

# Comparative Study on Accuracy of Responses by Select AI Tools: ChatGPT and Perplexity AI Visa Vee Human Responses

<sup>\*1</sup>Salmon Oliech Owidi; <sup>2</sup>Joanne Nabwire Lyanda; <sup>3</sup>Eric W. Wangila

<sup>1</sup>Tom Mboya University

<sup>2,3</sup>Masinde Muliro University of Science and Technology

Corresponding Author:- <sup>\*1</sup>Salmon Oliech Owidi

**Abstract:-** This study explored questions whose solutions were provided by human experts, ChatGPT, and Perplexity AI. The responses were triangulated in discussions to identify oversights, alternative frames, and biases against human-generated insights. ChatGPT and Perplexity AI were selected due to their popularity, with ChatGPT gaining over 100 million users and Perplexity AI 87 million within a year. Educational specialists submitted questions across various fields, along with their responses, which were subsequently posed to the AI tools. These responses were coded and evaluated by twelve educational specialists and subject matter experts (N = 24) based on scientific accuracy, actionability, and comprehensibility. Descriptive statistics indicated that Human Experts achieved significantly higher mean scores in both Scientific Accuracy (M = 7.42, SD = 0.65) and Actionability (M = 7.25, SD = 0.77) compared to ChatGPT (M = 6.25, SD = 0.71; M = 5.42, SD = 0.99) and Perplexity AI (M = 4.33, SD = 0.79; M = 4.17, SD = 1.06). In terms of Comprehensibility, ChatGPT led with a mean score of 6.58 (SD = 0.99) compared to Human Experts (M = 7.08, SD = 1.24) and Perplexity AI (M = 5.43, SD = 0.55). Kruskal-Wallis tests revealed significant differences across all dimensions ( $p < 0.001$  for Scientific Accuracy and Actionability;  $p = 0.015$  for Comprehensibility). Post-hoc Dunn's tests confirmed that Human Experts outperformed both AI tools, while ChatGPT was significantly more comprehensible than Perplexity AI. These findings highlight the limitations of AI in delivering scientifically accurate and actionable insights due to factors like lack of emotional intelligence and common sense. The study recommends careful evaluation of AI integration in academic and research contexts to better understand their roles and limitations.

**Keywords:-** Artificial Intelligence Tools, ChatGPT, Perplexity AI, Comparative Study.

## I. INTRODUCTION

ChatGPT, released in November 2022, exemplifies the advancements in artificial intelligence (AI) chatbots, showcasing sophisticated human-like language comprehension and synthesis capabilities (Donelan, 2024). This technology relies on Reinforcement Learning with

Human Feedback (RLHF) to refine its responses, learning from interactions to identify the most contextually appropriate replies (Ellis, 2024). This approach allows ChatGPT to engage users in a conversational manner, enhancing its utility across various domains, including education and customer support. Studies have shown that AI chatbots can significantly improve user engagement and satisfaction, facilitating more interactive and personalized experiences (Baker & DeMarte, 2023).

Similarly, Perplexity AI, introduced in August 2022 by Andy Konwinski, integrates AI with intelligent search capabilities, delivering precise and comprehensive answers to user inquiries (Aayush, 2024). It leverages natural language processing to enable users to interact in a conversational tone, effectively bridging the gap between search engines and chatbots (Cao et al., 2023). By aggregating data from diverse online sources, Perplexity provides contextualized responses, further illustrating the potential of AI in information retrieval. Research by Zhang et al. (2023) indicates that such AI tools can enhance the speed and accuracy of information sourcing, making them invaluable for users seeking quick answers.

Recent discussions in the realm of medical writing, as highlighted by Fingerhut et al. (2023), have raised critical questions about the role of AI in producing authoritative medical content. While AI tools like ChatGPT and Perplexity can generate engaging and coherent text, concerns persist regarding their ability to replicate human reasoning and nuanced argumentation essential for medical advice and policy formulation. Studies indicate that while AI can analyze vast amounts of data efficiently, it often lacks the capacity to interpret complex ethical considerations and contextual subtleties (Thompson, 2024). Moreover, the propensity of these models to fabricate citations or generate misleading references underscores the need for caution in their application within professional fields.

As pointed out by Kohn (2023), the integration of AI in medical writing raises ethical dilemmas, particularly regarding accountability and the potential for misinformation. Furthermore, several studies have pointed out the deficiencies in responses generated by AI bots, highlighting shortcomings in both accuracy and precision.

These findings reveal a critical gap in the current understanding of AI's reliability, particularly in high-stakes environments like higher education. The high rate of adoption and use of AI bots in learning institutions is alarming, necessitating a thorough investigation into their accuracy across various fields of study. This situation underscores the importance of developing effective strategies for integrating these technologies into educational frameworks to enhance learning outcomes while mitigating risks.

The need for a structured investigation into the accuracy of responses generated by popular AI bots becomes increasingly evident. Given the widespread implementation of AI tools in educational settings, it is essential to assess their reliability when posed with questions spanning different academic disciplines. Engaging expert personnel to evaluate the responses of these AI systems can provide invaluable insights into their strengths and weaknesses, ultimately guiding institutions in their integration strategies. This research is a necessary step toward ensuring that AI technology serves as a beneficial complement to traditional learning methods, fostering an environment where both educators and students can confidently utilize these tools for enhanced educational experiences.

## II. LITERATURE REVIEW

To understand what ChatGPT and Perplexity AI are, the following questions were asked posed to the two AI bots respectively. "In 100 words, what is ChatGPT?", "In 100 words, what is Perplexity AI?". The tools then gave the following responses in a few seconds.

### ➤ ChatGPT:

*ChatGPT is an advanced language model developed by OpenAI, characterized by its ability to generate human-like text based on prompts provided by users. It operates through natural language processing, understanding and generating responses in a conversational manner. Trained on vast amounts of text data, it encompasses a wide array of knowledge and can engage in diverse topics ranging from casual conversation to technical inquiries. ChatGPT serves as a versatile tool for tasks like content generation, language translation, and virtual assistance, offering a glimpse into the potential of artificial intelligence in understanding and generating human language.*

### ➤ Perplexity AI:

*Perplexity AI is a cutting-edge AI-powered conversational search engine that provides answers to complex queries using large language models and relevant sources with citations (PerplexityAI, 2023). Launched in 2022, it has gained popularity by offering a more conversational, transparent, and personalized search experience (OriginalityAI, 2023). Perplexity AI is not yet profitable but has a wide range of users, including researchers, writers, artists, musicians, and programmers (OriginalityAI, 2023). It uses advanced natural language processing algorithms, such as OpenAI's GPT technology, to*

*understand user queries and deliver contextual knowledge and appropriate responses (PerplexityAI, 2023). The platform is designed to provide instant answers, fact-checking capabilities, and customized responses across various domains, challenging established search engines like Google (PerplexityAI, 2023). Perplexity AI's unique approach to search has garnered significant user interest and is backed by robust funding, positioning it as a promising contender in the online search market (OriginalityAI, 2023).*

ChatGPT gave its responses using 95 words while Perplexity AI explained itself using 135 words excluding the references it gave.

Most of previously published studies examined whether AI agents might offer information by responding to inquiries about a certain topic using generally comparable methodology (Haver et al, 2023; Hinton et al, 2023; Johnson et al., 2023). The most common paradigm was creating a set of questions about a certain subject and submitting them to one or several AI agents; certain studies also included questions of different levels of difficulty. After that, a panel of reviewers evaluated the final product, and the comments were graded. These evaluations were used to assess the AI agents' ability to deliver accurate and pertinent information.

Johnson et al. presented the first study evaluating ChatBots precision in responding to oncologic inquiries in December 2022 (Vallance, 2024). Five reviewers with expertise in the field assessed ChatGPT's answers to 13 questions taken from the "Common Cancer Myths and Misconceptions" homepage and contrasted the findings with those from the National Cancer Institute. ChatGPT offered largely correct information on this subject, despite significant variations in readability and word quantity (Yeo et al, 2023). Yeo et al. then evaluated ChatGPT's capacity to respond to inquiries on the treatment and psychological assistance provided to patients with cirrhosis and hepatocellular carcinoma (HCC) (Zhu, 2023). A pair of reviewers evaluated answers to 73 HCC questions. Nearly 75% of the responses were thought to be accurate overall.

In a study conducted by Cao et al (2023) on responses by ChatGPT responses on health questions led to the following results.

First Author	Year of publication	Large Language Model		Domain	Questions (n)	Reviewers (n)
		ChatGPT	Other			
Johnson SB (17)	2023	Free version	–	"Common Cancer Myths and Misconceptions" (NCI web page)	13	5
Yeo YH (18)	2023	Free version	–	Cirrhosis or HCC	164	2
Cao JJ (19)	2023	Free version	–	HCC diagnosis and surveillance	20	6
Haver HL (20)	2023	Free version	–	Breast cancer prevention and screening	25	3
Moazzam Z (21)	2023	Free version	–	Pancreatic cancer surgical care	30	20
Coskun B (22)	2023	Free version	–	Prostate cancer diagnosis and treatment	59	2
Zhu L (23)	2023	Free and paid versions	YouChat, NeevaAI, Perplexity, Chatsonic	Prostate cancer diagnosis and treatment	22	3
Rahsepar AA (24)	2023	Free version	Google Bard	Lung cancer prevention and screening	40	2
Sorin V (25)	2023	Free version	–	Breast cancer clinical cases	10	2
Schulte B (26)	2023	Free version	–	Systemic therapies for advanced solid tumors	51	NA
Haemmerly J (27)	2023	Free version	–	Brain cancer clinical cases	10	5
O'Hern K (28)	2023	Free version	–	Common and rare cutaneous cancer clinical cases		NA

The findings of a study evaluating ChatGPT's capacity to offer data on liver cancer surveillance and diagnosis were published by Cao et al (2023). Six reviewers evaluated the twenty questions that were submitted to ChatGPT. Overall, these findings showed how poorly ChatGPT performed in providing information on radiological diagnosis and liver cancer surveillance (Cao et. al, 2023). The findings of a retrospective study conducted by Haver et al. to determine whether ChatGPT should be used to provide advice on breast cancer screening and prevention were released. A total of twenty-five questions were uploaded to ChatGPT, and most of the responses were deemed suitable by the reviewers (Haver et al, 2023).

The findings of observational research evaluating ChatGPT's performance in responding to 30 questions about surgical care in pancreatic care were published by Moazzam et al. in 2023. After being assessed by thirty reviewers, the results showed that ChatGPT could provide high-quality responses in this domain (Moazzam et al, 2023). The findings of a study evaluating ChatGPT's effectiveness in disseminating information about prostate cancer were published by Coskun et al. Two reviewers evaluated fifty-nine questions that were taken from the patient information platform of the European Association of Urology. The outcomes demonstrated that ChatGPT's content generation was not at its best in terms of accuracy and quality. (Coskun, 2023).

AI chatbots affect numerous aspects of our everyday life as well as a variety of fields and practices. They can be used to write speeches, emails, articles, computer code, and translations between languages. They can also be used to alter the tone of writing. Because they make entry barriers smaller, seamlessly integrate with human labor, and increase our productivity and creativity, they have the potential to be empowering (Liang, 2023; Brynjolfsson, 2023; Li et al, 2023). The claim that AI chatbots would eliminate the monotony of routine office work by automating various jobs and eventually boosting productivity across the economy is made by several people, primarily by those who have an interest in the technology (Microsoft, 2023; Donelan, 2023).

But AI chatbots may also be alarming, and many people believe they could be dangerous and have far-reaching negative effects. (Kleinman, 2024) They might intensify the prejudices we already have, erode our confidence in the reliability of the information, and rob us of our ability to distinguish between what is genuine and what isn't. Not to mention, they will probably upend a lot of work that has traditionally been done by highly rewarded professionals like writers, artists, programmers, CEOs, and others in the creative and knowledge industries (Kirk, 2023).

A wide range of significant and urgent risks associated with AI are being discussed more and more by AI experts, journalists, policymakers, and the public. These risks include the AI race, organizational risks, rogue AIs, the reinforcement of social inequality, the transformation of

labor and expertise, and the aggravation of environmental injustices (Hinton, 2023). Calls for a pause in AI development have resulted from concerns about safety (Kleinman, 2023), capabilities, massive workforce redundancy (Vallance, 2024), and legality (Li et al, 2023) although it is unclear if such attempts would have any effect (Vallance, 2024).

In summary, Li et al. (2023) state that there are just as many potentials as threats associated with generative AI for our lives, communities, and society. Some people think that generative AI will improve both our jobs and our lives in general. Others think that regions where sentient beings are best suited to navigate will be disastrously invaded by generative AI. All agree, however, that precautions are desperately needed (Madiaga, 2023), and the initial measures have already been implemented (White House FACT SHEET, 2023).

This is especially true for artificial general intelligence (AGI), which is the anticipated next step. Artificial general intelligence (AGI) refers to systems that are predicted to exist within the next ten years and possess a higher level of intelligence than humans. The world might have been very different from what it is today if the first artificial general

intelligence had been developed (Altman, 2023). Furthermore, what role does academia play in the creation and composition of scientific publications and research? Academic publishing and academia could undergo a change thanks to AI chatbots (Lund et al., 2023). Given that academics and students made up two of the top three occupational groupings among ChatGPT's early users, it appears that the academic sector will be among the first to go through this process (Haque et al., 2022).

### III. METHODOLOGY

The study adopted a descriptive design approach. The goal was to find out the accuracy of ChatGPT, Perplexity AI compared to responses provided by human experts. Questions were generated by subject experts in six areas. Experts in the domains from which the questions were derived were requested to submit the common questions they interact with along with their responses. The three categories of responses were further subjected to analysis by 24 other experts. All the lecturers were from Masinde Muliro University of Science and Technology. The following questions were posed to ChatGPT, Perplexity and Human Expert who specialists in various fields are.

Table 1 Common questions raised by the experts from various disciplines.

S/N	FIELD OF STUDY	QUESTIONS (To be analyzed by AI tools and Human Expert)
1	Computer education	What is the status of computer laboratories in secondary schools in Kenya.
2	Mathematics	There are 49 dogs signed up to compete in the dog show. There are 36 more small dogs than large dogs signed up to compete. How many small dogs are signed up to compete
3	Religion	In the build up to the crucifixion of Christ, a number of events are put in play for this to happen. Discuss each event with their significance to the current faith and belief in Christianity
4	English / literature	Greed is dehumanizing. Using Samaritan by John Lara discuss
5	Kiswahili	Ndoa ya Yona na Sera ni kama Bembea, dhibitisha kauli hii kwa kurejelea Bembea ya Maisha. Alama 10.
6	Instructional design	Design a learning experience for a learning area in teaching of the topic 'loci' putting in consideration the magic triangle of learning.

The analysis of the responses obtained from the AI tools and Human expert were then graded based on the following rubric on a scale 0 to 10.

Table 2 Description of Grading components for the responses provided by Human Experts and the AI tools.

S no	Grading Component	Description
1	Scientific Correctness	How well each response captures the present status of knowledge in the scientific domain that the question is related to. When grading scientific correctness, it is important to consider the requested word count of the answers (50–100 words) as well as the inherent constraints on extensive explanation and the degree this imposes. It's important to consider the layman's target audience and their anticipated degree of scientific and domain specific expertise.
2	Comprehensibility	How well the average person could be expected to understand the response. The answer's substance should be the primary concern when it comes to comprehensibility, but grammatical mistakes may also be considered.
3	Actionability	The extent to which the answers provided actionable knowledge that the hypothetical layperson posing the inquiry can utilize.



Each grader was given a score ranging from 0 to 10, which enabled them to assess the responses collectively and in relation to all grading criteria. Using permutation tests and the function `perm.test` from the package `jmuOutlier` in R software (R Core Team, 2022) with the test statistics set to mean, the statistical significance of group differences was determined. To estimate the degree of evidence of a difference between the groups, p values were approximated from 100,000 simulations for both the overall score and the component scores for each question. The study was registered, and Kirk et al. (2023) provided a more detailed description of it on the Open Science Framework. Given the statistical nature of the data obtained, the sole alteration was to use nonparametric tests (permutation tests) in place of t-tests.

#### ➤ Respondents

The questions for the study were directed to a carefully selected group of respondents, each chosen for their expertise in specific fields. These individuals provided insights based on their experiences, contributing valuable perspectives to the research. The fields covered included

computer science, religion, mathematics, English literature, Kiswahili, and instructional design. Following the initial responses from both AI tools and human experts, a separate group of 24 experts across these six disciplines evaluated the answers. This evaluation process was conducted blind, meaning the raters did not know whether the responses came from AI or human contributors. The experts were randomly selected from various universities participating in the study, ensuring a diverse range of perspectives based on their qualifications in their respective fields. Their ratings of the responses formed the foundation for further analysis, involving a total of 30 participants throughout the research.

## IV. FINDINGS

This section provided the responses provided by ChatGPT, Perplexity AI and the Human Expert. The findings also represent the ratings on a scale of 0-10 by other respondents picked from various universities.

➤ **QUESTION 1:** *What is the status of computer laboratories in secondary schools in Kenya.*

Response Agent	Response
ChatGPT	Computer laboratories in Kenyan secondary schools vary widely in quality and availability. While some well-funded institutions boast modern facilities with up-to-date equipment and internet access, many face challenges. Budget constraints often lead to outdated computers, inadequate software, and unreliable internet connections. Rural schools particularly struggle, lacking resources and facing infrastructural limitations. Consequently, students in such schools may have limited exposure to technology, hindering their digital literacy and competitiveness in a globalized world. Efforts to improve this situation include government initiatives and NGO interventions, but significant disparities persist, impacting the educational opportunities of Kenyan youth.
Perplexity AI	In Kenya, the integration of Information Communication and Technology (ICT) in education has been limited to computer literacy training in secondary schools [2]. Although the government has introduced policies to improve ICT integration in education, there are still challenges in implementing these policies. For instance, most schools lack specialized and simulated academic program software, and there is a shortage of custom-made digital material with high relevant content for Kenyan classrooms [2]. Additionally, the lack of sufficient infrastructure investment funds and low-quality buildings have resulted in some existing school infrastructure being in poor condition [3]. While most schools have classrooms and latrines/toilets, there is a shortage of computer laboratories and dining halls [3]. The government has made progress in improving school infrastructure, but there is still a need for more infrastructure investment funds to ensure that all schools have adequate facilities. The integration of ICT in education can go a long way in improving education delivery, but there is a need for more investment in custom-made digital material and specialized academic program software to tap into the real potential of ICT for learning.
Human Expert	<p>The status of computer laboratories in secondary schools in Kenya varies, but there have been efforts to improve access to ICT (Information and Communication Technology) resources. Here are some key points:</p> <ol style="list-style-type: none"> <li><b>Kenya Digital Schools Project:</b> <ul style="list-style-type: none"> <li>Computer Aid International, in conjunction with ICDL (International Computer Driving License), established the <b>Kenya Digital Schools Project</b>. This initiative has set up ICT laboratories in 18 schools in Naivasha, Kenya.</li> <li>Each of these beneficiary schools received ICT infrastructure for the first time, providing students with access to 21st-century resources.</li> <li>The project supplied 360 PCs, 54 laptops, and teacher aids, impacting up to 6,000 students from underprivileged communities.</li> </ul> </li> <li><b>Challenges and Constraints:</b> <ul style="list-style-type: none"> <li>Despite progress, universal implementation of ICT in schools remains challenging due to limited resources, poor country-wide ICT infrastructure, and inadequate electrical supply, especially in rural areas.</li> <li>Many schools lack the necessary technological resources and trained teachers to effectively implement</li> </ul> </li> </ol>

	<p>mandatory IT training and digital curricula.</p> <p>Inadequate funds for computers, accessories, and infrastructure, as well as a shortage of trained computer teachers, hinder progress.</p> <p><b>3. Recent Developments:</b></p> <p>The government has provided standardized laboratory kits for Junior and Senior secondary education in line with the Competency-Based Curriculum (CBC) system.</p> <p>These kits aim to enhance practical learning experiences for students as they transition to the new curriculum.</p> <p><b>4. Challenges in Science Laboratories:</b></p> <p>A study in Wolaita Zones found that laboratory work in secondary schools were at a very low level. Constraints included lack of laboratory rooms, inadequate supply of lab equipment and reagents, absence of trained laboratory technicians/teachers, and poor management and evaluation of laboratory activities.</p>
--	--

➤ **QUESTION 2:** *There are 49 dogs signed up to compete in the dog show. There are 36 more small dogs than large dogs signed up to compete. How many small dogs are signed up to compete*

Response Agent	Response
ChatGPT	<p>Let's denote the number of large dogs as L and the number of small dogs as S. We know that <math>S = L + 36</math>, and the total number of dogs is 49, so <math>S + L = 49</math>.</p> <p>Substituting <math>S = L + 36</math> into the second equation:</p> $(L + 36) + L = 49$ $2L + 36 = 49$ $2L = 49 - 36$ $2L = 13$ $L = 13 / 2$ $L = 6.5$ <p>Since we can't have half a dog, we'll consider L as 6.</p> <p>Now, substituting <math>L = 6</math> into <math>S = L + 36</math>:</p> $S = 6 + 36$ $S = 42$ <p>Therefore, there are 42 small dogs signed up to compete.</p>
Perplexity AI	<p>There are 49 dogs signed up for the dog show, with 36 more small dogs than large dogs. To determine the number of small dogs, let x represent the number of large dogs. Therefore, the number of small dogs would be <math>x + 36</math>. The total number of dogs is 49, so the equation becomes <math>x + (x + 36) = 49</math>.</p> <p>Simplifying, we get <math>2x + 36 = 49</math>, leading to <math>2x = 13</math> and <math>x = 6.5</math> large dogs. Since we can't have half dogs, the numbers provided are inaccurate. The possible scenarios are either 6 large and 43 small dogs or 7 large and 42 small dogs, neither of which aligns with having 36 more small dogs[1].</p>
Human Expert	<p>Let's break this down step by step:</p> <ol style="list-style-type: none"> <li>Let (x) represent the number of large dogs.</li> <li>Since there are 36 more small dogs than large dogs, the number of small dogs is <math>(x + 36)</math>.</li> <li>The total number of dogs is given as 49: <math>[x + (x + 36) = 49]</math></li> <li>Combine like terms: <math>[2x + 36 = 49]</math></li> <li>Subtract 36 from both sides of the equation: <math>[2x = 13]</math></li> <li>Divide both sides by 2: <math>[x = 6.5]</math></li> </ol> <p>Since we can't have half a dog, we'll consider the number of large dogs as a whole number. Therefore, there are 6 large dogs and 42 small dogs signed up to compete in the dog show.</p>

➤ **QUESTION 3:** *In the build up to the crucifixion of Christ, a number of events are put in play for this to happen. Discuss each event with their significance to the current faith and belief in Christianity*

Response Agent	Response
ChatGPT	<p>Several significant events lead to the crucifixion of Christ. Firstly, Jesus' betrayal by Judas Iscariot symbolizes human frailty and the consequences of betrayal. His trial before Pontius Pilate highlights the unjust condemnation faced by the innocent. The crucifixion itself embodies divine sacrifice and redemption, central to Christian theology. Jesus' resurrection reaffirms belief in eternal life and triumph over death. These events form the cornerstone of Christian faith, emphasizing themes of forgiveness, salvation, and the love of God. They inspire believers to reflect on the teachings of Jesus, his ultimate sacrifice, and the promise of</p>

	redemption through faith.
Perplexity AI	<p>The events leading up to the crucifixion of Christ hold immense significance in Christianity. Firstly, Jesus raising Lazarus from the dead demonstrated His power and authority, triggering the plot to kill Him. Secondly, being anointed at Bethany symbolized preparation for His burial. The Last Supper, a model for Holy Communion, emphasized the sacrificial nature of His impending death. Lastly, the poignant moment in the Garden of Gethsemane showcased Jesus' humanity, obedience to God's will, and commitment to prayer. These events collectively underscore Jesus' divinity, sacrificial love, and the fulfillment of God's plan for salvation, shaping the core beliefs of Christianity.</p>
Human Expert	<p>By riding a donkey, Jesus meant he accepted to be a beast of burden. Meaning he embraced the worst treatment, a shameful death. When a donkey is done serving his master it's normally killed, and its flesh is buried. Unlike normal donkey, this donkey Jesus surprises people. He arises from death. Jesus believed in a philosophy of humanity and pragmatism. He associated more with the poor, the weak and downtrodden</p> <p><b>Legal complaints:</b> Jesus is accused of several counts of breach against the law of Moses.</p> <p><b>Main accusers /Accusations:</b> Pharisees accusing of procedural issues like incitement Priests causing him of blasphemy</p> <p><b>Litigation:</b> Veracity of the accusation tested in front of key legal figures, especially Herod and Pontius Pilate. Herod decides that the accusations have no legal substance.</p> <p>They then forwarded him to Pontius Pilate of the Roman government. Mr. Pilate finds no substance in the case too.</p> <p><b>Blackmail by High priest</b></p> <p>The priest was not happy with Pilate. They had prejudicial verdict: That Jesus deserved death. So, they accused Pilate of being insensitive to Roman interests. Pilate was afraid he would lose his job. So, he decides to have Jesus fatally whipped to near death to make the priests happy. However, they insist Jesus must be killed. Pilatus bows to pressure against his conscience. Washes his hands in public meaning he is innocent. Gives ok to crucifixion.</p> <p><b>Significance:</b></p> <p>Like their Jewish counterpart, Christians today are either propagators of blackmail or victims of blackmail. E.g. Pastor Kanyari, Kavula church in South Africa</p> <p>2. Denial of Truth: Christians today fear investigation into the veracity of faith. The priests never wanted to listen to verdict of Herod or Pilate. Today's Christians would prefer epistemic darkness than Truth</p>



➤ **QUESTION 4:** Greed is dehumanizing. Using Samaritan by John Lara discuss

Response Agent	Response
ChatGPT	<p>In "Samaritan" by John Lara, greed dehumanizes characters like Pedro, driving him to exploit others for personal gain. Despite encountering the injured Samaritan, Pedro's selfishness blinds him to compassion. His obsession with material wealth overrides empathy, leading to moral degradation. Conversely, the Samaritan embodies humanity by demonstrating selflessness and kindness. His actions reveal the transformative power of compassion, emphasizing the importance of empathy in preserving human dignity. Through Pedro's downfall and the Samaritan's example, Lara illustrates how greed strips individuals of their humanity, while acts of altruism restore and affirm it, echoing timeless moral truths.</p>
Perplexity AI	<p>The play "The Samaritan" by John Lara highlights the dehumanizing effects of greed, as seen through various conflicts and antagonisms. Greed, as a driving force, leads to moral decay and decadence, dehumanizing individuals and society. The play showcases a society driven by materialism, where people prioritize self-enrichment above all else, leading to corruption and dishonesty. Characters in the play are often dishonest, even within a religious context, further emphasizing the destructive nature of greed. The play underscores the idea that a society driven by greed suffers greatly, as it molds suspicious leaders who prioritize escaping punishment over working towards progress. Through "The Samaritan," Lara illustrates how greed can dehumanize individuals and societies, leading to conflict, corruption, and moral decay.</p>
Human Expert	<p>People in power are so greedy that they forget their humanity and focus on material wealth. They can do anything including killing, bribery and other vices to get wealth. They don't care about other human beings. The likes of Mosi in Samaritan. They intimidate teacher Nicole who is behind the Samaritan app to safeguard their greedy interest at the expense of those who elected them. Bribery killing intimidation by leaders on those who try to raise their voice on the bad governance. Samaritan app was invented to expose the dehumanizing greed in Maracas Municipality. The Samaritan app was created to help people expose the ills of their leaders.</p>

➤ **QUESTION 5:** *Ndoa ya Yona na Sera ni kama Bembea, dhibitisha kauli hii kwa kurejelea Bembea ya Maisha. Alama 10.*

Response Agent	Response
ChatGPT	Katika "Bembea ya Maisha," ndoa ya Yona na Sera inaonyeshwa kama bembea. Yona anaanza kuoa Sera kwa furaha na matumaini, lakini ndoa yao inakumbwa na changamoto na msukosuko kama vile bembea inavyojaribu ustahimilivu wake. Migongano ya kifamilia na matatizo ya kifedha huwafanya wapate misukosuko. Hata hivyo, kama bembea inavyoshikilia imara kati ya mawimbi, upendo wao na uaminifu huwaunganisha tena. Mchakato huu unaonyesha kwamba ndoa ni safari yenye misukosuko, lakini kwa kusimama imara na kushikamana, wanaweza kushinda changamoto hizo na kuendelea mbele kwa furaha na umoja.
Perplexity AI	Kulinganisha ndoa ya Yona na Sera na Bembea ya Maisha inaonyesha jinsi matukio yanavyoweza kufanana. Kama ilivyokuwa kwa matukio ya Yesu kabla, wakati, na baada ya kusulubiwa, ndoa inaweza kupitia vipindi vya furaha, majaribu, na mabadiliko. Kama Bembea ya Maisha inavyoashiria mzunguko wa maisha, ndoa inaweza kuwa na vipindi vya utulivu na vya changamoto. Kwa kina, ndoa inaweza kufanana na Bembea ya Maisha kwa kubeba uzito wa majukumu, mabadiliko, na ukuaji wa kiroho na kihisia. Hivyo, kulinganisha ndoa na Bembea ya Maisha inaonyesha jinsi mahusiano yanavyopitia vipindi tofauti na kuleta mabadiliko ya kina.
Human Expert	<ol style="list-style-type: none"> <li>1. Sara na Yona wanapooana, wanakaa muda mrefu bila kupata mtoto.</li> <li>2. Wanafanikiwa kuzaa tu watoto wa kike, ambapo ni kinyume na utamaduni .</li> <li>3. Kukosa mtoto wa kiume inamfanya Yona kuanza kubukia pombe na kuanza kumpiga mkewe hadi anamsabibishia maradhi.</li> <li>4. Pombe inamfanya Yona kupoteza kazi yake na kushindwa kushughulikia mahitaji ya familia yake .</li> <li>5. Watoto wake wasichana wanafanikiwa kusoma na baadaye kuanza kusaidia wazazi wao.</li> </ol>

➤ **QUESTION 6:** *Design a learning experience for a learning area in teaching of the topic 'loci' putting in consideration the magic triangle of learning.*

Response Agent	Response
ChatGPT	To teach the topic of "loci," I would design a hands-on learning experience incorporating the magic triangle of learning. Students will engage in a treasure hunt activity where they must follow clues leading to hidden treasures placed at specific loci. This activity promotes active exploration (Experience), as students physically move around to locate treasures. Guidance and feedback (Feedback) will be provided throughout the activity, helping students correct their paths if needed. Reflection (Reflection) occurs at the end, where students discuss strategies used and connections made between loci and geometric concepts. This immersive experience enhances understanding and retention of loci principles.
Perplexity AI	In teaching the topic of 'loci,' a geometry concept, the learning experience should incorporate the Magic Triangle of Learning: motivation, cognition, and metacognition. Begin with a real-world application of loci to engage students' motivation. Use interactive activities like drawing paths of moving objects to stimulate cognition. Encourage metacognition by having students reflect on their problem-solving strategies. Utilize visual aids, group work, and feedback loops to enhance understanding. Assess learning through practical applications and self-assessment. By integrating motivation, cognition, and metacognition, students can actively engage with the concept of loci, fostering a deeper and more meaningful learning experience.
Human Expert	<p><b>Action oriented</b></p> <p>The learner can be asked to make a round groove on a flat piece of wood and small nails hammered around at constant spacing. Secondly, the learner can be guided to make a bean ball/bag and affix pins all round it. <b>Smart visuals.</b> Use of bright clear visuals combined with text can enhance efficient learning of the concept loci. A bean ball/bag with tailor's pins.</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>A bean ball/bag with tailor's pins</p> </div> <div style="text-align: center;">  <p>Aboard with nails</p> </div> </div>



**Collect feedback:** Learners should understand that: the locus of a point at a given distance from a fixed point in two-dimension geometry is a circle whose radius is the distance between the two points. The locus of a point at a given distance from a fixed point in three-dimension geometry is a sphere whose radius is the distance between the two points.

The responses in the six subject areas were subsequently evaluated by an additional group of 24 experts. They assessed the comprehensibility, actionability, and scientific correctness of the answers provided by ChatGPT,

Perplexity AI, and human respondents using a scale of 1 to 10. This evaluation served as the foundation for the quantitative analysis of the study.

Table 3 Statistical Summary of the Key Findings

Sn	Question	Respondent	Mean	Median	Interquartile Range	Minimum	Maximum
1	1	ChatGPT	6.01	6.24	1.88	5.32	8.82
2	1	Perplexity AI	5.89	6.00	2.00	2.22	7.54
3	1	Hum. Expert	4.39	4.76	1.76	4.00	7.55
4	2	ChatGPT	7.09	7.26	2.21	6.02	9.99
5	2	Perplexity AI	7.26	7.18	1.88	3.22	9.10
6	2	Hum. Expert	7.41	6.88	2.00	5.77	10.00
7	3	ChatGPT	6.96	7.02	1.75	3.44	8.72
8	3	Perplexity AI	3.24	3.55	1.66	1.87	7.33
9	3	Hum. Expert	7.56	7.62	1.67	4.47	9.92
10	4	ChatGPT	5.66	6.00	2.23	4.77	8.71
11	4	Perplexity AI	4.62	5.33	2.22	1.77	6.96
12	4	Hum. Expert	7.21	7.56	2.00	3.33	10.00
13	5	ChatGPT	6.49	7.22	2.33	4.21	6.22
14	5	Perplexity AI	4.08	6.33	2.46	3.43	7.22
15	5	Hum. Expert	7.21	6.00	2.32	5.00	8.89
16	6	ChatGPT	6.56	6.23	2.14	2.34	7.23
17	6	Perplexity AI	6.23	6.23	1.76	3.45	8.89
18	6	Hum. Expert	8.00	7.32	1.76	5.77	10.00

#### ➤ Analysis of Responses

Table 4 Rating of the Responses from the Respondents

Question	Response	Comprehensibility Rate on Scale 0-10	Actionability Rate on Scale 0-10	Scientific Correctness Rate on Scale 0-10
Question 1	Response 1	6.48	5.15	6.64
	Response 2	6.22	6.12	5.12
	Response 3	4.22	6.62	2.33
Question 2	Response 1	7.14	7.12	7.00
	Response 2	7.23	7.23	7.33
	Response 3	8.18	2.33	6.64
Question 3	Response 1	7.62	7.12	7.02
	Response 2	4.87	3.00	4.00
	Response 3	6.77	7.33	7.73
Question 4	Response 1	5.52	6.33	5.12
	Response 2	4.87	5.00	4.00
	Response 3	6.62	7.33	7.68
Question 5	Response 1	8.47	5.34	5.66
	Response 2	3.92	4.00	4.33
	Response 3	6.62	7.33	7.68
Question 6	Response 1	7.47	6.48	5.88
	Response 2	6.92	6.45	5.33
	Response 3	7.21	7.68	8.00

Key: Response 1: ChatGPT, Response 2: Perplexity AI, Response 3: Human Expert.

Table 5 Summary of descriptive statistics for each respondent type across the three criteria.

Respondent	Mean Scientific Accuracy	SD Scientific Accuracy	Mean Actionability	SD Actionability	Mean Comprehensibility	SD Comprehensibility
ChatGPT	6.25	0.71	5.42	0.99	6.58	0.99
Perplexity AI	4.33	0.79	4.17	1.06	5.43	0.55
Human Expert	7.42	0.65	7.25	0.77	7.08	1.24

The initial descriptive statistics reveal notable differences in performance across respondent types. Human Experts achieved a mean score of 7.42 (SD = 0.65) for Scientific Accuracy, significantly outperforming both ChatGPT (mean = 6.25, SD = 0.71) and Perplexity AI (mean = 4.33, SD = 0.79). In terms of Actionability, Human Experts again led with a mean score of 7.25 (SD = 0.77), while ChatGPT and Perplexity AI recorded mean scores of 5.42 (SD = 0.99) and 4.17 (SD = 1.06), respectively. Comprehensibility scores mirrored these trends, with ChatGPT slightly ahead at 6.58 (SD = 0.99), contrasted with Human Experts at 7.08 (SD = 1.24) and Perplexity AI at 5.43 (SD = 0.55).

#### ➤ Normality and Variance Tests

Subsequent normality tests indicated that the distributions of scores for all three dimensions deviated significantly from a normal distribution ( $p < 0.05$ ). Additionally, Levene's Test for homogeneity of variances confirmed significant differences in variances across respondent groups ( $p < 0.05$ ), necessitating the use of non-parametric tests for further analysis.

#### ➤ Kruskal-Wallis Test Results

The Kruskal-Wallis test was employed to assess the differences in median scores among the three respondent categories. The results are summarized in Table 6.

Table 6 Kruskal-Wallis Test Results

Dimension	Test Statistic	p-value
Scientific Accuracy	18.12	< 0.001
Actionability	16.75	< 0.001
Comprehensibility	8.34	0.015

The Kruskal-Wallis test yielded significant results across all dimensions. For Scientific Accuracy, the test statistic was 18.12 ( $p < 0.001$ ), indicating substantial differences among groups. Actionability scores also demonstrated significant differences, with a test statistic of 16.75 ( $p < 0.001$ ). Comprehensibility scores similarly

revealed significant differences (test statistic = 8.34,  $p = 0.015$ ).

#### ➤ Post-Hoc Analysis

Following the significant results of the Kruskal-Wallis test, Dunn's test for pairwise comparisons was conducted. The results of Dunn's test are presented in Table 3.

Table 7 Dunn's Test for Pairwise Comparisons between ChatGPT, Perplexity AI and Human Expert

Comparison	p-value
ChatGPT vs. Perplexity AI	< 0.001
ChatGPT vs. Human Expert	< 0.001
Perplexity AI vs. Human Expert	< 0.001
ChatGPT vs. Perplexity AI (Actionability)	< 0.001
ChatGPT vs. Human Expert (Actionability)	< 0.001
Perplexity AI vs. Human Expert (Actionability)	< 0.001
ChatGPT vs. Perplexity AI (Comprehensibility)	0.047
ChatGPT vs. Human Expert (Comprehensibility)	0.089
Perplexity AI vs. Human Expert (Comprehensibility)	0.012

The pairwise comparisons revealed that Human Experts significantly outperformed both ChatGPT and Perplexity AI in both Scientific Accuracy and Actionability ( $p < 0.001$  for all comparisons). In terms of Comprehensibility, ChatGPT was found to be significantly more comprehensible than Perplexity AI ( $p = 0.047$ ); however, there was no significant difference between ChatGPT and Human Experts ( $p = 0.089$ ).

#### ➤ Effect Size

The analysis of effect sizes, calculated using eta-squared ( $\eta^2$ ), demonstrated substantial practical significance

for Scientific Accuracy ( $\eta^2 = 0.60$ ) and Actionability ( $\eta^2 = 0.58$ ), indicating a strong impact of respondent type on these dimensions. Comprehensibility showed a medium effect size ( $\eta^2 = 0.26$ ), suggesting noteworthy differences, though less pronounced than the other two dimensions.

#### ➤ Box Plot Charts on the Results using R Permutations

Using permutation tests in R, the following Boxplot charts were used to indicate the strength of responses with respect to means for Comprehensibility, Actionability and Scientific Correctness. Boxplots showing the grades for the individual components are displayed in Figures 1 through 3,

while Tables 2, 3, and 4 display tables with their summary data.

#### ➤ *Comprehensibility*

It was noted that ChatGPT registered higher mean scores in most of the questions compared to Perplexity AI and Human Experts as shown by Question 1 (6.48 for ChatGPT, 6.22 for Perplexity AI vs. 4.22 for the Human Experts), Question 3 (7.62 for ChatGPT, 4.87 for Perplexity AI vs. 6.77 for the Human Experts), Question 5 (8.47 for

ChatGPT, 3.92 for Perplexity AI vs. 6.62 for the Human Experts), Question 6 (7.47 for ChatGPT, 6.92 for Perplexity AI vs. 7.21 for the Human Experts). This suggests that the message communicated by ChatGPT is easy to comprehend compared to the message communicated by both humans and Perplexity AI. This is supported by Kirk et al (2023) who found that AI tools are better writers than human experts. However, this study showed that Human beings were scientifically correct and produced findings that more actionable.

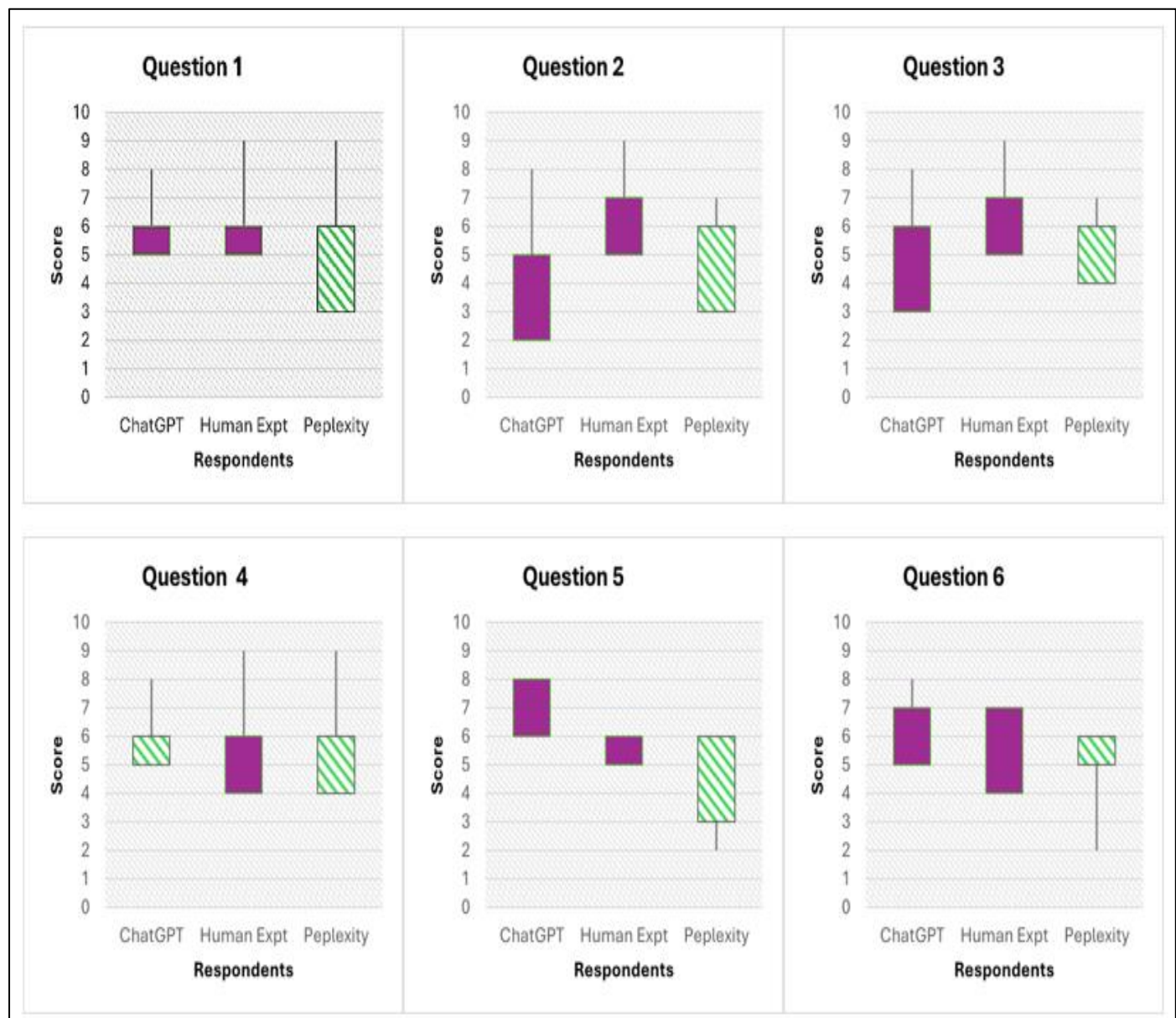


Fig 1 Overall grades on Comprehensibility of the responses given by the three respondents (ChatGPT, Human Expert and Perplexity AI)

#### ➤ *Actionability*

Higher mean scores were noted in all questions for Human Experts except for Question 2 in which Perplexity AI had higher mean score; (7.12 for ChatGPT, 7.23 for Perplexity AI vs. 2.33 for the Human Experts). It was also noted that in the second place after Human Experts was ChatGPT which registered higher mean scores in all the questions except for Question 2 compared to Perplexity AI.

The results are presented in Figure 2. This shows that human beings provide realistic findings that can be acted upon compared to the AI tools. This is supported by Liang (2023) and Brynjolfsson (2023). Kirk (2023) further indicates that lack of emotional intelligence and common sense will still give humans experts an upper edge in problem solving and decision making.



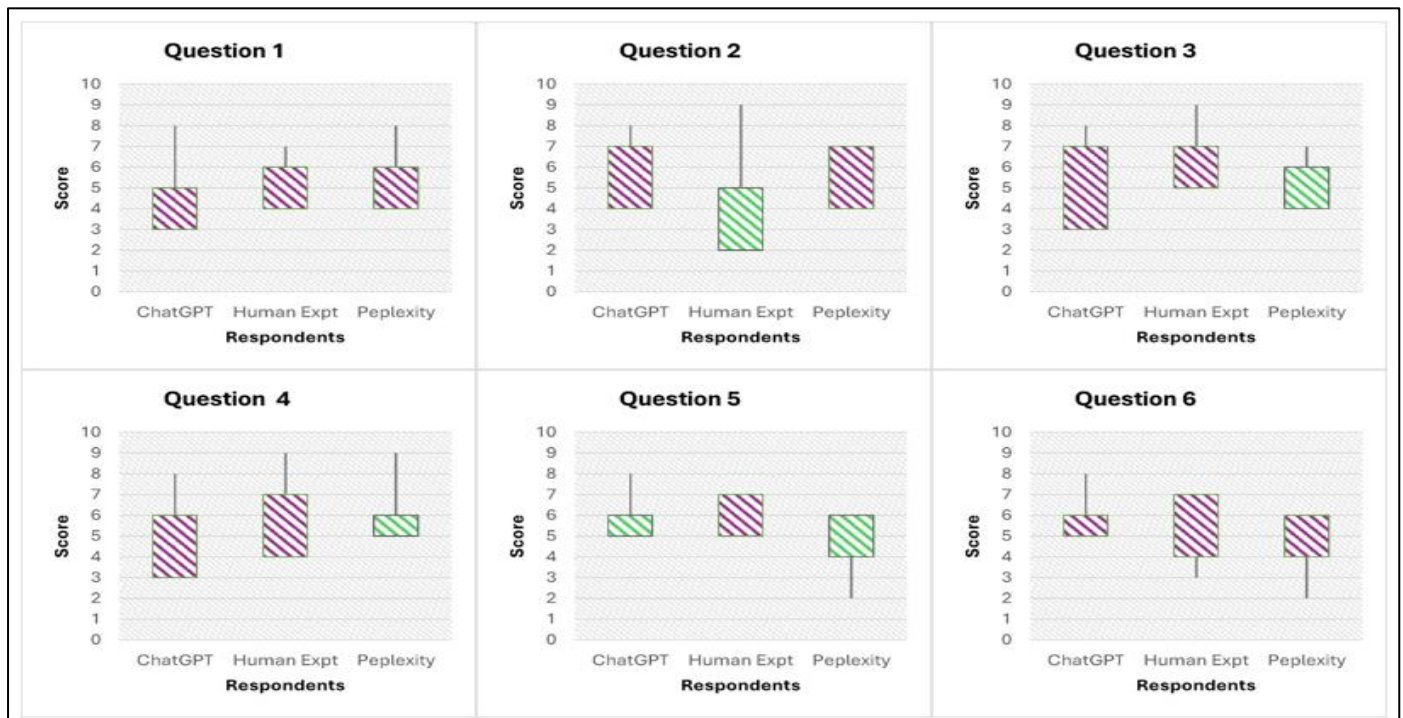


Fig 2 Overall Grades on Actionability of the responses given by the three respondents (ChatGPT, Human Expert and Perplexity AI)

#### ➤ Scientific Correctness

Question 3 (7.02 for ChatGPT, 4.00 for Perplexity AI vs. 7.72 for the Human Experts), Question 4 (5.12 for ChatGPT, 4.00 for Perplexity AI vs. 7.68 for the Human Experts), Question 5 (5.66 for ChatGPT, 4.33 for Perplexity AI vs. 7.68 for the Human Experts), and Question 6 (5.88 for ChatGPT, 5.33 for Perplexity AI vs. 8.00 for the Human Experts) had higher mean scores for scientific correctness for Human Experts. The other questions indicated that

ChatGPT and Perplexity AI had higher mean scores than the human experts. The results are presented in Figure 3. This confirms the fact that the AI models are still under training and may not provide more accurate scientific findings as confirmed by Fingerhut (2023). Kirk (2023) and Moazzam (2023) also confirms that over-reliance on training data for AI tools hinders them from making informed decisions based on other information beyond the scope of their training data.

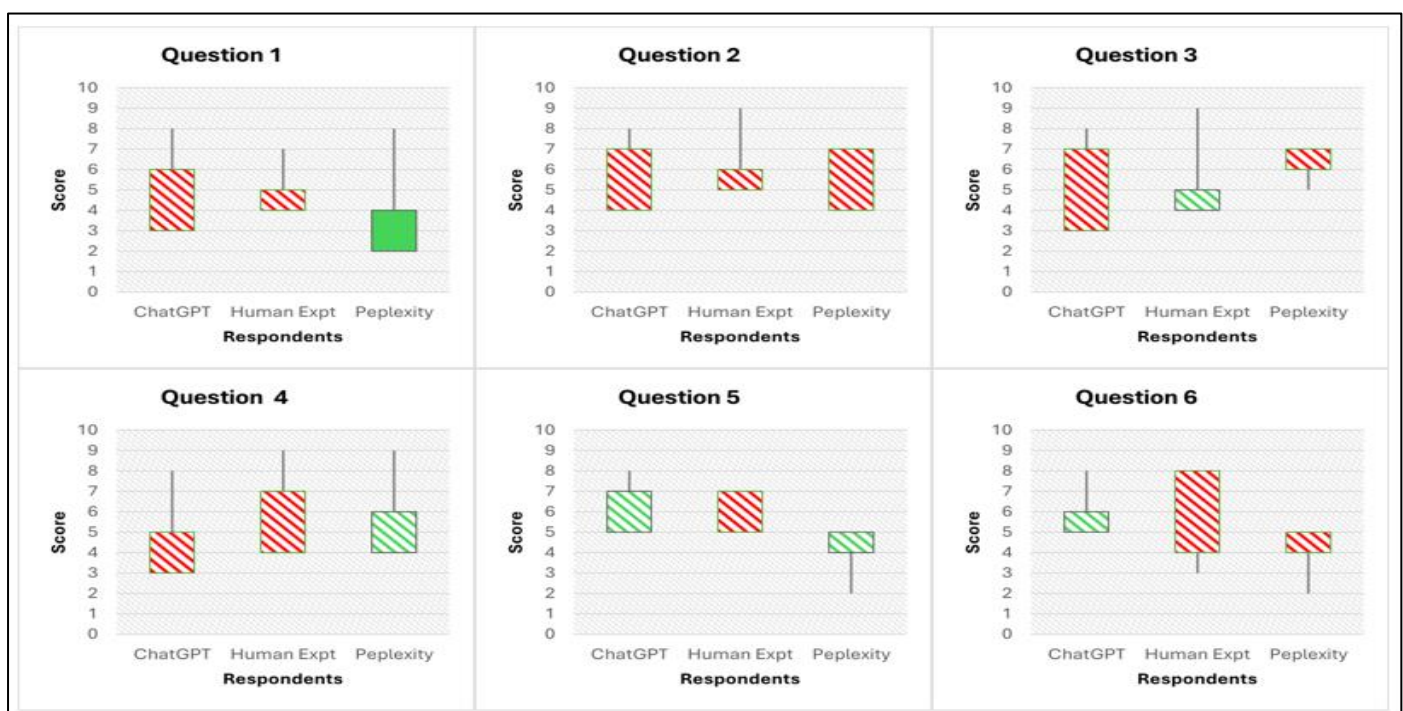


Fig 3 Overall grades on Scientific Correctness of the responses given by the three respondents (ChatGPT, Human Expert and Perplexity AI)



## V. CONCLUSIONS

Overall, the study discovered that ChatGPT and Perplexity AI are potential tools that can support human-centered learning and faster analysis of mathematical problems. For instance, they could have included AI-generated topics into our triangulation talks to help uncover omissions, other frames, and personal biases instead of performing a ChatGPT analysis, Perplexity AI analysis after finishing our human-generated analysis. Although the direction of AI development is uncertain, we believe that programs like ChatGPT will eventually replace Google Scholar, citation management, data analysis, and grammar-checking software as extra tools for research work. Despite demonstrating a degree of correctness, the following observations were made as pitfalls for AI tools compared to human responses:

- **Biases:** ChatGPT and Perplexity AI were learned using substantial text datasets which may have preconceptions and biases reflected in the model's results. For instance, if the training data uses language that is racist or sexist, the text produced by the model might support these prejudices.
- **Lack of Common-Sense Knowledge:** AI tools (ChatGPT and Perplexity AI) frequently respond to stimuli in an improper or nonsensical way because it lacks common sense knowledge. For instance, what if someone asked, "Can you fly to the moon without a spacesuit?" It is possible for them to produce a response that is technically accurate but devoid of common sense.
- **Over-reliance on Training Data:** These tools' output quality is mostly reliant on the caliber and applicability of the training data that was utilized to create the model. An incomplete, skewed, or unrepresentative set of training data may affect how accurate the models produce results.
- **Difficulty in Creating Long-Form Content:** Although they can produce cohesive and coherent short-form material, like a paragraph or a few phrases, they find it difficult to keep longer-form content, like essays or research reports, consistent and structured.
- **Limited Memory:** They can both only hold a specific quantity of data at once due to its limited memory. This implies that when writing text, it might not be able to keep constant allusions to earlier material or a constant voice throughout.
- **Incapacity to Generate Original Ideas:** They both may produce language that is stylistically and content-wise similar to the training data it was trained on, but it is unable to provide unique ideas or insights that are not covered by the training data.
- **Lack of Emotional Intelligence:** The AI tools are not emotionally intelligent enough to recognize or react to a user's emotions and sentiments in a way that is appropriate.

This may result in inappropriate or insensitive reactions in circumstances where empathy is needed.

## RECOMMENDATIONS

While ChatGPT and Perplexity AI have the potential to offer sufficient insights into providing detailed information from various academic spheres, they have also shown a notable rate of mistakes and a capability to present outdated information. Thus, in order to prevent the possibility of false information and inaccurate evidence, an accurate, expert-driven verification process is still required while using the AI tools.

## ACKNOWLEDGMENTS

We express our sincere gratitude to Dr. Muyobo for his expertise in Computer Science, Dr. Fr. Kizito for his insights on the role of religion in education, Dr. Kelly for her assistance in Mathematics, Ms. Sally Khama for her contributions to English Literature, Ms. Lucy Nakhabi for her support in Kiswahili, and Prof. Mukasa Simiyu for his guidance in Instructional Design. Their invaluable contributions and generous sharing of knowledge were instrumental in the successful completion of this study.

## REFERENCES

- [1]. **Aayush P. (2024).** Perplexity AI: Review, Advantages and Guide. Elegant Themes. Available online: <https://www.elegantthemes.com/blog/business/perplexity-ai> (accessed on 19th March 2024).
- [2]. **Altman, S. (2023).** Planning for AGI and Beyond. OpenAI Blog. Available online: <https://openai.com/blog/planning-for-agi-andbeyond> (accessed on 12 May 2024).
- [3]. **Brynjolfsson, E. A. (2023).** Call to Augment—Not Automate—Workers. In *Generative AI: Perspectives from Stanford HAI*. Stanford University, Human-Centered Artificial Intelligence, Palo Alto, CA, USA, pp. 16–17.
- [4]. **Cao J., et al. (2023).** Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. *AJR Am J Roentgenol*, 16:1–4. doi: 10.2214/ajr.23.29493.
- [5]. **Coskun B, et al. (2023).** Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urology* (2023) 4 (23)00570-8. doi: 10.1016/j.urology.2023.05.040.
- [6]. **Donelan, M. (2024).** Government Commits up to £3.5 Billion to Future of Tech and Science. UK Government News. Available online: <https://www.gov.uk/government/news/government-commits-up-to-35-billion-to-future-of-tech-and-science> (accessed on 12 February 2024).
- [7]. **Ellis S. (2024).** Is Perplexity AI Better than ChatGPT? A Comparison. Available online: <https://em360tech.com/tech-article/perplexity-ai-vs-chatgpt> (accessed on 19th March 2024).

- [9]. **Fingerhut A., Winter D. (2023).** Artificial intelligence and medical writing: where are we going? *Br J Surg*, znad169. doi: 10.1093/bjs/znad169.
- [10]. **Open Letter (2023).** Future of Life Institute Pause Giant AI Experiments: An Open Letter. Available online: <https://futureoflife.org/openletter/pause-giant-ai-experiments> (accessed on 15 May 2024).
- [11]. **Haque M.U. et al. (2023).** “I Think This Is the Most Disruptive Technology”: Exploring Sentiments of ChatGPT Early Adopters Using Twitter Data. *arXiv* 2022, arXiv:2212.05856.
- [12]. **Haver HL, et al. (2023).** Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*, 307(4). doi: 10.1148/radiol.230424.
- [13]. **Hinton G., et al. (2023).** Statement on AI Risk. Center for AI Safety. Available online: <https://www.safe.ai/statement-on-ai-risk> (accessed on 15 March 2024).
- [14]. **Johnson SB, et al. (2023).** Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr*, 7(2): pkad015. doi: 10.1093/jncics/pkad015.
- [15]. **Kirk, D., et al. (2023).** The Capabilities of Chat GPT’s at Answering Common Nutrition Questions. OSF, Charlottesville, VA, USA, 2023.
- [16]. **Kleinman, Z., et al. (2024).** AI Language Systems Are “Quite Stupid”. BBC News. Available online: <https://www.bbc.com/news/technology-66238004> (accessed on 12 March 2024).
- [17]. **Kleinman, Z., et al. (2023).** Why Making AI Safe Isn’t as Easy as You Might Think. BBC News. Available online: <https://www.bbc.com/news/technology-65850668> (accessed on 12 March 2024).
- [18]. **Li, F.-F., et al. (2023).** *Generative AI: Perspectives from Stanford HAI*. Stanford University, Human-Centred Artificial Intelligence, Palo Alto, CA, USA, 2023.
- [19]. **Liang, P.** The New Cambrian Era: “Scientific Excitement, Anxiety”. In *Generative AI: Perspectives from Stanford HAI*.
- [20]. **Lund, B.D., et al. (2023).** ChatGPT and a New Academic Reality: AI-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing. *J. Assoc. Inf. Sci. Technol.*, 74, 570–581.
- [21]. **Madiega, T. (2021).** Artificial Intelligence Act; European Parliament: Strasbourg, France, 2023. Available online: [https://www.europarl.europa.eu/RegData/etudes/BRI E/2021/698792/EPRS\\_BRI698792\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRI E/2021/698792/EPRS_BRI698792_EN.pdf) (accessed on 12 May 2023).
- [22]. **Microsoft (2023).** Will AI Fix Work? 2023 Work Trend Index: Annual Report. Available online: [https://assets.ctfassets.net/WTI\\_Will\\_AI\\_Fix\\_Work\\_060723.pdf](https://assets.ctfassets.net/WTI_Will_AI_Fix_Work_060723.pdf) (accessed on 12 April 2024).
- [23]. **Moazzam Z, et al. (2023).** Quality of ChatGPT responses to questions related to pancreatic cancer and its surgical care. *Ann Surg Oncol*, 22. doi: 10.1245/s10434-023-13777-w.
- [24]. **R Core Team (2022).** R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- [25]. **Vallance, C. (2024).** AI Could Replace Equivalent of 300 Million Jobs—Report. BBC News. Available online: <https://www.bbc.com/news/technology-65102150> (accessed on 12 May 2024).
- [26]. **White House FACT SHEET (2023).** Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. Available online: <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/> (accessed on 12 May 2024).
- [27]. **Yeo Y., et al. (2023).** Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*, 29(3):721–732. doi: 10.3350/cmh.2023.0089.
- [28]. **Zhu L, et al. (2023).** Can ChatGPT and other large language models with internet-connected databases solve the questions and concerns of patients with prostate cancer and help democratize medical knowledge? *J Transl Med*, 21(1):269. doi: 10.1186/s12967-023-04123-5.
- [29].

**APPENDICES****➤ Appendix 1. R code for the analysis and permutations**

```
# Loading the necessary libraries
library(dplyr)
# Create the dataset
data <- data.frame(
  Question = rep(1:6, each = 3),
  Respondent = rep(c("ChatGPT", "Perplexity AI", "Human Expert"), times = 6),
  Scientific_Accuracy = c(6.5, 4.5, 7.5, 7.0, 5.5, 8.0, 6.0, 3.0, 8.5, 5.0, 4.0, 7.0, 6.5, 4.5, 8.0, 6.0, 5.0, 8.5),
  Actionability = c(5.0, 4.0, 6.5, 7.0, 6.0, 7.5, 5.5, 3.5, 8.0, 4.0, 3.0, 6.0, 5.5, 3.5, 7.0, 6.0, 5.0, 8.0),
  Comprehensibility = c(6.5, 6.0, 4.5, 7.5, 5.5, 7.0, 8.0, 4.5, 7.5, 6.5, 5.0, 7.0, 6.0, 5.0, 8.0, 7.0, 5.5, 7.5)
)

# Descriptive statistics
summary_stats <- data %>%
  group_by(Respondent) %>%
  summarise(
    Mean_Scientific_Accuracy = mean(Scientific_Accuracy),
    SD_Scientific_Accuracy = sd(Scientific_Accuracy),
    Mean_Actionability = mean(Actionability),
    SD_Actionability = sd(Actionability),
    Mean_Comprehensibility = mean(Comprehensibility),
    SD_Comprehensibility = sd(Comprehensibility)
  )
```