findings and explore the potential of Large Language Model-assisted learning in different educational contexts. This could include investigating the impact of AI-assisted learning on students' critical thinking and problem-solving skills, which are crucial for the future workforce.

6. Prepare students for the AI-integrated workforce: Given the increasing prevalence of AI in the workforce, it is recommended that universities incorporate AI literacy into their curriculums. This would equip students with the necessary skills to navigate and contribute effectively in an AI-integrated workforce.

7. Foster human-AI synergy in learning: As the future of work will increasingly involve complex problem-solving tasks that require the synergy of human and AI, it is critical that universities foster this synergy in their learning environments. This could involve integrating AI-assisted learning tools into group projects and assignments, thereby providing students with practical experience of working with AI.

8. Advocate for policy changes: Given the implications of the findings, it is recommended that educators and policymakers advocate for policy changes that support the integration of AI-assisted learning tools in university education. This could involve lobbying for increased funding

for AI research and development in education, or advocating for policy changes that support the training of educators in the use of AI-assisted learning tools.

In conclusion, these recommendations aim to leverage the potential of Large Language Model-assisted learning to enhance university students' learning outcomes and knowledge retention, and to prepare them for the future workforce. The rapid integration of AI systems into industries underscores the urgency of these recommendations. It is hoped that these recommendations will guide educators, policymakers, and researchers in their efforts to enhance university education in the era of AI.

REFERENCES

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. Journal of educational psychology, 51(5), 267.

Baker, R. S. J., & Inventado, P. S. (2014). Chapter X: Educational Data Mining and Learning Analytics. Comput. Sci, 7, 1-16.

Basham, James D., et al. "An operationalized understanding of personalized learning." Journal of Special Education Technology 31.3 (2016): 126-136.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Campbell, D., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally.

Chan, C. K. Y., & Tsi, L. H. (2023). The AI Revolution in Education: Will AI Replace or Assist Teachers in Higher Education?. arXiv preprint arXiv:2305.01185.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. Ieee Access, 8, 75264-75278.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. Computers and Education: Artificial Intelligence, 1, 100002.

Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Sage publications.

Driscoll, M. P. (2000). Psychology of learning for instruction. Needham. MA: Allyn & Bacon.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. Current Directions in Psychological Science, 23(5), 374-380.

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil transactions on benchmarks, standards and evaluations, 2(4), 100089.

Hoffman, J., Owen, J., & Calvert, S. L. (2021). Parasocial relationships with conversational agents: Experiences of parents. Human-Computer Interaction, 36(6), 548-580.

Hiremath, P. S., Kumar, P., Bhirud, S. G., & Mujawar, A. B. (2018). The development of a chatbot for an educational system using deep learning. International Journal of Engineering and Technology, 7(3.7), 390-393.

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own?. Memory, 17(4), 471-479.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist, 41(2), 75-86.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., & Mattick, A. (Year). OpenAssistant Conversations -- Democratizing Large Language Model Alignment.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. Review of educational research, 86(1), 42-78.

Maghsudi, S., Lan, A., Xu, J., & van Der Schaar, M. (2021). Personalized education in the artificial intelligence era: what to expect next. IEEE Signal Processing Magazine, 38(3), 37-50.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, 50-60.

Lee, M. K., Kavya, P., & Lasser, J. (2021). The design and implications of social interactions with an intelligent virtual agent. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-21.

Novak, J. (1998). Learning, creating and using knowledge. Concept Maps™ as facilitative tools in schools and in corporations. London: Lawrence Erlbaum.

OpenAI, R. (2023). GPT-4 technical report. arXiv, 2303-08774.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational psychologist, 38(1), 1-4.

Pfeffer, O. P., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., ... & Kasneci, G. (2021). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. arXiv preprint arXiv:2107.07650.

Piaget, J. (1970). Science of education and the psychology of the child. Trans. D. Coltman.

Rana, S. (2023). AI and GPT for Management Scholars and Practitioners: Guidelines and Implications. FIIB Business Review, 12(1), 7-9.

Ramadan, R. A., Farah, R. S., & El Essrawi, M. A. (2021). The emergence of conversational agents in redefining companionship and interdependence for individuals with special needs. ACM Transactions on Accessible Computing (TACCESS), 15(1), 1-22.

Roll, I., Wiese, E. S., Long, Y., Aleven, V., & Koedinger, K. R. (2014). Tutoring self- and co-regulation with intelligent tutoring systems to help students acquire better learning skills. Design recommendations for intelligent tutoring systems, 2, 169-182.

Schunk, D. H. (2012). Learning theories an educational perspective. Pearson Education, Inc.

Shadish, W. R. (2002). Revisiting field experimentation: field notes for the future. Psychological methods, 7(1), 3.

Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions.

Stock, J. H., & Watson, M. W. (2015). Introduction to econometrics (3rd updated edition). Age (X3), 3(0.22).

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285.

Sweller, J., Ayres, P., Kalyuga, S., Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. Cognitive load theory, 71-85.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenç”on, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora

Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, et al. (292 additional authors not shown) (2023). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv preprint arXiv:2211.05100 [cs.CL]. Retrieved from https://arxiv.org/abs/2211.05100

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Vygotsky, L. S., & Cole, M. (1978). Mind in society: Development of higher psychological processes. Harvard university press.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. Educational and Psychological Measurement, 67(2), 219-238.

Woolf, B. P. (2010). A roadmap for education technology.