AI | MDPI

Article

Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard

Vagelis Plevris [1], George Papazafeiropoulos [2] and Alejandro Jiménez Rios [3,*]

[1] Department of Civil and Environmental Engineering, Qatar University, Doha P.O. Box 2713, Qatar; vplevris@qu.edu.qa
[2] School of Civil Engineering, National Technical University of Athens, 15780 Athens, Greece; gpapazafeiropoulos@yahoo.gr
[3] Department of Built Environment, Oslo Metropolitan University, 0166 Oslo, Norway
* Correspondence: alejand@oslomet.no

Abstract: In an age where artificial intelligence is reshaping the landscape of education and problem solving, our study unveils the secrets behind three digital wizards, ChatGPT-3.5, ChatGPT-4, and Google Bard, as they engage in a thrilling showdown of mathematical and logical prowess. We assess the ability of the chatbots to understand the given problem, employ appropriate algorithms or methods to solve it, and generate coherent responses with correct answers. We conducted our study using a set of 30 questions. These questions were carefully crafted to be clear, unambiguous, and fully described using plain text only. Each question has a unique and well-defined correct answer. The questions were divided into two sets of 15: Set A consists of "Original" problems that cannot be found online, while Set B includes "Published" problems that are readily available online, often with their solutions. Each question was presented to each chatbot three times in May 2023. We recorded and analyzed their responses, highlighting their strengths and weaknesses. Our findings indicate that chatbots can provide accurate solutions for straightforward arithmetic, algebraic expressions, and basic logic puzzles, although they may not be consistently accurate in every attempt. However, for more complex mathematical problems or advanced logic tasks, the chatbots' answers, although they appear convincing, may not be reliable. Furthermore, consistency is a concern as chatbots often provide conflicting answers when presented with the same question multiple times. To evaluate and compare the performance of the three chatbots, we conducted a quantitative analysis by scoring their final answers based on correctness. Our results show that ChatGPT-4 performs better than ChatGPT-3.5 in both sets of questions. Bard ranks third in the original questions of Set A, trailing behind the other two chatbots. However, Bard achieves the best performance, taking first place in the published questions of Set B. This is likely due to Bard's direct access to the internet, unlike the ChatGPT chatbots, which, due to their designs, do not have external communication capabilities.

Keywords: chatbot; AI; logic; mathematics; ChatGPT; GPT-3.5; GPT-4; Google Bard

check for updates

Citation: Plevris, V.; Papazafeiropoulos, G.; Jiménez Rios, A. Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI* **2023**, *4*, 949–969. https://doi.org/10.3390/ai4040048

Academic Editors: Kenji Suzuki, Antony Bryant, Roberto Montemanni, Min Chen, Paolo Bellavista and Jeanine Treffers-Daller

Received: 4 September 2023
Revised: 12 October 2023
Accepted: 16 October 2023
Published: 24 October 2023

1. Introduction and Purpose of the Study

In the realm of early chatbot development, a significant milestone was reached with the inception of ELIZA [1], an innovative software program crafted by Joseph Weizenbaum in the 1960s. ELIZA stands as a prominent figure in the domain of natural language processing and artificial intelligence, and it garnered widespread acclaim for its capacity to emulate conversations between a psychotherapist and a patient, thus representing one of the initial manifestations of AI-driven human–computer interactions [2]. ELIZA's pioneering methodology in replicating conversational dialogues laid the cornerstone for subsequent advancements in chatbot technology and thereby set the stage for the comprehensive evaluation and comparison of contemporary chatbot systems, a pursuit that our study diligently undertakes.

A chatbot is a computer program or artificial intelligence (AI) system which is meticulously engineered to partake in dialogues with human users through textual or voice-based exchanges [3]. Chatbots find applicability across a spectrum of functions, including, but not limited to, furnishing customer support, addressing inquiries, proffering recommendations, or even facilitating informal discourse [4]. They are typically designed to mimic human-like conversation patterns and can be found on messaging platforms, websites, and mobile applications. They are AI-driven programs designed to engage in natural language conversations with users. They often use natural language processing (NLP) and machine learning (ML) algorithms to understand and respond to user inputs more effectively.

In November 2022, the landscape of chatbot technology underwent a transformative evolution with the introduction of ChatGPT by OpenAI. This milestone marked a significant departure from prior models, owing to its proficiency in generating responses that exhibit enhanced human likeness and contextual coherence. ChatGPT excels in engaging in prolonged and contextually intricate conversations, overcoming the limitations inherent in earlier chatbots, which were characterized by succinct and contextually constrained responses. Users are empowered to specify their desired style or tone of generated content through prompts, affording them heightened control over the chatbot's output. With an expanded model size and an extensive corpus of training data, ChatGPT boasts an augmented lexicon and knowledge repository, enabling it to address a broader spectrum of topics and queries. These collective advancements firmly position ChatGPT as a more adept, versatile, and context-aware chatbot that can adeptly meet the evolving expectations and demands of users in natural language interactions. The chatbot garnered considerable attention for its comprehensive responses and articulate elucidations across diverse domains of knowledge; it achieved a notable milestone by attracting one million users within just five days of its launch, thus establishing a world record. In January 2023, it reached over 100 million users, making it the fastest growing consumer application to date [5,6].

Given their revolutionary nature, chatbots have numerous uses in many different areas, including the potential to accelerate and enhance scientific research and boost technological development [7]. They can be put to the test in various fields to evaluate their understanding and problem-solving abilities, and they can even be tested against actual professional examinations [8]. In the present study, we put three chatbots, (i) ChatGPT-3.5, (ii) ChatGPT-4, and (iii) Google Bard (or simply Bard), to the test with math and logic problems to determine their capacity to:

1. **Understand the problem**: Chatbots must be able to accurately interpret the user's input, recognize the type of problem being posed, and identify the relevant information needed to solve it.
2. **Apply appropriate algorithms or methods**: Chatbots need to utilize appropriate problem-solving techniques or mathematical algorithms to tackle the given problem. This may involve arithmetic, algebra, calculus, or logical reasoning, depending on the complexity of the problem.
3. **Generate a coherent response and the correct answer**: Once a solution is derived, chatbots should present the answer in a clear and concise manner, making sure the response is relevant and easy for the user to understand. Also, the final answer to the question should be mathematically correct.

In the scholarly discourse, a multitude of inquiries and surveys emerged mere months after the debut of the ChatGPT and Bard chatbots, aimed at evaluating their mathematical proficiencies and their capacity to provide substantive assistance to professional mathematicians [9]. These assessments subjected ChatGPT to a battery of tests encompassing a spectrum of logical reasoning domains. These tests spanned from the chatbot's adeptness in addressing computational inquiries, such as intricate integral calculations, to its ability to furnish solutions to mathematical proofs featuring gaps or missing steps. Furthermore, examinations extended to the chatbot's competence in resolving mathematical challenges culled from Olympiad-level problems, as well as its aptitude in cross-domain reasoning—exemplified by the ability to identify the requisite theorems for substantiating a given theorem [9]. The assessment of ChatGPT's mathematical capabilities was conducted utilizing a novel dataset introduced within the same research study [9]. The outcomes of these evaluations demonstrated that ChatGPT's proficiency in mathematics falls significantly below the level of an average graduate student specializing in the field. Another investigation revealed a stark contrast in ChatGPT's performance, which was contingent upon whether it was tasked with providing explanations or additional contextual information alongside its answers, in comparison to scenarios where it was instructed to furnish answers in isolation, devoid of any supplementary text [10]. In the former scenario, its probability of failure was estimated at 20%, while in the latter, it surged to 84%. The evaluation of ChatGPT's performance in this study was conducted employing the DRAW-1K dataset [11–13], which was meticulously arranged to encompass 1000 algebraic word problems. These problems were semi-automatically annotated to facilitate the assessment of automated solvers, and they encompassed not only the problems and their corresponding solutions but also the template algebraic equations necessary for solving such word problems [10].

An additional endeavor aimed at mitigating the inherent limitations of ChatGPT in solving intricate mathematical conundrums was undertaken and is expounded upon in [14]. This research delves into the systematic deficiencies of ChatGPT in addressing complex open-domain queries. In this context, the study classifies ChatGPT's shortcomings into four distinct categories, namely comprehension, factualness, specificity, and inference. Furthermore, the study identifies three pivotal competencies implicated in the occurrence of quality assurance lapses, namely knowledge memorization, knowledge association, and knowledge-based reasoning. The findings presented in [14] culminate in the inference that augmenting the model with refined external knowledge, providing cues for knowledge association, and offering guidance for reasoning can potentiate the model's capacity to furnish more accurate responses to inquiries.

Despite the ever-expanding body of research dedicated to the examination and evaluation of ChatGPT in several domains [7,8,15], the authors are not aware of comparable investigations being conducted for the other prominent chatbot systems, such as Bard for instance. Furthermore, there is a dearth of studies that undertake a comparative analysis of different chatbots, particularly studies that concern the accuracy of their responses. The present study endeavors to address this gap by conducting a performance assessment of diverse chatbots based on their responses to a meticulously crafted set of logically formulated and mathematically oriented questions. The objective of these evaluations is to facilitate a comprehensive appraisal of the capabilities and limitations of various chatbot systems, offering valuable insights for researchers, developers, and end users alike. Such assessments shed light on the domains in which chatbots excel and the areas where enhancements are required. Ultimately, these endeavors contribute to the continuous evolution of more advanced, proficient, and user-centric AI chatbot systems.

The research questions that the present study attempts to address are the following:

- **Performance Evaluation**: How do ChatGPT-3.5, ChatGPT-4, and Google Bard perform in solving a diverse range of math and logic problems, and what understanding, problem-solving ability, and accuracy do they have?
- **Comparative Analysis**: What are the key differences in the performance of these chatbot systems in addressing different types of math and logic problems, and how are these differences manifested?
- **Contextual Performance**: How does the performance of the chatbots vary based on the complexity and nature of the problems, and what role does the context of the problem play in their capabilities?
- **Comparison Across Generations**: How does ChatGPT-4 compare to its predecessor, ChatGPT-3.5, in terms of improvements in accuracy, problem solving, and contextual understanding?

- **External Resources Impact**: What impact does Google Bard's access to the internet and Google search have on its performance, particularly when addressing problems with publicly available solutions online?
- **User Experience**: How do users perceive these chatbots in an educational context, and what insights can be drawn regarding user preferences and trust in chatbot-generated responses?

The full dataset of the study, which includes the full set of the 30 questions, together with the correct answer for each one of them, an explanation of the solution, and the responses of the chatbots, can be found in [16]. A preprint of this paper was previously published [17]. The present version of the paper is more thorough and complete, offering some additional analysis and updated information about the chatbots and their latest improvements since the preprint version. The remainder of the paper is organized as follows: Section 2 discusses the three chatbots used in the study, the technologies behind them, and the differences between the three models. Section 3 presents the methodology of the study; this is followed by the discussion of the individual answers to each question in Section 4. Section 5 presents and discusses the performance of the three chatbots in the 30 questions. Finally, Section 6 presents the conclusions of this work, a relevant discussion, and some future directions.

2. Chatbots Used in the Study

In this study, we used three chatbots: (i) ChatGPT-3.5, (ii) ChatGPT-4, and (iii) Google Bard. All these chatbots rely on a large language model (LLM), which is a type of AI model designed to understand and generate human-like text by leveraging deep learning (DL) techniques. These models are "large" in the sense that they have a massive number of parameters, often ranging from hundreds of millions to hundreds of billions, which allows them to capture complex language patterns and relationships.

2.1. ChatGPT-3.5 and ChatGPT-4

GPT, short for Generative Pre-trained Transformer, is a state-of-the-art language model developed by OpenAI. It comes in two main versions today, GPT-3.5 and GPT-4. On the other hand, ChatGPT is a chatbot app, powered by GPT, which is optimized for dialogue. Consequently, ChatGPT refers to a chatbot which is powered by GPT (any version), ChatGPT-3.5 refers to the chatbot which is powered by GPT-3.5, and, similarly, ChatGPT-4 refers to the chatbot which is powered by GPT-4.

ChatGPT is designed to generate human-like text responses based on the input it receives. At a high level, ChatGPT works by using DL techniques to understand and generate text. It is trained on a massive amount of data from the internet and other sources; this allows it to learn the patterns and structures of language. During training, the model predicts the next word in a sentence given the previous words, and this process is repeated many times in order for it to learn the relationships between words and the overall context of the text. When a user interacts with ChatGPT, they provide it with a prompt or a question. The model then analyzes the input and generates a response based on its understanding of the given text and the patterns it has learned. The response is generated by sampling from a probability distribution over the vocabulary of possible words; the context and the likelihood of each word in the given context are considered.

The training process for ChatGPT involves training on a vast amount of text data, but it is important to note that the model does not have real-time access to the internet or current events. As of today (September 2023), the knowledge and information available to ChatGPT are based on the data it was trained on; these data had a cutoff date in September 2021 [18]. As a result, the model may not be aware of recent developments or be able to provide real-time information.

In scientific terms, ChatGPT operates through the following key components and processes:

1. **Transformer architecture**: The transformer model, proposed by Vaswani et al. in 2017 [19], is the backbone of ChatGPT. It uses self-attention mechanisms to process input data in parallel rather than sequentially, allowing it to efficiently handle long-range dependencies on the text.
2. **Pre-training**: ChatGPT is pre-trained on a large corpus of text from the internet. During this unsupervised learning phase, it learns the structure and statistical properties of the language, including grammar, vocabulary, and common phrases. The objective is to predict the next word in a sentence, given the context of the preceding words.
3. **Fine-tuning**: After the pre-training phase, ChatGPT is fine-tuned on a smaller dataset, often containing specific conversational data. This supervised learning phase involves training the model to generate appropriate responses in a conversational setting. The model learns from human-generated input–output pairs and refines its ability to provide contextually relevant and coherent responses.
4. **Tokenization**: When ChatGPT receives input text, it tokenizes the text into smaller units, such as words or subwords. These tokens are then mapped to unique IDs, which are used as input for the model.
5. **Encoding and decoding**: The transformer architecture consists of an encoder and a decoder. The encoder processes the input tokens, while the decoder generates the output tokens sequentially. Both the encoder and decoder rely on self-attention mechanisms and feed-forward neural networks to process and generate text.
6. **Attention mechanisms**: Attention mechanisms enable the model to weigh the importance of different parts of the input when generating a response. This helps ChatGPT to focus on the most relevant information in the input text and generate coherent and contextually appropriate responses.
7. **Probability distribution**: The model's output is a probability distribution over the vocabulary for the next token in the sequence. The token with the highest probability is chosen, and this process is repeated until the model generates a complete response or reaches a predefined maximum length.
8. **Beam search or other decoding strategies**: To generate the most likely response, ChatGPT uses decoding strategies like beam search, which maintains a set of top-*k* candidate sequences at each time step. These strategies help in finding a balance between fluency and coherence while minimizing the risk of generating nonsensical or overly verbose outputs.

By combining these components and techniques, ChatGPT can understand and generate human-like text, making it a powerful tool for various applications, such as conversational agents, content generation, and question-answering systems. ChatGPT can work with various languages, to some extent. While it is primarily trained on English text data, it also learns from multilingual text sources during its pre-training phase. As a result, it can understand and generate text in several languages, such as Spanish, French, German, Chinese, Arabic [20], and more. However, it is important to note that ChatGPT's proficiency in different languages may vary depending on the amount and quality of training data available for each language. Its performance is typically better for languages with a larger presence on the internet and in the training data. Naturally, it exhibits the best performance in the English language.

It is important to note that while ChatGPT demonstrates impressive coherence in generating responses, it can occasionally produce incorrect or nonsensical answers. Essentially, it is a statistical model that relies on data patterns and likely lacks true understanding or human-like knowledge. The term "hallucination" [21] describes the phenomenon where AI systems generate outputs that are unrealistic, incorrect, or nonsensical. Although these outputs may resemble human creativity or imagination, they are solely generated by the AI algorithm. The term "hallucination" is used metaphorically to emphasize that the AI is producing content that may lack a direct basis in reality or accurate data. Therefore, it is crucial to critically evaluate and fact-check information provided by ChatGPT or any similar language model. OpenAI is actively refining and improving language models such as ChatGPT, and future iterations may address some of the limitations and challenges associated with the current generation of models.

GPT-3.5 vs. GPT-4

GPT-4 is an advanced version of its predecessor, GPT-3.5. It was released on 14 March 2023. Some of the improvements and advanced features of GPT-4 include [18]:

1. **Increased model size**: GPT-3 has 175 billion parameters that allow it to take an input and give a text output that best matches the user request. GPT-4 has far more parameters, although the exact number is not known, leading to improved language understanding and generation capabilities. OpenAI has not given information about the size of the GPT-4 model [22].
2. **Enhanced context understanding**: GPT-4 can handle longer text inputs and maintain context better, which allows more coherent and relevant responses.
3. **Improved fine-tuning**: GPT-4 has been fine-tuned on more diverse and specific tasks, allowing it to perform better across a wider range of applications.
4. **Better handling of ambiguity**: GPT-4 is better at resolving ambiguous input and providing clearer, more accurate responses.
5. **More robust language support**: GPT-4 has been trained on a broader range of languages and can better handle multilingual tasks and code switching.
6. **Enhanced safety and ethical considerations**: GPT-4 has been designed with more robust safety measures to prevent harmful outputs, ensuring better alignment with human values.
7. **Domain-specific knowledge**: GPT-4 has been trained on more specific knowledge domains, allowing it to provide more accurate information and support specialized tasks.

These are general features and improvements over the previous version. The actual performance of GPT-4 may still vary depending on the specific task, input, or context. ChatGPT-3.5 is free for all users and does not have any limitations in its use, while ChatGPT-4 currently has a cap of 50 messages (previously 25) every 3 h, and it is provided as a paid service (ChatGPT Plus), with a 20 USD/month subscription.

2.2. Google Bard

Google Bard is an LLM chatbot developed by Google AI. It is trained on a massive dataset of text and code, including Wikipedia, Books, Code, Stack Overflow, Google Search, and other publicly available datasets. It can generate text, translate languages, write different kinds of creative content, and answer user questions in an informative way. Bard is free to use. It was first released for US and UK users on 21 March 2023, and it was initially available only in English. In July 2023, it became available in more languages and places. As of September 2023, it is available in more than 40 languages, including Arabic, Chinese, German, Hindi, and Spanish, and in over 230 countries and territories around the world. According to Google, Bard is powered by several technologies, including:

- **Natural language processing (NLP)**: NLP is a field of computer science that deals with the interaction between computers and human (natural) languages. NLP is used in Bard to understand and process the text that the user inputs.
- **Machine learning (ML)**: ML is a field of computer science that gives computers the ability to learn without being explicitly programmed. ML is used in Bard to train the model on the massive dataset of text and code.
- **Deep learning (DL)**: DL is a subset of ML that uses artificial neural networks to learn from data. DL is used in Bard to train the model to generate text, translate languages, write different kinds of creative content, and answer user questions in an informative way.

As of September 2023, Bard is still officially under development, but it has learned to perform many kinds of tasks, such as: (i) Following user instructions and completing user requests thoughtfully; (ii) using its knowledge to answer questions in a comprehensive and informative way, even if they are open ended, challenging, or strange; and (iii) generating different creative text formats of text content, like poems, code, scripts, musical pieces, email, letters, etc. Bard has direct access to the internet and its data are constantly being updated; so, it is always learning new things.

2.3. Differences between ChatGPT and Bard

A key difference between ChatGPT and Bard is that Bard has access to Google search, and it continually draws data from the internet; so, it has the latest information. On the other hand, ChatGPT does not have direct access to the internet or any specific data source. Its knowledge is based on the vast amount of text data that it was trained on, which includes web pages, books, articles, and other textual sources. The training data of ChatGPT were last updated in September 2021. As a result, ChatGPT does not have knowledge of the latest events and developments in any scientific fields or any other fields.

3. Methodology of the Study

In this study, we use 30 questions describing mathematics and logic problems that have a unique correct answer. These questions are fully described with plain text only, without the need for any images or special text formatting. They cover various categories of logical and mathematical problems, including arithmetic, algebraic expressions, basic logic puzzles, and complex mathematical problems. They were selected based on several criteria such as:

- **Clarity and Unambiguity**: All the questions were designed to be clear and unambiguous to ensure that the chatbots could comprehend the problem statements accurately.
- **Diversity**: We aimed to include a diverse set of problems to evaluate the chatbots across various mathematical and logical domains.
- **Availability**: The questions were divided into two sets, with Set A consisting of 15 "Original" problems that were not readily available online and Set B consisting of 15 "Published" problems that could be found online, often with solutions. This division allowed us to assess the chatbots' ability to handle both novel and publicly available problems.
- **Well-Defined Correct Answers**: Each question had a unique and well-defined correct answer, which made it possible to objectively evaluate the chatbots' responses for correctness.

Each question is posed three times to each chatbot. For ChatGPT-3.5 and ChatGPT-4, we had to click the button "Regenerate response" two times to receive three answers. With Bard, things were simpler as it automatically provides three "draft" answers. The first answer is displayed to the user, while to see the other two, one needs to click the "View other drafts" button. The full set of the 30 questions, together with the correct answer for each one of them, an explanation of the solution, and the $30 \times 3 \times 3 = 270$ detailed answers of the chatbots can be found in the relevant published dataset [16]. The structure of the dataset is the following: The problems of Set A are presented first, followed by the ones of Set B. First, the problem is stated, together with its correct answer and an explanation of it, where needed. Then, the responses of the chatbots are presented, starting with the three responses of ChatGPT-3.5 and followed by the responses of ChatGPT-4 and Bard. Each response is marked as "Correct" (highlighted with green color) or "NOT Correct" (highlighted with red color) in the dataset manuscript [16].

A hypothesis that needs to be tested is that the chatbots that rely primarily on online "ready-to-use" information and online search may be better in answering the questions of Set B, but they will have problems with those of Set A. It must be noted that we tried to avoid ambiguities and to make the problems as clear as possible. Therefore, we do not try to check the abilities of the chatbots in handling ambiguous or ill-posed problems. We also do not engage in any kind of dialogue with the chatbot for any question, and we do not allow it to ask clarifying questions. To keep things simple and fair, we do not provide any feedback to any answer given by a chatbot. Thus, we do not check the ability of a chatbot to learn from user feedback. In general, chatbots should be able to learn from the user's feedback and to improve their problem-solving abilities over time, but this is not an aim of the present study. The focus of the study is on checking the ability of chatbots to (i) understand the problem at hand, (ii) apply appropriate algorithms or methods, and (iii) generate a coherent response and the correct answer. All the questions were posed to the chatbots in May 2023. Since then, there may have been developments and improvements in the way that chatbots reply to these or similar prompts.

4. Discussion of the Individual Answers to Each Question

In this section, we provide each full question, its correct answer, the score of each chatbot, and some discussion on the responses given. Each question was posed three times, and as a result, the chatbots gave three answers for each question. A score of *k-l-m* means *k* correct answers for ChatGPT-3.5, *l* correct answers for ChatGPT-4, and *m* correct answers for Bard in their three attempts. We keep this order for the three chatbots throughout the manuscript.

4.1. Set A: "Original" Questions

This is the set of 15 "Original" questions, denoted as A01 to A15, which cannot be found online and have not been published previously, at least not with the same wording.

Questions A01 and A02 (Scores: 3-1-1 and 0-0-0)

> *A01. "Solve the following cubic equation: x^3 - 13*x^2 + 50*x - 56 = 0";*
>
> *A02. "Solve the following cubic equation: 100*x^3 - 1340*x^2 + 5389*x - 6660 = 0".*

The correct answer to question A01 is "**x = 2, x = 4, x = 7**". The correct answer to question A02 is "**x = 2.5, x = 3.7, x = 7.2**". Both questions are of a similar nature: the numerical values of x that satisfy the given cubic equation must be found. All three chatbots seem to understand the nature of the problem. In terms of the methods or algorithms used to solve the problem, ChatGPT-3.5 implements the rational roots theorem five out of six times and Cardano's formula once. ChatGPT-4 attempts to provide a solution by using the rational roots theorem, a graphical solution, and a code snippet in python 66.7%, 16.7%, and 16.7% of the time, respectively. Finally, Bard uses factor lists five times and the rational roots theorem once. All the implemented methods or algorithms can correctly lead to a right answer; thus, it could be said that the chatbots have chosen a proper way to give an answer.

The first problem (A01) has three integer roots. ChatGPT-3.5 had the best performance in this relatively easy task. ChatGPT-3.5 managed to give three correct solutions, while ChatGPT-4 failed in the first attempt, got it right in the second, and failed again in the third. Bard found the correct solution in the first attempt but missed the other two. It is impressive that although these models are so complex and can find solutions to difficult problems, they can fail at such an easy task. In addition, it must be highlighted that normal reasoning appears not to work for them. After solving this exercise, a human would easily check the solution by substituting the found values for x in the equation to see if the equation was satisfied. This is not the case with AI chatbots. They did not bother to check their final solution.

Problem A02 has three roots which are not integers; so, this task is a bit harder than the previous one. All the models failed to give a correct solution in all their attempts at this question.

Question A03 (Score: 1-1-0)

> *"A closed club of professional engineers has 500 members. Some members are "old members" while the others are "new members" (subscribed within one year from now). An event was organized where old members had to pay $200 each for their participation while new members had to pay $140 each. The event was successful and while all new*

> *members came, only 70% of the old members attended. What is the amount of money (in $) that was collected from all members for this event?"*

The correct answer is "**70,000**". This is a relatively "wordy" question where the chatbots need to understand not only the numbers given, but also the relationship created with the words between those numbers. The question is only asked at the end. As all the chatbots tried to provide an amount of money as an answer, it could be said that they correctly understood the question. Although all three attempted to provide a solution by performing some mathematical calculations, it seems that only ChatGPT chatbots have the capacity to correctly define "unknowns" and assign given data to variables, which could be considered as the "correct" methodology. On the other hand, Bard seems to work with the numbers given without assigning these numbers an actual problem-context meaning.

This question was very challenging for all the chatbots. Only ChatGPT-3.5 and ChatGPT-4 got it correct once. Bard failed in all its attempts. The challenge with this question is that we do not really know the number of the old members and the number of new members. Nevertheless, these are not asked by the problem. The question simply asks the amount of money collected, which does not require knowing the exact number of old and new members. This was the tricky part of this question, and it caused problems for the chatbots.

Question A04 (Score: 1-3-0)

> *"The sum of three adults' ages is 60. The oldest of them is 6 years older than the youngest. What is the age of each one of them? Assume that an adult is at least 18 years old."*

The correct answer is "**18, 18, 24**". Again, ChatGPT-3.5 got it correct once, ChatGPT-4 got it correct all three times as it clearly understood the problem. Bard still failed in all the attempts it made. This is a straightforward mathematical problem but with some constraints included. All three chatbots attempted to compute the age of the three adults, which somehow shows their understanding of the question asked. Furthermore, they tried to use algebra and could assign numerical values to the variables (a quality which Bard did not show in the previous question A03) to come up with a solution; thus, they implemented a "proper" methodology.

Question A05 (Score: 3-3-1)

> *"A decade ago, the population of a city was 55,182 people. Now, it is 170% larger. What is the city's current population?"*

The correct answer is "**148,991 people**". Interestingly, only ChatGPT seemed to fully understand the problem, whereas Bard did not even correctly identify the city's current population in two out of three attempts. On the other hand, the three chatbots seemed to implement appropriate methods or algorithms to give an answer to the question. Both ChatGPT models got this right in all their attempts. Strangely, Bard failed two out of three times, even though the problem seemed clear and simple. In these two attempts, Bard failed to understand that 170% larger means 270% of the original value (not 170% of it).

Question A06 (Score: 1-3-0)

> *"What is the precise sum of 523,654,123 and 7,652,432,852,136?"*

The correct answer is "**7,652,956,506,259**". In this case, all three chatbots seemed to understand the problem, which is a simple mathematical addition. Strangely, ChatGPT-3.5 fails to provide a correct result in two out of three attempts, and Bard failed in all three attempts. ChatGPT-4 got it correct all three times. It is strange that two chatbots failed in this rather simple calculation. Table 1 shows the responses of ChatGPT-3.5 and Bard in this question. These models failed to predict one or two digits of the result, while the other digits were correct, which is also strange.

Table 1. Responses of ChatGPT-3.5 and Bard in the addition question A06.

Correct Calculation and Result	**523,654,123** **+7,652,432,852,136**
	7,652,956,506,259
ChatGPT-3.5 Attempt #1	✔ **7,652,956,506,259**
ChatGPT-3.5 Attempt #2	✗ 7,65**3**,956,506,259
ChatGPT-3.5 Attempt #3	✗ 7,65**3**,956,506,259
Bard Attempt #1	✗ 7,652,956,5**15**,259
Bard Attempt #2	✗ 7,652,956,50**5**,259
Bard Attempt #3	✗ 7,652,956,50**5**,259

Green color indicates a correct answer. The red color highlights the mistake in the answer.

Question A07 (Score: 2-3-3)

> *"You decide to make a road-trip with your new car. The distance between City A and City B is 120 km. When you travel from A to B, your average speed is slow, 60 km/h. When you travel from B to A, your average speed is high, 120 km/h. What is the average speed for the whole trip A to B to A (with return to City A)?"*

The correct answer is "**80 km/h**". Many people can get confused thinking that the answer is 90 km/h, which is the average of 60 km/h and 120 km/h, but this is not a correct approach. This question presents one of the best chatbot performances so far (88.9% correct responses). All three chatbots seemed to understand the problem correctly and to implement appropriate methods/algorithms (algebra-based) to come up with an answer. The chatbots succeeded in providing the correct answer in all their attempts, except for ChatGPT-3.5, which failed once. Similar questions can be found online, but not with the exact wording of this problem.

Question A08 (Score: 1-3-3)

> *"If Tom has 35 marbles and I have 12 marbles, and then Tom gives me 9 marbles, how many more marbles does Tom have than I?"*

The correct answer is "**5**". After a careful examination of the performance of ChatGPT-3.5, it could be observed that it systematically failed in two out of three attempts. The reason for this is that the chatbot failed to "understand" that in this context "giving" also means "losing" (although this issue was not present in the last attempt of the chatbot). Although this appears to be a very easy question, ChatGPT-3.5 failed in two of its three attempts, while the other two chatbots got it correct all three times.

Question A09 (Score: 3-3-3)

> *"Tom's father has three children. The younger child's name is Erica. The middle child's name is Sam. What is the name of the older child?"*

The correct answer is "**Tom**". All the chatbots managed to correctly interpret the problem, implement a suitable method/algorithm to find a solution, and give a correct answer to this simple question in all three of their attempts, thus showing a remarkably high performance when dealing with purely logical problems.

Question A10 (Score: 3-3-1)

> *"A woodworker normally makes a certain number of parts in 11 days. He was able to increase his productivity by 3 parts per day, and so he not only finished the job 2 days earlier, but in addition he made 9 extra parts. How many parts does the woodworker normally make per day?"*

The correct answer is "**9**". Both versions of ChatGPT managed to give correct answers to this question in all three of their attempts. Bard was correct only in its first attempt and failed in the other two attempts. On this occasion, all the chatbots seemed to understand the problematic, which is related to the computation of the number of parts the woodworker originally produced per day. Moreover, all three use variables and apply basic algebra operations to find the solution, which could indeed be considered an appropriate methodology.

Question A11 (Score: 2-3-0)

> *"Think of a number. Add 5, double the result, then subtract 12, then take half of the result and finally subtract the initial number. What is the result?"*

Here the correct answer is "**−1**". In this question, ChatGPT-3.5 got two out of three, GPT-4 got them all correct, while Bard failed three times. This is a simple, yet interesting logico-mathematical problem that requires one to follow a step-by-step process and apply basic operations (+, −. *, /) in each one to find the correct solution. All three chatbots seemed to correctly understand the problematic. Furthermore, both ChatGPT chatbots applied a correct multistep algorithm to find a solution. On the other hand, Bard tried to pose the problem as an equation, sometimes even generating a greater number of variables than required.

Question A12 (Score: 3-2-3)

> *"If one and a half hens lay one and a half eggs in one and a half days, how many eggs do 9 hens lay in 9 days?"*

The correct answer is "**54**". In this question, ChatGPT-3.5 and Bard got it correct three times, while GPT-4 failed once and got it correct two times. This is a relatively simple question in mathematical terms, but it is posed in a tricky way in terms of language, as the "*one and a half*" string is repeated several times, which may pose a challenge to a chatbot. Although all three chatbots seemed to correctly "understand" the problematic and apply an adequate methodology to solve it, on one occasion ChatGPT-4 seemed to get "confused" in the process. It adequately determined that one hen laid one egg per day, but then it assigned the original number mentioned in the question (1.5), and the result turned out to be erroneous. Variations of this problem can be found online, but not with this exact wording.

Question A13 (Score: 0-2-0)

> *"Find a 4-digit number so that the last four digits of the number squared is the number itself."*

The correct answer is "**9376**". Variations of this problem can be found online, but not with this exact wording. In this question, ChatGPT-3.5 and Bard failed three times, while GPT-4 got it correct two times. The one time it failed, it gave 0625 as an answer, stating that "Note that 0625 is technically a 4-digit number, although it may appear as a 3-digit number (625) in some representations due to the leading zero." This shows that GPT-4 understood the problem and that the solution it provided might not have been what was expected. It tried to defend its answer with some reasoning, which is interesting. It must be noted that in its second attempt, ChatGPT-4 stated that "such a number is called a "Kaprekar number"", which is not correct as a definition. Nevertheless, ChatGPT-4 ended up with the correct answer at the end.

Question A14 (Score: 0-3-0)

> *"The number of water lilies on a lake doubles every two days. If there is initially one water lily on the lake, it takes exactly 50 days for the lake to be fully covered with water lilies. In how many days will the lake be fully covered with water lilies, if initially there were two water lilies (identical with the previous one) on it?"*

The correct answer is "**48**". Variations of this problem can be found online, but not with this exact wording. In most similar problems, it is stated that the number doubles every day. Both ChatGPT-3.5 and Bard failed three times in this problem. GPT-4 got it correct three times. However, by the answer given by all three chatbots, it could be said that all of them correctly understood the question (they all "acknowledge" the fact that the number of lilies doubles every second day); only ChatGPT-4 seemed to apply the correct "logic" or algorithm to come up with correct answers. It is also interesting to note how Bard may be completely biased in this specific problem due to the large number of online appearances of this type of problem when it is stated that the number of lilies would double every day. In such a scenario, the correct answer would be 49, which was the answer given by Bard in all three attempts; in other words, it seemed to give the "right" answer to the "wrong" question.

Question A15 (Score: 1-3-3)

> *"There are 25 handball teams playing in a knockout competition (i.e., if you lose a match, you are eliminated and do not continue further). What is the minimum number of matches (in total) they need to play to decide the winner?"*

The correct answer is "**24**". Variations of this problem can be found online, but not with this exact wording. The concept of such a tournament is one that can be found broadly on the internet. ChatGPT-4 and Bard got it correct three times, while ChatGPT-3.5 failed two times.

4.2. Set B: "Published" Questions

This is a set of 15 "Published" questions (problems), denoted as B01 to B15, which can be found online. The first nine questions were taken from [23]. Questions 10 and 15 are from [24], while questions 11 and 12 are from [25]. Finally, question 13 comes from [26] and question 14 is taken from [27].

Question B01 [23] (Score: 2-3-2)

> *"A bad guy is playing Russian roulette with a six-shooter revolver. He puts in one bullet, spins the chambers and fires at you, but no bullet comes out. He gives you the choice of whether or not he should spin the chambers again before firing a second time. Should he spin again?"*

The correct answer is "**Yes**". ChatGPT-3.5 got it right two times in three attempts, GPT-4 got it right in all three attempts, while Bard missed the first but got it right in the other two. From a purely mathematic point of view, this is a simple problem of probabilities, i.e., whether the outcome was desirable or not would determine whether a higher or a lower probability percentage was selected as the answer. Nevertheless, it could be argued that the wording used turns the question into an ethical one, by describing a person as "bad", with the corresponding assumptions that would imply. This could explain the fact that not all the answers were correct as this ethical dilemma may have confused the chatbots.

Question B02 [23] (Score: 2-3-3)

> *"Five people were eating apples, A finished before B, but behind C. D finished before E, but behind B. What was the finishing order?"*

The correct answer is "**CABDE**". This is an easy problem. Both GPT-4 and Bard give correct answers in all their attempts, while ChatGPT-3.5 missed it once. All three chatbots seemed to "understand" the problematic correctly as they all attempted to put the given letters in order. Furthermore, the method applied by the chatbots also seemed to be correct as the correct order of the letters can be known based on the information provided by the different statements of the question; this was the information used by the chatbots when attempting to provide a right answer.

Question B03 [23] (Score: 0-0-1)

> *"A man has 53 socks in his drawer: 21 identical blue, 15 identical black and 17 identical red. The lights are out, and he is completely in the dark. How many socks must he take out to make 100 percent certain he has at least one pair of black socks?"*

The correct answer is "**40**". This problem is not so easy to solve, although once given the solution everybody can understand that it makes sense. In this question, both the GPT models failed in all their attempts. Strangely, ChatGPT-3.5 gave answers which were more reasonable and closer to being correct. Bard got it right once, in its first attempt. Regardless of the low performance of the chatbots (an overall success rate of only 11.1%), they all seemed to have some understanding of the problem. Both ChatGPT-3.5 and Bard seemed to apply an adequate algorithm, which consisted of discarding all the socks of a different color to the one of interest (although they failed to compute the right number), whereas ChatGPT-4 applied the wrong logic to find the answer.

Question B04 [23] (Score: 1-1-3)

> *"Susan and Lisa decided to play tennis against each other. They bet $1 on each game they played. Susan won three bets and Lisa won $5. How many games did they play?"*

The correct answer is "**11**". This is a relatively easy problem, which nevertheless caused trouble for the chatbots. Both ChatGPT-3.5 and GPT-4 failed two times and predicted the outcome correctly once, while Bard got it correct in all three attempts. The three chatbots appear to have correctly understood the situation. Both ChatGPT chatbots implemented a similar solving methodology based on solving a couple of equations algebraically (although not always successfully), whereas Bard provided the correct answer in three attempts apparently purely out of self-logic "reasoning", which may be a hint that the chatbot found the solution to the problem posted online in the referenced source or elsewhere.

Question B05 [23] (Score: 3-3-3)

> *"Jack is looking at Anne. Anne is looking at George. Jack is married, George is not, and we don't know if Anne is married. Is a married person looking at an unmarried person?"*

The correct answer is "**Yes**". This is a relatively easy problem which caused no trouble for the chatbots. All of them got it correct, three times. All the chatbots "understood" the problem correctly and applied adequate methods/algorithms to solve it. Once again, the high performance of all three chatbots was ascertained while facing purely logical problems (as was the case with question A09).

Question B06 [23] (Score: 0-2-3)

> *"A girl meets a lion and unicorn in the forest. The lion lies every Monday, Tuesday and Wednesday and the other days he speaks the truth. The unicorn lies on Thursdays, Fridays and Saturdays, and the other days of the week he speaks the truth. "Yesterday I was lying," the lion told the girl. "So was I," said the unicorn. What day is it?"*

The correct answer is "**Thursday**". In this question, ChatGPT-3.5 failed three times. GPT-4 got it correct two times and failed once, while Bard got it right in all its attempts. The three chatbots seemed to "understand" the problematic in all the attempts; they "know" that they must give a day of the week as an answer to the problem. Moreover, they used the statements of the problem to "reason" and to come up with an answer, which is similar to what a human would do. Once again, this is a purely logical problem, but contrary to the cases of questions A09 and B05, the accuracy performance of the chatbots was considerably poorer.

Question B07 [23] (Score: 0-3-1)

> *"Three men are lined up behind each other. The tallest man is in the back and can see the heads of the two in front of him; the middle man can see the one man in front of him; the man in front can't see anyone. They are blindfolded and hats are placed on their heads, picked from three black hats and two white hats. The extra two hats are hidden and the blindfolds removed. The tallest man is asked if he knows what color hat he's wearing; he doesn't. The middle man is asked if he knows; he doesn't. But the man in front, who can't see anyone, says he knows. How does he know, and what color hat is he wearing?"*

The correct answer is "**Black**". In this question, ChatGPT-3.5 failed three times and GPT-4 got it right three times. Bard got it right only in the third attempt. This is again a purely logical problem, although expressed in a relatively long text question, which may increase the level of difficulty for an LLM. All three chatbots seemed to understand the problematic correctly and applied logical "reasoning" to come up with a solution. Nevertheless, their performance was far from satisfactory, except for ChatGPT-4.

Question B08 [23] (Score: 1-1-3)

> *"A teacher writes six words on a board: "cat dog has max dim tag." She gives three students, Albert, Bernard and Cheryl each a piece of paper with one letter from one of the words. Then she asks, "Albert, do you know the word?" Albert immediately replies yes. She asks, "Bernard, do you know the word?" He thinks for a moment and replies yes. Then she asks Cheryl the same question. She thinks and then replies yes. What is the word?"*

The correct answer is "**Dog**". In this question, only Bard got it correct three times, while the other two chatbots failed two times and got it correct only once. The complete reasoning of ChatGPT-3.5 in its second attempt did not appear to be 100% correct. Nevertheless, it came up with the correct answer in the end, and for this reason, we consider the answer to be finally "Correct" in this case.

Question B09 [23] (Score: 0-0-3)

> *"There are three people (Alex, Ben and Cody), one of whom is a knight, one a knave, and one a spy. The knight always tells the truth, the knave always lies, and the spy can either lie or tell the truth. Alex says: "Cody is a knave." Ben says: "Alex is a knight." Cody says: "I am the spy." Who is the knight, who the knave, and who the spy?"*

The correct answer is "**Alex: knight, Ben: spy, Cody: knave**". In this question, only Bard got it correct three times, while the other two chatbots failed in all three of their attempts. All three chatbots understood that this was a word puzzle and implemented an adequate reasoning strategy based on the statements of the question in order to attempt to provide the correct answer. Interestingly, none of the ChatGPT chatbots got any attempt correct, whereas Bard showed 100% accuracy. This may be because Bard was able to locate the right answer in the online source.

Question B10 [24] (Score: 0-0-3)

> *"Kenny, Abby, and Ned got together for a round-robin pickleball tournament, where, as usual, the winner stays on after each game to play the person who sat out that game. At the end of their pickleball afternoon, Abby is exhausted, having played the last seven straight games. Kenny, who is less winded, tallies up the games played: Kenny played eight games. Abby played 12 games. Ned played 14 games. Who won the fourth game against whom?"*

The correct answer is "**Ned beat Kenny in the fourth game**". In this question, only Bard got it correct three times, while the other two chatbots failed in all three of their attempts.

Questions B11 [25] and B12 [25] (Scores: 3-3-2 and 2-3-2)

> *B11. "The distance between two towns is 380 km. At the same moment, a passenger car and a truck start moving towards each other from different towns. They meet 4 h later. If the car drives 5 km/h faster than the truck, what are their speeds?"*
>
> *B12. "A biker covered half the distance between two towns in 2 h 30 min. After that he increased his speed by 2 km/h. He covered the second half of the distance in 2 h 20 min. Find the distance between the two towns."*

The correct answer to question B11 is "**Truck's speed: 45 km/h, Car's speed: 50 km/hr**". In this relatively easy question, ChatGPT-3.5 and GPT-4 got it correct in all of their attempts, while Bard made a mistake in its second attempt and got it correct in the other two. The problematic posed by the question is correctly identified by all three chatbots, which can be seen by the fact that they all tried to give speeds as correct answers to the problem. Furthermore, they all implemented a correct methodology to find a suitable solution, based on assigning the unknown speeds of the vehicles to variables and using algebraic operations to come up with the answer.

The correct answer to B12 is "**140 km**". In this easy question, only GPT-4 managed to give three correct answers. The other two models made one mistake and got it correct in the other two attempts.

Question B13 [26] (Score: 2-3-1)

> *"Rhonda has 12 marbles more than Douglas. Douglas has 6 marbles more than Bertha. Rhonda has twice as many marbles as Bertha has. How many marbles does Douglas have?"*

The correct answer is "**24**". Although this appears to be a very easy question, only GPT-4 managed to give three correct answers. This is a relatively simple mathematical question where the solution can be found through assigning variables to the unknowns and solving a relatively simple system of equations, which all three chatbots seemed to understand in all their attempts. Nevertheless, ChatGPT-3.5 got it correct two times and failed once, while Bard, strangely, failed in two attempts and got it correct only once.

Question B14 [27] (Score: 3-3-3)

> *"15 workers are needed to build a wall in 12 days. How long would it take to 10 workers to build the same wall?"*

The correct answer is "**18 days**". This question posed a relatively easy mathematical problem where all the chatbots managed to give correct answers in all three of their attempts. For all three chatbots, the perfect scoring reflects a good problematic understanding and an adequate implementation of a method/algorithm to come up with an answer.

Question B15 [24] (Score: 0-0-3)

> *"A hen and a half lay an egg and a half in a day and a half. How many eggs does one hen lay in one day?"*

The correct answer is "**⅔ of an egg**". Although this problem is similar to the previous one (B14), only Bard got it correct three times in this case. Strangely, the other two models failed in all their attempts. This trend in the performance of the chatbots turns out to be quite interesting. It could be said that all three chatbots understood the problem and correctly attempted to provide the number of eggs as an answer. For the ChatGPT chatbots, we are given the impression that the relatively "complex wording" of the question (the "and a half" string and the words "hen", "egg", and "day" are repeated several times in a quite short string) may have "confused" the chatbots, thus causing their failure to provide the correct answer. On the other hand, Bard's perfect score may be again due to the fact that the answer can be easily found online. This question is similar to question A12 from Set A, where the score was 3-2-3 for the three chatbots and Bard again got it correct three times. It is interesting that the ChatGPT chatbots had a much better performance in question A12 in comparison to this question, question B15, despite the similar nature of both questions.

5. Performance of the Chatbots

Due to space limitations, the detailed responses of the chatbots in all the problems are not included in this paper, but they can be found in the published open access full dataset [16]. Table 2 presents the scores of each chatbot in the first set of questions (Set A, 15 "Original" problems), the scores for each question, and the relevant sums. Each chatbot receives 1 point for a correct answer and 0 points for an incorrect answer. For illustrative purposes, the correct answers in the table are highlighted with a green color, while the incorrect ones are highlighted with a red color.

Table 2. Responses of the three chatbots to the questions of Set A (“Original” problems) *.

	ChatGPT-3.5				**ChatGPT-4**				**Bard**			
Question	**#1**	**#2**	**#3**	**SUM**	**#1**	**#2**	**#3**	**SUM**	**#1**	**#2**	**#3**	**SUM**
A01	1	1	1	3	0	1	0	1	1	0	0	1
A02	0	0	0	0	0	0	0	0	0	0	0	0
A03	0	1	0	1	0	1	0	1	0	0	0	0
A04	1	0	0	1	1	1	1	3	0	0	0	0
A05	1	1	1	3	1	1	1	3	0	1	0	1
A06	1	0	0	1	1	1	1	3	0	0	0	0
A07	1	0	1	2	1	1	1	3	1	1	1	3
A08	0	0	1	1	1	1	1	3	1	1	1	3
A09	1	1	1	3	1	1	1	3	1	1	1	3
A10	1	1	1	3	1	1	1	3	1	0	0	1
A11	0	1	1	2	1	1	1	3	0	0	0	0
A12	1	1	1	3	1	0	1	2	1	1	1	3
A13	0	0	0	0	0	1	1	2	0	0	0	0
A14	0	0	0	0	1	1	1	3	0	0	0	0
A15	0	1	0	1	1	1	1	3	1	1	1	3
SUM	8	8	8	**24**	11	13	12	**36**	7	6	5	**18**
Percentage	53.3%	53.3%	53.3%	**53.3%**	73.3%	86.7%	80.0%	**80.0%**	46.7%	40.0%	33.3%	**40.0%**

* 1 (green color) means “Correct”, 0 (red color) means “Not correct”. Bold shows the overall performance of each chatbot (sum of correct answers in 45 attempts and the relevant percentage).

In the Set A questions, ChatGPT-4 ranked first by providing 36 correct answers out of 45 attempts, resulting in an 80.0% success rate. ChatGPT-3.5 followed in second place, with 24 correct answers (53.3%), and Bard came in third with 18 correct answers (40.0%). Only in one question, A09, did all the LLMs answer correctly in all three attempts, accounting for 6.7% of the questions. ChatGPT-4 achieved an “all correct” score (three out of three attempts) in 10 out of 15 questions (66.7%), while both ChatGPT-3.5 and Bard achieved this score in 5 out of 15 questions (33.3%). Only by seeing the colors of the table, where green indicates a correct answer and red indicates an incorrect answer, can one understand that ChatGPT-4 was quite successful in most of the problems. Question A02 was the only question where all the LLMs scored zero, i.e., they did not manage to give a correct answer in any attempt. On the other hand, question A09 was the only question where all the LLMs scored 3, giving correct answers in all of their attempts.

In a similar manner, Table 3 presents the scores of each chatbot in the second set of questions (Set B, 15 “Published” problems), the scores for each question, and the relevant sums. In Set B, Bard came first, giving 36 correct answers out of 45 attempts (80% success rate), while GPT-4 came second, managing to give 28 correct answers (62.2%), and ChatGPT-3.5 gave 19 correct answers (42.2%). The success rate of Bard was impressive in these problems. Only in two questions (13.3%), B05 and B14, were all the models correct in all three of their attempts, while Bard got “all correct” (three out of three attempts) in 9 out of 15 questions (60%). Similarly, ChatGPT-3.5 got “all correct” in 3 questions (20%) and GPT-4 got “all correct” in 8 out of 15 questions (53.3%).

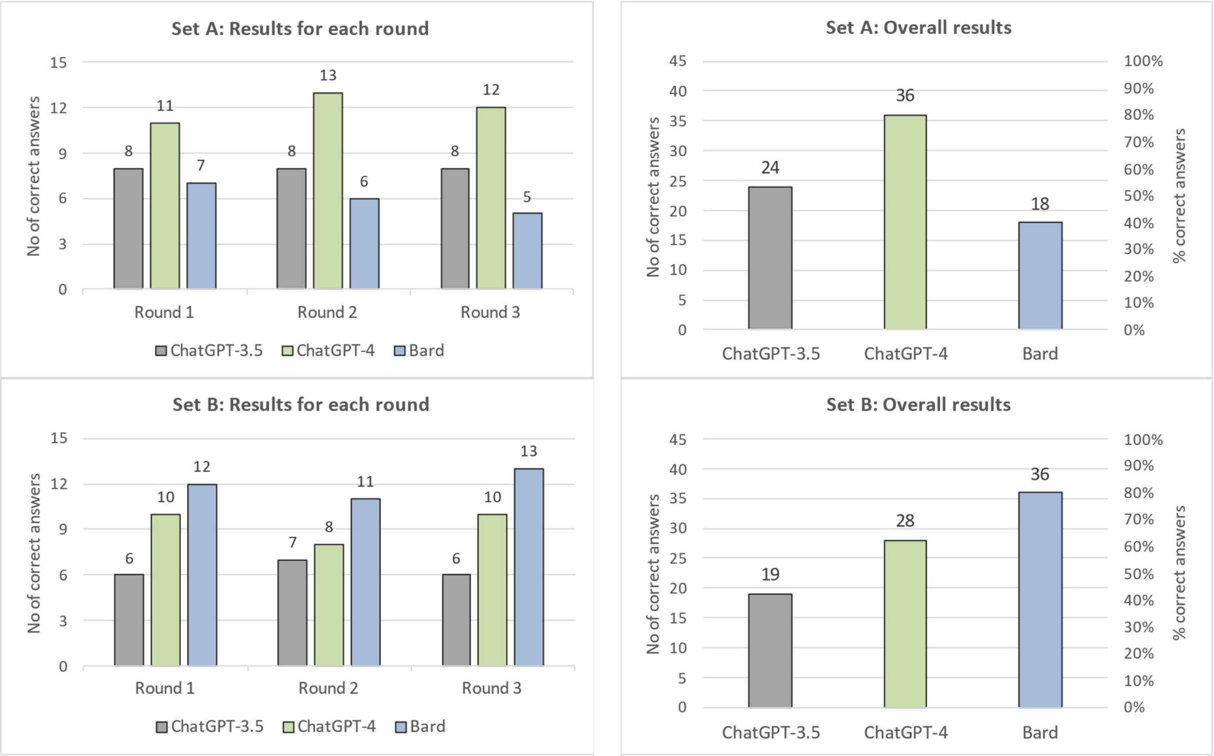
Figure 1 presents an illustration of the same results as the number of correct answers each chatbot gave for every set, in each of the three rounds (left column—(a)), and overall (right column—(b)). Comparing the performance of the chatbots in the questions of Set B with Set A, we see that ChatGPT-3.5 fell from 24 correct answers in Set A to 19 correct answers in Set B (20.8% decrease), and similarly, the performance of ChatGPT-4 fell from 35 correct answers to 28 (20% decrease), which shows a consistency and that the problems of Set B were probably harder than the ones of Set A. Impressively, the performance of Bard went up from 18 correct answers for the “Original” Set A questions to 36 for the

"Published" Set B questions, which is a remarkable improvement, despite the probable increased difficulty of these problems.

Table 3. Responses of the three chatbots to the questions of Set B ("Published" problems) *.

	ChatGPT-3.5				ChatGPT-4				Bard			
Question	**#1**	**#2**	**#3**	**SUM**	**#1**	**#2**	**#3**	**SUM**	**#1**	**#2**	**#3**	**SUM**
B01	1	1	0	2	1	1	1	3	0	1	1	2
B02	1	0	1	2	1	1	1	3	1	1	1	3
B03	0	0	0	0	0	0	0	0	1	0	0	1
B04	0	0	1	1	1	0	0	1	1	1	1	3
B05	1	1	1	3	1	1	1	3	1	1	1	3
B06	0	0	0	0	1	0	1	2	1	1	1	3
B07	0	0	0	0	1	1	1	3	0	0	1	1
B08	0	1	0	1	0	0	1	1	1	1	1	3
B09	0	0	0	0	0	0	0	0	1	1	1	3
B10	0	0	0	0	0	0	0	0	1	1	1	3
B11	1	1	1	3	1	1	1	3	1	0	1	2
B12	0	1	1	2	1	1	1	3	1	1	0	2
B13	1	1	0	2	1	1	1	3	0	0	1	1
B14	1	1	1	3	1	1	1	3	1	1	1	3
B15	0	0	0	0	0	0	0	0	1	1	1	3
SUM	6	7	6	**19**	10	8	10	**28**	12	11	13	**36**
Percentage	40.0%	46.7%	40.0%	**42.2%**	66.7%	53.3%	66.7%	**62.2%**	80.0%	73.3%	86.7%	**80.0%**

* 1 (green color) means "Correct", 0 (red color) means "Not correct". Bold shows the overall performance of each chatbot (sum of correct answers in 45 attempts and the relevant percentage).

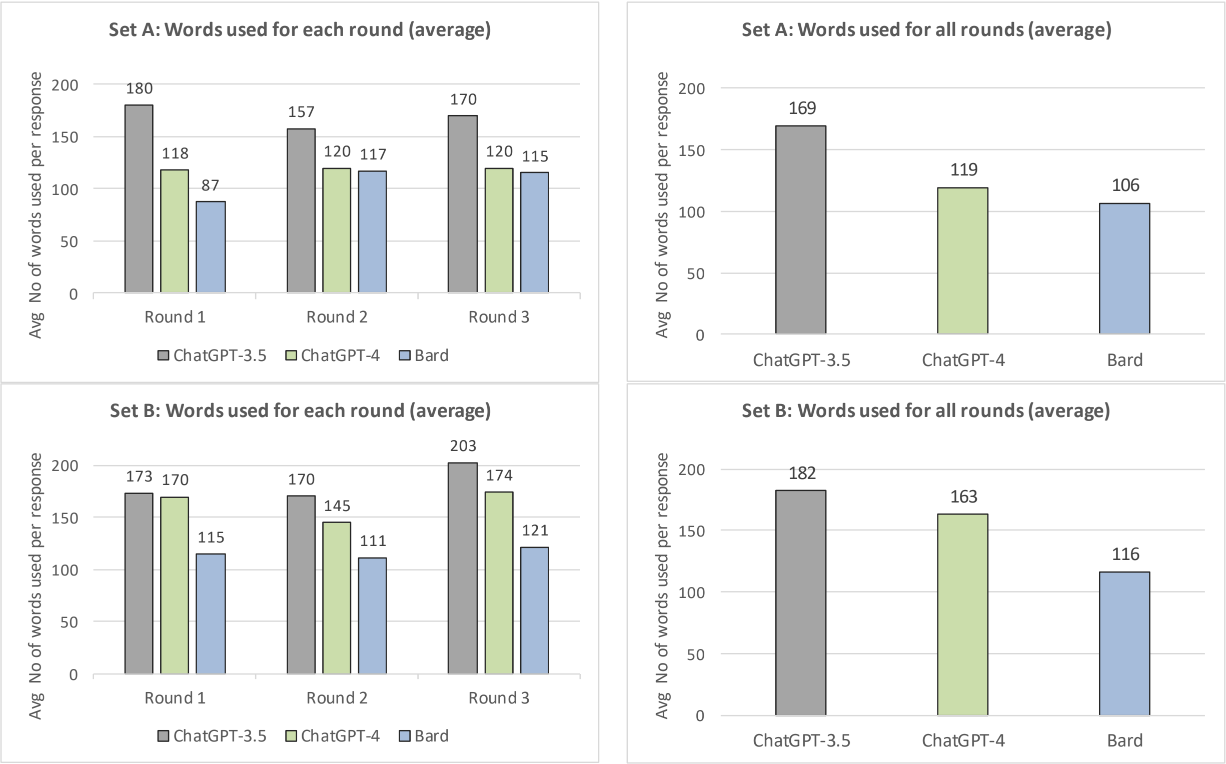


(**a**) (**b**)

Figure 1. Performance of each chatbot for the two sets of questions: (**a**) results for each of the three rounds, (**b**) overall results (all rounds).

It is obvious that Bard is much better at handling questions that have been already published and answered online than original questions that have not yet been published. The same is not true for the other two models, which showed a worse performance in the "Published" questions in comparison to the "Original" questions. This is probably because (i) the published questions were in fact harder than the original questions and because (ii) the two GPT chatbots did not have direct access to the internet, in contrast to Bard, which did. Indeed, there is an important difference between Bard and the ChatGPT chatbots. Bard can access Google's search engine, while ChatGPT (both versions) has no internet access and has only been trained on information available up to 2021.

Figure 2 presents a comparison of the three chatbots in terms of the number of words used for each generated response. We see that ChatGPT-3.5 has used the most words in its responses, with an average of 169 words per response for Set A and 182 words per response for Set B. Similarly, ChatGPT-4 has used 119 words for Set A and 163 words for Set B. We see that both chatbots have used more words for the problems of Set B, with the increase being 7.7% for ChatGPT-3.5 and 37.0% for ChatGPT-4. Bard used the lowest number of words, with an average of 106 words per response for Set A and 116 for Set B (an increase of 9.4%).



(**a**) (**b**)

Figure 2. Average number of words used per response for each chatbot for the two sets of questions: (**a**) results for each of the three rounds, (**b**) results for all rounds.

6. Discussion, Conclusions and Future Research Directions

In this study, we compared the performances of three leading chatbot systems, namely ChatGPT-3.5, ChatGPT-4, and Google Bard, in solving math and logic problems. Our aim was to evaluate their understanding, problem-solving abilities, and overall effectiveness in tackling a diverse range of problems. Our findings revealed that all three chatbots demonstrated, to some extent, an ability to understand and process math and logic problems, with some exceptions and limitations. These models can be used to solve basic mathematics and logic problems as they have learned to perform simple calculations and understand logical concepts from their training data. However, their capabilities in this domain have certain limitations. For straightforward arithmetic, algebraic expressions, or basic logic puzzles, they may provide accurate solutions, though not every single time. For more complex mathematical problems or advanced logic tasks, their performance may not be as reliable.

ChatGPT-4 clearly outperformed ChatGPT-3.5 in terms of accuracy and handling complex problems. The improved performance of ChatGPT-4 can be attributed to its larger model size, enhanced context understanding, and better fine-tuning compared to its predecessor. Nevertheless, ChatGPT-3.5 is much faster than ChatGPT-4, which is rather slow in generating its responses, which, again, is probably due to its larger model size. Bard on the other hand is fast and generates three responses at once, but it showed the poorest performance in the set of the original problems, although it exhibited the best performance in the set of published problems which can be found on the internet. Bard has direct access to the internet and to Google search, which gives it a competitive advantage when it comes to problems that have been published online, together with their solutions.

Overall, our study demonstrated the progress made in the field of AI chatbots, with the three chatbots showcasing notable advancements in reasoning and in solving math and logic problems. However, there remains room for improvement in terms of accuracy, handling complex problems, and natural language understanding, as certain limitations were observed in all three chatbots. Complex mathematical problems and those requiring advanced logical reasoning still posed challenges. Even some simple problems appeared to be challenging or hard for the chatbots. Moreover, occasional errors in understanding the problem or misinterpreting user input were observed, highlighting the need for further improvement in natural language understanding. It is essential to understand that these chatbots are primarily language models, not specialized tools for mathematics or logic. While they can demonstrate some problem-solving abilities in these areas, dedicated software or specialized models would be better suited for more complex or advanced tasks in mathematics and logic. Future research should prioritize the resolution of the limitations mentioned in the study and delve into methods for improving the chatbots' learning and problem-solving abilities. This includes developing specialized algorithms and models that enable chatbots to tackle complex mathematical problems and advanced logical reasoning tasks with greater accuracy and efficiency.

Another problem we observed is the so-called "AI hallucination" effect, where in many cases the solution the chatbots provide is very long, detailed, and written in a "professional" way, but it still may be completely wrong or nonsensical when examined more carefully. This may fool a human into thinking that such a detailed and long solution would be correct; so, extra caution is needed when we use such tools for solving similar exercises. In other words, a chatbot will rarely claim that it does not know the answer to a problem, and will not state its confidence in its solution, like a human would normally do. It will simply give an answer, and the user is not able to know whether this answer can be considered trustworthy or not. The problem becomes more prominent as models become more truthful and as users build trust in them. Future research should explore methods to detect and rectify this "AI hallucination" effect, providing more dependable outputs.

Another issue has to do with the consistency of the responses. In many cases, a model would correctly respond once but would miserably fail in the very next attempt. There is no guarantee that in a given attempt the model will get it correct. This is particularly a problem for questions where we do not know the exact answer and rely on the response of the chatbot to provide it. Future studies could focus on developing mechanisms for chatbots to estimate their confidence levels in responses. Implementing confidence scores or disclaimers when responses are generated can help users gauge the reliability of the chatbot's solutions.

In addition, in the future it is important to apply formal methods in the design, development, and verification of chatbot systems. Formal methods are a set of mathematical techniques and tools used in computer science and engineering to specify, develop, and verify software and hardware systems. These methods have been used to verify the correctness of the smart contract code, which can help to prevent costly errors and security breaches [28,29]. Formal methods can play a role in ensuring the correctness, reliability, and safety of chatbots' responses.

Last but not least, it has to be noted that the developments in the field of chatbots and LLMs are extremely fast, and the situation is dynamic and changes all the time. The questions of this research work were posed to chatbots in May 2023. Since then, chatbots have been developed and improved, and they are constantly receiving updates. For example, on 7 June 2023, Google announced that Bard was getting better at mathematical tasks, coding questions, and string manipulation through a new technique called implicit code execution that helps Bard detect computational prompts and run code in the background. As a result, it can respond more accurately to mathematical tasks, coding questions, and string manipulation prompts [30].

Author Contributions: Conceptualization, V.P.; methodology, V.P., G.P. and A.J.R.; software, V.P.; validation, V.P., G.P. and A.J.R.; formal analysis, V.P. and A.J.R.; investigation, V.P. and G.P.; resources, V.P. and G.P.; data curation, V.P., G.P. and A.J.R.; writing—original draft preparation, V.P., G.P. and A.J.R.; writing—review and editing, V.P., G.P. and A.J.R.; visualization, V.P., G.P. and A.J.R.; supervision, V.P.; funding acquisition, A.J.R. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Oslo Metropolitan University (OsloMet).

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset containing the 30 questions, correct answers, explanations, and the full answers of all chatbots in all questions can be found in [16]. A preprint of this work has been published in [17].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
2. Kuhail, M.A.; Alturki, N.; Alramlawi, S.; Alhejori, K. Interacting with educational chatbots: A systematic review. *Educ. Inf. Technol.* **2023**, *28*, 973–1018. [CrossRef]
3. Nguyen, H.D.; Tran, D.A.; Do, H.P.; Pham, V.T. Design an Intelligent System to automatically Tutor the Method for Solving Problems. *Int. J. Integr. Eng.* **2020**, *12*, 211–223. [CrossRef]
4. Tatai, G.; Csordás, A.; Kiss, Á.; Szaló, A.; Laufer, L. Happy Chatbot, Happy User. In *Intelligent Virtual Agents. 4th International Workshop, IVA 2003*; Rist, T., Aylett, R.S., Ballin, D., Rickel, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2792, pp. 5–12. [CrossRef]
5. Hu, K. ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note. 2023. Available online: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (accessed on 19 May 2023).
6. Bryant, A. AI Chatbots: Threat or Opportunity? *Informatics* **2023**, *10*, 49. [CrossRef]
7. Cheng, H.-W. Challenges and Limitations of ChatGPT and Artificial Intelligence for Scientific Research: A Perspective from Organic Materials. *AI* **2023**, *4*, 401–405. [CrossRef]
8. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef] [PubMed]
9. Frieder, S.; Pinchetti, L.; Griffiths, R.-R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P.C.; Chevalier, A.; Berner, J. Mathematical Capabilities of ChatGPT. *arXiv* **2023**, arXiv:2301.13867. [CrossRef]
10. Shakarian, P.; Koyyalamudi, A.; Ngu, N.; Mareedu, L. An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP). *arXiv E-Prints* **2023**, arXiv:2302.13814.
11. Upadhyay, S.; Chang, M.-W. Annotating Derivations: A New Evaluation Strategy and Dataset for Algebra Word Problems. *arXiv* **2017**, arXiv:1609.07197.
12. Upadhyay, S.; Chang, M.-W.; Chang, K.-W.; Yih, W.-t. Learning from Explicit and Implicit Supervision Jointly For Algebra Word Problems. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; Association for Computational Linguistics: Kerrville, TX, USA, 2016; pp. 297–306.
13. Lan, Y.; Wang, L.; Zhang, Q.; Lan, Y.; Dai, B.T.; Wang, Y.; Zhang, D.; Lim, E.-P. MWPToolkit: An Open-Source Framework for Deep Learning-Based Math Word Problem Solvers. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 13188–13190. [CrossRef]
14. Zheng, S.; Huang, J.; Chang, K.C.-C. Why Does ChatGPT Fall Short in Answering Questions Faithfully? *arXiv* **2023**, arXiv:2304.10513.

15. Lai, U.H.; Wu, K.S.; Hsu, T.-Y.; Kan, J.K.C. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. *Front. Med.* **2023**, *10*, 1240915. [CrossRef] [PubMed]
16. Plevris, V.; Papazafeiropoulos, G.; Jiménez Rios, A. Dataset of the study: "Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard". *Zenodo* **2023**. [CrossRef]
17. Plevris, V.; Papazafeiropoulos, G.; Jiménez Rios, A. Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *arXiv* **2023**, arXiv:2305.18618.
18. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [CrossRef]
20. Alruqi, T.N.; Alzahrani, S.M. Evaluation of an Arabic Chatbot Based on Extractive Question-Answering Transfer Learning and Language Transformers. *AI* **2023**, *4*, 667–691. [CrossRef]
21. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
22. Heaven, W.D. GPT-4 is Bigger and Better Than ChatGPT—But OpenAI Won't Say Why. 2023. Available online: https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/ (accessed on 19 May 2023).
23. Parade. 25 Logic Puzzles That Will Totally Blow Your Mind, But Also Prove You're Kind of a Genius. 2023. Available online: https://parade.com/970343/parade/logic-puzzles/ (accessed on 11 May 2023).
24. Feiveson, L. These 20 Tough Riddles for Adults Will Have You Scratching Your Head. 2022. Available online: https://www.popularmechanics.com/science/math/a31153757/riddles-brain-teasers-logic-puzzles/ (accessed on 11 May 2023).
25. math10.com. Math Word Problems and Solutions—Distance, Speed, Time. 2023. Available online: https://www.math10.com/en/algebra/word-problems.html (accessed on 11 May 2023).
26. Wolfram Alpha LLC. Examples for Mathematical Word Problems. 2023. Available online: https://www.wolframalpha.com/examples/mathematics/elementary-math/mathematical-word-problems (accessed on 12 May 2023).
27. 15 Workers Are Needed to Build a Wall in 12 Days. How Long Would 10 Workers Take to Build the Wall? 2023. Available online: https://www.quora.com/15-workers-are-needed-to-build-a-wall-in-12-days-how-long-would-10-workers-take-to-build-the-wall (accessed on 12 May 2023).
28. Krichen, M.; Lahami, M.; Al–Haija, Q.A. Formal Methods for the Verification of Smart Contracts: A Review. In Proceedings of the 2022 15th International Conference on Security of Information and Networks (SIN), Sousse, Tunisia, 11–13 November 2022; pp. 1–8. [CrossRef]
29. Abdellatif, T.; Brousmiche, K. Formal Verification of Smart Contracts Based on Users and Blockchain Behaviors Models. In Proceedings of the 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), Paris, France, 26–28 February 2018; pp. 1–5. [CrossRef]
30. Krawczyk, J.; Subramanya, A. Bard Is Getting Better at Logic and Reasoning. 2023. Available online: https://blog.google/technology/ai/bard-improved-reasoning-google-sheets-export/ (accessed on 4 September 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.