IEPG SUMMIT'23
PENSANDO O FUTURO COM INOVAÇÃO E SUSTENTABILIDADE

GPT-4 *vs* Bard: A Performance Evaluation in Statistics

Alex Fernandes de Souza – alex154237@gmail.com
Pedro Paulo Balestrassi – pedro@unifei.edu.br

Keywords: Chatbot, Performance Evaluation, Artificial Intelligence.

1 INTRODUCTION

In recent years, the rise of chatbots driven by artificial intelligence has transformed the way businesses and individuals interact with technology. These chatbots, also known as virtual assistants, are computer programs designed to simulate conversations with human users, especially over the internet. They have become essential tools in various sectors, from customer service to education, streamlining processes and providing quick and accurate responses.

Among the chatbots in the market, Bard and GPT-4 stand out for their advanced text and image reading capabilities, thanks to cutting-edge technologies. Both promise to revolutionize human interaction in areas where statistics are crucial, from medicine to engineering. Thus, the research question arises: Can we trust GPT-4 and Bard when addressing statistical challenges in these areas?

Therefore, the aim of this study was to conduct a thorough comparative assessment between GPT-4 and Bard. A statistical test containing 20 questions was applied, aiming not only to identify which of the two performs better but also to understand the nuances and specificities of each in the analysis and interpretation of information, whether textual or in graphical forms.

2 DEVELOPMENT

2.1 Literature review

Artificial Intelligence (AI) refers to the simulation of human intelligence processes by computer systems. This involves the acquisition of information (learning), reasoning to use

this information (logic), and self-regulation to improve accuracy (optimization). In the realm of digital communication, AI predominantly manifests through chatbots, which are software programs designed to simulate conversations with users, replicating human interactions (Aggarwal *et al.*, 2023).

The Generative Pre-trained Transformer (GPT) series stands as a prime example of an AI-driven chatbot. Originating from OpenAI, the GPT was built using transformers, which are deep learning models specialized in understanding and generating natural language. Since its inaugural version, the GPT has undergone various evolutions, with notable enhancements in its text comprehension and generation capabilities. The model was pre-trained on vast datasets, allowing it to acquire a comprehensive grasp of human language (Waisberg *et al.,* 2023).

Bard, although lesser-known in some circles than GPT, represents another significant innovation in the chatbot world. Just like GPT, Bard is powered by cutting-edge technologies in natural language processing, but may have specific nuances in its approach and capabilities, depending on its architecture and the data it was trained on. While many chatbots primarily focus on textual language, the distinction of these two systems lies in their unique ability to read and interpret images. This skill to integrate visual information with textual contexts marks a milestone in the chatbot field, positioning Bard and GPT-4 as leaders in innovation and potential within the realm of Artificial Intelligence (Toyama *et al.,* 2023).

The ability of these bots to process information swiftly and accurately makes them invaluable tools in mathematical and statistical analyses. They can process vast amounts of data in fractions of a second, identify patterns, and provide insights that might be inaccessible or time-consuming for humans. Specifically, their ability to learn and adapt to new information makes them dynamic tools in fast-changing environments, such as statistical analysis (Waisberg *et al.*, 2023).

The potency and accuracy of chatbots, like GPT-4 and Bard, in statistical analysis have profound implications across various fields. In the industry, they can assist in decision-making, forecasting market trends, optimizing supply chains, and pinpointing inefficiencies.

In education, they can tailor learning, identifying areas of struggle for students, and adapting to provide targeted resources and support. The capability to analyze and interpret vast amounts of data meaningfully can transform how decisions are made, leading to more informed and effective solutions (Egli, 2023).

2.2 Methodology

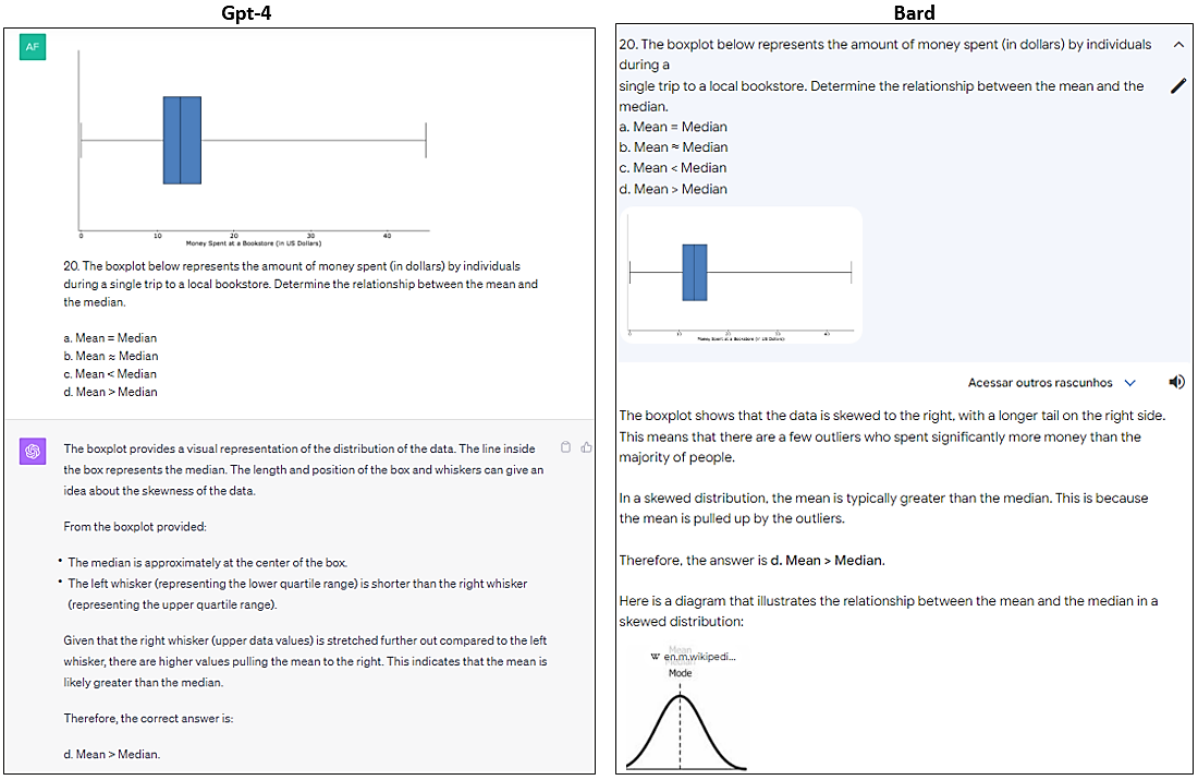
Each chatbot underwent a replication of the same statistical test containing 20 questions, derived from the "Arkansas Council of Teachers of Mathematics 2018 - Regional Statistics Exam[1]". This test features questions with graphs and tables, heightening the challenge for the chatbots' comprehension. These responses were then tabulated and compared to the correct answers previously established for each question in the questionnaire. The primary objective of this comparison was to investigate the accuracy rate of each chatbot.

In addition to the accuracy analysis, we sought to determine if there were significant statistical differences between the performances of the two chatbots. For this analysis, a hypothesis test was conducted. Furthermore, to understand the agreement between the answers of the two bots, an analysis using the Kappa concordance coefficient was conducted. This metric offers a deeper understanding of the consistency between the responses provided by the two systems under assessment.

2.3 Results and Discussion

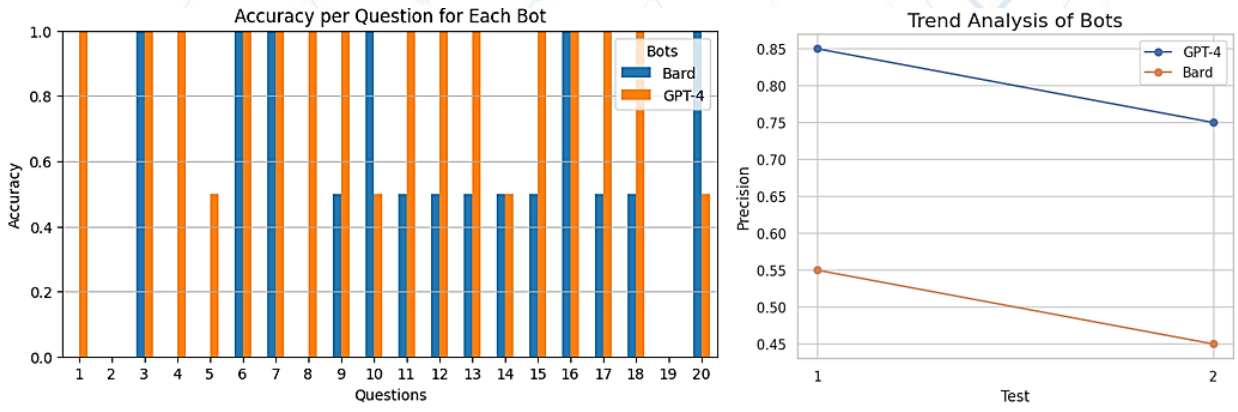
Both GPT-4 and Bard excel in reading and interpreting visual content like images, a defining trait that sets them apart in chatbot technology. Their ability to integrate and analyze graphical elements significantly enhances user interaction. Figure 1 vividly demonstrates this capability, showcasing how they respond to queries with graphical elements.

Figure 1 – Reading of questions containing graphs using GPT-4 and Bard.



Based on Figure 1, it is observed that in both cases, the chatbots correctly answered the question, which has "d" as the correct option. However, overall, the global results reveal a significant difference between the two analyzed tools. Figure 2 graphically displays the accuracy of the bots, highlighting that each tool was tested twice.

Figure 2 – Number of correct answers.



Examining the bar chart in Figure 2, the y-axis ranges from 0 to 1, and the x-axis represents the question numbers. A score of 1 on the y-axis signifies two correct answers and 0.5 represents one correct answer. Consequently, it is evident that GPT-4 achieved a higher number of correct answers (32 out of 40), while Bard answered 20 accurately.

Upon analyzing the 'Trends Analysis of Bots' graph in Figure 2, it is observed that GPT-4's accuracy was 0.85 in Test 1 and 0.75 in Test 2, indicating an assertiveness of 85% and 75% in the first and second tests, respectively. Bard achieved an accuracy of 55% in the first test and 45% in the second, thereby demonstrating GPT-4's superior performance, maintaining a high number of correct answers in both tests.

Thus, it is also noted that GPT-4 maintains a certain consistency in its responses, as does Bard. A Hypothesis Test using t-statistics (T-Test) was then conducted, as shown in Figure 3, to determine if there is a significant difference between the means of two groups. This can be useful in scenarios such as comparing the performance of two different chatbots. The following results were obtained.

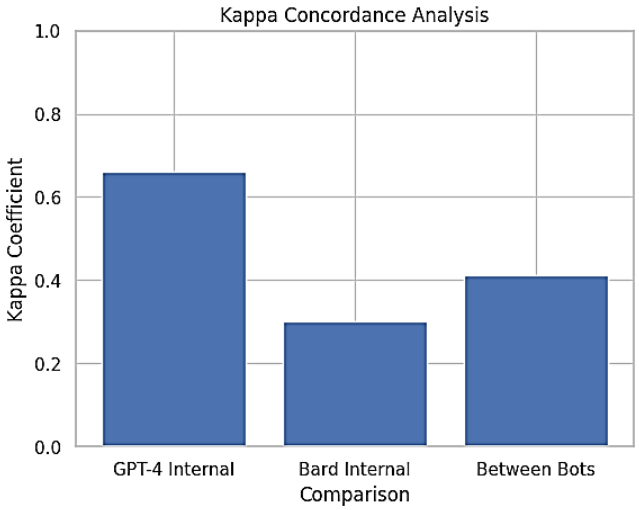
Figure 3 - Hypothesis test results comparing Bard and GPT-4

```
T-Statistic: 4.242640687119286
P-Value: 0.05131670194948619
```

The T-Statistic value indicates a difference between the bots; however, the P-Value suggests that this difference is on the threshold of statistical significance. Based on these results, one might infer that there is a trend suggesting varying performance between the chatbots, but this variance isn't statistically significant at the traditional 0.05 level.

Furthermore, the agreement between the responses within each chatbot, as well as between them, was examined. This exploration aims to determine whether the answers follow a consistent pattern or if the bots respond randomly. The outcomes of this examination are presented in Figure 4.

Figure 4 - Agreement analysis using the Kappa Coefficient.



GPT-4 demonstrated high internal consistency when answering the statistical test twice, suggesting that it is highly reliable in its responses. In contrast, Bard exhibited moderate consistency, which might indicate some variability in its answers to the same set of questions. When comparing the two bots, they only showed fair agreement, suggesting that while they might concur on some answers, there's a notable discrepancy between their responses.

3 CONCLUSION

Based on the conducted tests, it's clear that GPT-4 showcased superior performance in terms of accuracy and reliability. The Kappa analysis, which evaluates response agreement, further underscored GPT-4's remarkable consistency, displaying near-perfect agreement when subjected to the same test twice. In contrast, while Bard exhibited moderate agreement, it did not approach GPT-4's level of consistency.

ACKNOWLEDGMENTS

To the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the Doctoral scholarship awarded.

REFERENCES

AGGARWAL, Abhishek et al. Artificial Intelligence–Based Chatbots for Promoting Health Behavioral Changes: Systematic Review. **Journal of Medical Internet Research**, v. 25, p. e40789, 2023.

EGLI, Adrian. ChatGPT, GPT-4, and Other Large Language Models: The Next Revolution for Clinical Microbiology?. **Clinical Infectious Diseases**, p. ciad407, 2023.

JOHNSON, Douglas et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. **Research square**, 2023.

TOYAMA, Yoshitaka et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. **Japanese Journal of Radiology**, p. 1-7, 2023.

WAISBERG, Ethan et al. GPT-4: a new era of artificial intelligence in medicine. **Irish Journal of Medical Science (1971-)**, p. 1-4, 2023.