ARTIFICIAL INTELLIGENCE IN THE ANALYSIS OF EDUCATIONAL RESEARCH QUANTITATIVE DATA: RELIABILITY OF DATA ANALYST GPT (CHATGPT) COMPARED TO SPSS AND JAMOVI

INTELIGÊNCIA ARTIFICIAL NA ANÁLISE DE DADOS QUANTITATIVOS DE PESQUISA EDUCACIONAL: CONFIABILIDADE DO DATA ANALYST GPT (CHATGPT) COMPARADO AO SPSS E JAMOVI

INTELIGENCIA ARTIFICIAL EN EL ANÁLISIS DE DATOS CUANTITATIVOS DE INVESTIGACIÓN EDUCATIVA: CONFIABILIDAD DE DATA ANALYST GPT (CHATGPT) COMPARADO CON SPSS Y JAMOVI

Cassio SANTOS[1]
e-mail: cassiosantos@ie.ulisboa.pt

How to reference this paper:

SANTOS, C. Artificial Intelligence in the Analysis of Educational Research Quantitative Data: Reliability of Data Analyst GPT (ChatGPT) compared to SPSS and JAMOVI. **Nuances: Estudos sobre Educação**, Presidente Prudente, v. 35, n. 00, e024013, 2024. e-ISSN: 2236-0441. DOI: https://doi.org/10.32930/nuances.v35i00.10682

| **Submitted**: 20/06/2024
| **Revisions required**: 15/07/2024
| **Approved**: 12/08/2024
| **Published**: 11/10/2024

[1] Unidade de Investigação e Desenvolvimento em Educação e Formação (UIDEF), Instituto de Educação (IE), Universidade de Lisboa (ULisboa), Lisboa – Portugal. Professor and Researcher.

ABSTRACT: The integration of Artificial Intelligence (AI) into the educational and research landscape marks a transformative era, offering unparalleled opportunities for enhancing the way earn and conduct research. This article explores the potential of the AI-based language model, Data Analyst GPT, developed by OpenAI, as a reliable tool for conducting quantitative data analysis. The methodology involved employing Data Analyst GPT and two standard statistical software packages, SPSS and JAMOVI, to conduct an end-to-end statistical analysis on a typical educational data set, covering several standard statistical tests such as normality, correlation analysis (Pearson's and Spearman's), Categorical Variables Analysis, and mean comparison tests (Test t, ANOVA, Tukey, Mann-Whitney U and Kruskal-Wallis), and their results were compared. The results demonstrate a consistency comparable to that of standard statistical software.

KEYWORDS: Data Analyst GPT. ChatGPT. SPSS. JAMOVI. Artificial Intelligence (AI).

RESUMO*: A incorporação da Inteligência Artificial (IA) no cenário educacional e de pesquisa marca uma era transformadora, oferecendo oportunidades sem precedentes para aprimorar a forma como aprendemos e realizamos pesquisas. Este artigo explora o potencial do modelo de linguagem baseado em IA, Data Analyst GPT, desenvolvido pela OpenAI, como uma ferramenta confiável para realizar análises de dados quantitativos. A metodologia envolveu o uso do Data Analyst GPT e de dois softwares estatísticos padrão, SPSS e JAMOVI, para realizar uma análise estatística completa em um conjunto de dados educacionais típico, abrangendo vários testes estatísticos padrão, como testes de normalidade, análise de correlação (Pearson e Spearman), análise de variáveis categóricas e testes de comparação de médias (teste t, ANOVA, Tukey, Mann-Whitney U e Kruskal-Wallis), e seus resultados foram comparados.*

PALAVRAS-CHAVE*: Data Analyst GPT. ChatGPT. SPSS. JAMOVI. Inteligência Artificial (IA).*

RESUMEN*: La incorporación de la Inteligencia Artificial (IA) en el ámbito educativo y de investigación marca una era transformadora, ofreciendo oportunidades sin precedentes para mejorar la forma en que aprendemos y realizamos investigaciones. Este artículo explora el potencial del modelo de lenguaje basado en IA, Data Analyst GPT, desarrollado por OpenAI, como una herramienta confiable para llevar a cabo análisis de datos cuantitativos. La metodología involucró el uso de Data Analyst GPT y dos softwares estadísticos estándar, SPSS y JAMOVI, para realizar un análisis estadístico completo en un conjunto de datos educativos típico, abarcando varias pruebas estadísticas estándar, como pruebas de normalidad, análisis de correlación (Pearson y Spearman), análisis de variables categóricas y pruebas de comparación de medias (prueba t, ANOVA, Tukey, Mann-Whitney U y Kruskal-Wallis), y sus resultados fueron comparados. Los resultados demuestran una consistencia comparable a la de los software estadísticos estándar.*

PALABRAS CLAVE*: Data Analyst GPT. ChatGPT. SPSS. JAMOVI. Inteligencia Artificial (IA).*

Introduction

Artificial Intelligence (AI) is now central to various societal sectors, with stakeholders crafting guidelines focused on ethics (European Commission, 2019; UNESCO, 2021), research responsibility (European Commission, 2024) and educational strategies tailored for educators (European Commission, 2022). Numerous higher education institutions, including Stanford University (2021), contribute to this discourse by developing comprehensive guides. Major institutions published guides for the ethical use of AI, indicating a concerted effort towards responsible AI integration on a global scale (European Commission, 2019; UNESCO, 2021).

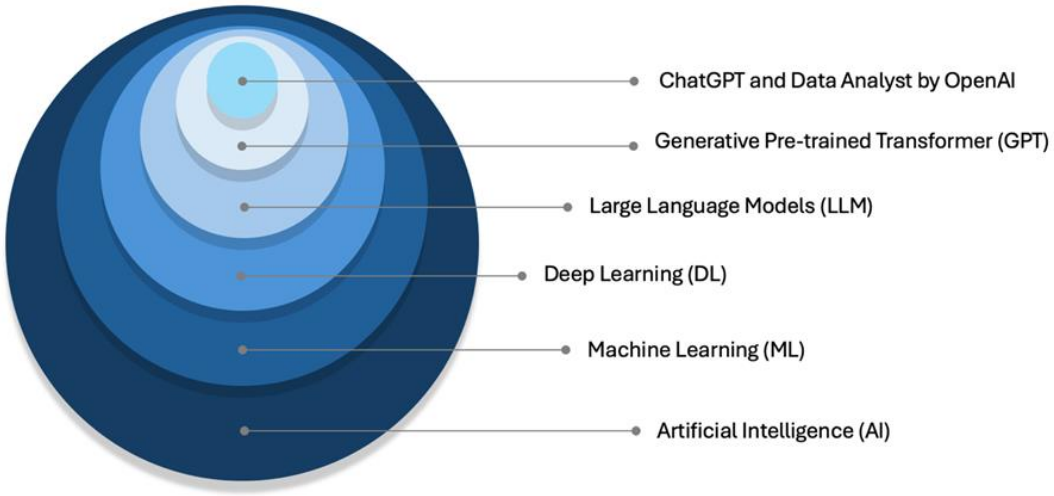
UNESCO has expressed its concern regarding Artificial Intelligence (AI) in higher education. This concern encompasses everything; from the development of Quick Start Guides (UNESCO, 2023a) to more specific issues, such as the use of ChatGPT, as well as broader topics to help stakeholders make better use of AI (UNESCO, 2023b). Additionally, the development of frameworks focused on AI-related competencies is also highlighted (Ehlers *et al.*, 2023).

In the European context, notably, landmark achievements include the European Commission's approval of the world's first guidelines for AI usage (European Commission, 2021), as well as the formulation of Ethics Guidelines for Trustworthy AI (European Commission, 2019).

The evolution of OpenAI's Generative Pre-Trained Transformer (GPT) series began with the inaugural launch in 2018, followed by the GPT-3 model which debuted in 2020, which represented a significant advancement in AI capabilities. Leveraging the foundation laid by GPT-3, ChatGPT emerged in 2022 as a purpose-built platform tailored for conversational AI and chatbot applications (Li *et al*., 2024; Singh-Harjit, Singh-Avneet, 2023).

The relationship between AI, Machine Learning (ML), Deep Learning (DL), Large Language Models (LLM) and GPT is one of hierarchy and specialisation. The ML (Lary *et al.*, 2016) provides the overarching framework for learning from data; DL (Alzubaidi *et al.*, 2021) offers advanced techniques for learning from complex, high-dimensional datasets; LLM (Chang, 2023; Fan *et al.*, 2023; Li *et al.*, 2024; Naveed *et al.*, 2023) processes and generates natural language at a large scale to facilitate natural human-computer interaction; and GPT, which uses DL and LLM, provides users with coherent and contextually relevant answers for their questions (Gimpel *et al.*, 2023). The Data Analyst GPT is a personalised version of the ChatGPT, optimised for data analysis. Figure 1 illustrates such a hierarchy of specialisation.

Figure 1. Hierarchy and Specialisation of Artificial Intelligence.



Source: Prepared by the author (2024)

The implementation of AI in education has had a significant impact, evidenced by improvements in the efficiency of the educational process, the promotion of global learning, the personalisation of learning, the creation of more intelligent content and the optimisation of educational management in terms of effectiveness and efficiency (Montenegro-Rueda *et al.*, 2023). The relationship between AI and research in higher education is two-fold: "The first relates to research on AI, whereas the second is about research using or supported by AI tools Research" (UNESCO, 2023b, p. 38). AI can process large volumes of data (Gimpel *et al.*, 2023), automatically learn to identify complex patterns and hidden trends, and it has the flexibility to adapt to different types of data and research contexts. Therefore, it can enhance the understanding and interpretation of quantitative data in the field of education.

There is a vast amount of research in the academic literature focused on using AI in education (Al-Ghonmein, Al-Moghrabi, 2024; Crawford *et al*., 2024; Ding *et al*., 2023; Jia, Tu, 2024). However, studies exploring this technology's potential in analysing data are still relatively scarce (Huang *et al*., 2024; Mohammadi, Nguyen, 2024; Sufi, 2024; Walter, 2024).

Following the directives laid out in the document "Living Guidelines on the Responsible Use of Generative AI in Research" (European Commission, 2024), it is crucial to underline the importance of responsibility and integrity on the part of researchers concerning scientific output supported by AI. This document emphasises the need for researchers to ultimately remain responsible for the scientific content generated or supported by AI tools, to adopt a critical stance, and be aware of the inherent limitations of generative AI, such as biases and inaccuracies.

Transparency in the use of these AI tools is also a key point highlighted in the research guidelines (European Commission, 2024). Researchers are encouraged to describe which generative AI tools have been used in their research processes, including information like the name, version, and date of the tool, and how it influenced the research process. Proper documentation of inputs (prompts) and outputs, whenever relevant, is encouraged to promote openness and replicability of research. Lastly, these guidelines encourage researchers to engage in continuous learning about the proper use of generative AI tools. Given the rapid development of these technologies and the constant emergence of new applications, researchers need to stay updated on best practices, participate in training, and share knowledge with colleagues and other stakeholders, in order to to maximise the benefits of these advanced tools for research.

Several institutions have expressed their concerns about privacy, confidentiality, and intellectual property rights, either concerning the fact that "models such as ChatGPT are opaque to the dataset that has been used to train them" (UNESCO, 2021, p. 7) or when sharing sensitive or protected information with AI tools, "researchers remain mindful that generated or uploaded input (text, data, prompts, images, etc.) could be used for other purposes, such as the training of AI models" (European Commission, 2024, p. 6).

> In research, ChatGPT can assist in data analysis and summarising large sets of data, which can help researchers quickly and easily identify patterns and insights that would be difficult to uncover manually. Additionally, the model can be used to generate research proposals, literature reviews, and other research-related documents (Atlas, 2023, p. 24).

This article aims to investigate the reliability of Data Analyst GPT, the personalised and optimised version of the ChatGPT-4o for data analysis, providing an intelligent and versatile conversational interface for analysing quantitative research data. Its performance will be compared with that obtained by using two standard statistical software packages, SPSS and JAMOVI.

Methodology

This section details study procedures introducing GPT Data Analyst and benchmark software. In Data Analyst GPT, custom prompts were developed to run the calculations, and SPSS and JAMOVI tests were conducted according to the procedures outlined in their respective support manuals.

Data Analyst GPT

The AI-based tests were conducted using the Data Analyst GPT, the ChatGPT version optimised for data analysis, using the Plus subscription plan. In ChatGPT, the "Data Analyst" GPT is accessed through the "Explore GPTs" section, where an Excel file (*xlsx) containing the dataset to be analysed can be uploaded. At the time of this writing, the GPT-4o model was the most advanced in the GPT series.

Standard statistical software as a benchmark

Two standard statistical software packages were selected to be used as a benchmark, namely the SPSS (version 29.0.2.0 [20]) and the JAMOVI (version 2.3.21.0), both operating on a Mac OS system. The SPSS is a widely used statistical software that allows various types of analysis, transformations and output forms (Alili; Krstev, 2019). The JAMOVI2 (R Core Team, 2021; The Jamovi Project, 2022) is a popular free and open-source statistical software, which was adopted by the research community due to its ease-of-use and comprehensive suite of statistical functions, from basic analyses to advanced univariate and multivariate techniques (Algthami, Hussin, 2022; Marek *et al*., 2023).

Dataset

The dataset adopted in this article is based on already published research. The original dataset was adapted, and new variables were added to cover a wider range of the statistical tests needed to verify the reliability of the Data Analyst GPT. The dataset contains a variety of variables, which allow the testing of different hypotheses and scenarios.

Statistical tests

This article covers the statistical tests most used in educational research, namely normality tests, correlation analysis, categorical variables analysis and mean comparison tests.

[2] https://www.jamovi.org/about.html

Normality

Two complementary approaches were employed for assessing the normality of the data distribution, namely the Shapiro-Wilk statistical analysis and the visual inspection of the data distribution using graphical analysis with boxplots and histograms. This provides a robust assessment of normality, allowing the visual identification of asymmetries, outliers, and the general shape of the distribution. Both analyses were carried out on the "points_1" and "points_2" variables.

The Shapiro-Wilk test was used to compare the data from a sample to a set of data that follows a normal distribution, i.e., with the same mean and standard deviation. In this test, non-significant results ($p>0.05$) indicate that the distribution of the sample data does not significantly differ from a normal distribution, suggesting that the data follow a normal distribution. Conversely, a significant result ($p<0.05$) means that the distribution of the data is significantly different from a normal distribution, implying that the data does not follow a normal distribution (Dancey; Reidy, 2020; Field, 2024).

The visual approach allows researchers and analysts to conduct a detailed and intuitive inspection of the data distribution, facilitating the identification of important characteristics, such as skewness and kurtosis. By using specific graphs, such as boxplots and histograms, it is possible to observe patterns, trends, and deviations that might not be evident through purely numerical or statistical methods (Field, 2024).

A boxplot is an effective graphical representation that highlights in evidence the essential characteristics of a dataset, which is especially useful when the data adhere to a normal distribution. At its centre is the median, neatly contained within a box. This box's upper and lower boundaries represent the upper and lower quartiles, respectively, demarcating the interquartile range that encompasses the central 50% of the data points. Projecting from the box, whiskers extend to the highest and lowest data points, delineating the data's overall spread. Similarly, a histogram serves as a graphical tool that depicts the frequency distribution of a dataset. It facilitates the visualisation of data distribution by illustrating the occurrence frequency of each value. This visualisation is achieved by segmenting the dataset into defined intervals, or "bins," and tallying the observations within these bins. These bins are designed to be sequential, distinct, and uniform in size (Field, 2024).

Prompt in Data Analyst GPT: "I need a Shapiro-Wilk test conducted on the 'point_1' data column, with the results presented in an APA format data table. This table should include the test statistic, the p-value (rounded to three decimal places), and the degrees of freedom.

Additionally, please generate a boxplot and histogram for the 'points_1' column to visually assess its distribution".

Prompt in Data Analyst GPT: "I need a Shapiro-Wilk test conducted on the 'point_2' data column, with the results presented in an APA format data table. This table should include the test statistic, the p-value (rounded to three decimal places), and the degrees of freedom. Additionally, please generate a boxplot and histogram for the 'points_2' column to visually assess its distribution".

Correlation Analysis

The Spearman and Pearson tests were employed to analyse the correlations present in the data, namely variables "points_1" and the "number_of_devices".

Correlation tests are used to assess both the strength and the direction of the association between two quantitative variables. The Spearman correlation test, also known as rho (ρ), is preferably used in situations where the data do not satisfy normality assumptions or when dealing with ordinal variables, provide a robust measure of correlation that does not assume a specific linear relationship. On the other hand, the Pearson correlation, symbolised by r, is indicated for data that exhibit a normal distribution and a linear relationship, providing a measure of the strength and direction of that linearity. Both tests range from -1 to 1, where values close to -1 or 1 indicate a strong linear relationship, whether negative or positive, respectively (Dancey; Reidy, 2020; Field, 2024).

Spearman's coefficient on variable "point_1"

Prompt in Data Analyst GPT: *"I need a Spearman's coefficient conducted on the 'points_1' data column between the "number_of_devices" data column, with the results presented in an APA format data table. This table should include the test statistic, the p-value (rounded to three decimal places)".*

Pearson's coefficient on variable *"point_2"*

Prompt in Data Analyst GPT: *"I need a Pearson's coefficient conducted on the 'points_2' data column between the "number_of_devices" data column, with the results presented in an*

APA format data table. This table should include the test statistic, the p-value (rounded to three decimal places)".

Categorical Variables Analysis

The Chi-square test was employed on the categorical variables' proficiency_level' and 'situation' to check their independence. The Chi-square test is a statistical method used to compare observed frequencies with expected frequencies across different categories of a categorical variable. It helps to determine if there are significant differences between categories, namely, if the observed frequencies deviate significantly from the expected frequencies by chance. It is widely used in research to test hypotheses about the association or independence between categorical variables (Dancey; Reidy, 2020; Field, 2024).

Prompt In Data Analyst GPT: *"I need a Chi-square conducted on the 'proficiency_level' data column between the "situation" data column, with the results presented in an APA format data table. This table should include the test statistic, the p-value (rounded to three decimal places) and the degrees of freedom".*

Mean Comparison Tests

Factors with two groups

The Mann-Whitney U test for non-parametric variables and the Test t for parametric variables were employed to analyse the dataset when there are factors with two groups.

The independent sample Test t is predicated on the assumption that the underlying populations from which the samples are drawn have normal distributions with equal variances, making it a rigorous tool for examining hypotheses about mean differences in a controlled, comparative context (Dancey; Reidy, 2020; Field, 2024). Unlike the Test t, the Mann-Whitney U does not assume the normality of distributions or equality of variances between the groups, making it particularly useful for data that do not meet parametric assumptions (Dancey; Reidy, 2020; Field, 2024).

Mann-Whitney U on variable "point_1"

Prompt in Data Analyst GPT: *"I need the Mann-Whitney U test conducted on the 'points_1' data column between the "gender" data column, with the results presented in an APA format data table, being 0 for Male and 2 for Female. This table should include the test statistic, the p-value (rounded to three decimal places)."*

Test t on the variable *"point_2"*

Prompt in Data Analyst GPT: *"I need test t conducted on the 'points_2' data column between the "gender" data column, with the results presented in an APA format data table, being 0 for Male and 2 for Female. This table should include the test statistic, the p-value (rounded to three decimal places) and the degrees of freedom."*

More than two groups

For the cases where there are more than two groups, the Kruskal-Wallis test was employed for non-parametric variables and the Analysis of Variance (ANOVA) for parametric variables.

The Kruskal-Wallis test is a non-parametric alternative to the one-way ANOVA for comparing more than two groups. It is used when the assumptions of the ANOVA are not met, particularly when the data is not normally distributed. This test assesses whether the median ranks of two or more groups differ significantly from each other.

Post-hoc tests are statistical comparisons conducted after an ANOVA to determine which specific groups differ from each other. These tests are necessary when an ANOVA indicates significant differences among group means, but the ANOVA itself does not specify which groups differ significantly. This article employed the Tukey method due to its ability to control the Type I error rate well across all pairwise comparisons (Dancey; Reidy, 2020; Field, 2024).

Kruskal-Wallis test on the variable "point_1"

Prompt in Data Analyst GPT: *"I need the Kruskal-Wallis test conducted on the 'points_1' data column between the "level_of_education" data column, with the results presented in an APA format data table, being 2 for bachelor's, 3 for master's and 4 for PhD. This table should*

include the test statistic, the p-value (rounded to three decimal places) and the degrees of freedom."

ANOVA on variable "point_2"

Prompt in Data Analyst GPT: *I need an ANOVA test conducted on the 'points_2' data column between the "level_of_education" data column, with the results presented in an APA format data table, being 2 for bachelor's, 3 for master's and 4 for PhD. I also need Levene's test on this data in an APA format data table. These tables should include the test statistic, the p-value (rounded to three decimal places) and the degrees of freedom.*

and

Prompt in Data Analyst GPT: *I need an ANOVA test conducted on the 'points_2' data column between the "situation" data column, with the results presented in an APA format data table, being 1 for employed, 2 for retired, 3 unemployed and 4 for student, for. I also need Levene's test on this data in an APA format data table. These tables should include the test statistic, the p-value (rounded to three decimal places) and the degrees of freedom. If a statistically significant difference is identified, perform the Tukey post-hoc test.*

Results

For comparative purposes, the same statistical tests were also carried out using two standard statistical software, SPSS and JAMOVI, allowing us to compare the outcomes directly with Data Analyst GPT.

Normality tests

Two distinct approaches were adopted for analysis performing normality tests: statistical analysis and graphical analysis.

Statistical Analysis

The Shapiro-Wilk test was to verify the normality of distributions, as can be seen in Table 1.

Table 1 - Shapiro-Wilk test

Software	*"point_1"*			*"point_2"*		
	statistic	**df**	***p-value***	**statistic**	**df**	***p-value***
Data Analyst GPT	0.994	845	<.001	0.998	845	0.555
SPSS	0.994	846	<.001	0.998	846	0.555
JAMOVI	0.994	-	<.001	0.998	-	0.555

Source: Prepared by the author (2024)

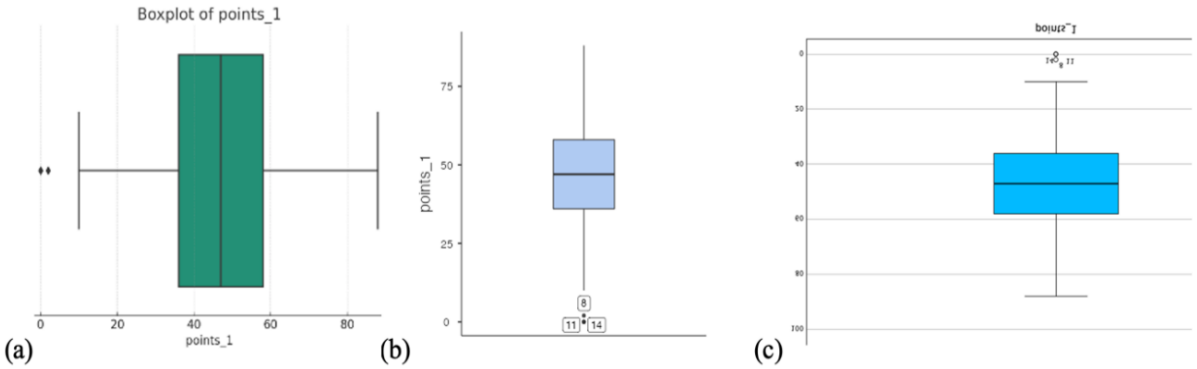
The findings indicate consistency in the Data Analyst GPT results compared to SPSS and JAMOVI for statistical values and p-values from the Shapiro-Wilk test. Nonetheless, a notable difference in the Degrees of Freedom (df) was observed, with Data Analyst GPT documenting 845, SPSS showing 846, and JAMOVI omitting this detail in both variables "point_1" and "point_2".

When selecting which statistical tests to use, it is important to know if the sample follows a normal distribution. In this case, the results obtained by statistical analysis show that the "point_1" variable does not follow a normal distribution ($p<0.05$), indicating a possible asymmetrical distribution or excess kurtosis. On the other hand, the "point_2" variable shows characteristics of normality ($p>0.05$), suggesting that its distribution is consistent with that of a normal distribution. This differentiation is crucial for the choice of statistical tests, guaranteeing the validity and reliability of the analyses.

Graphic Analysis

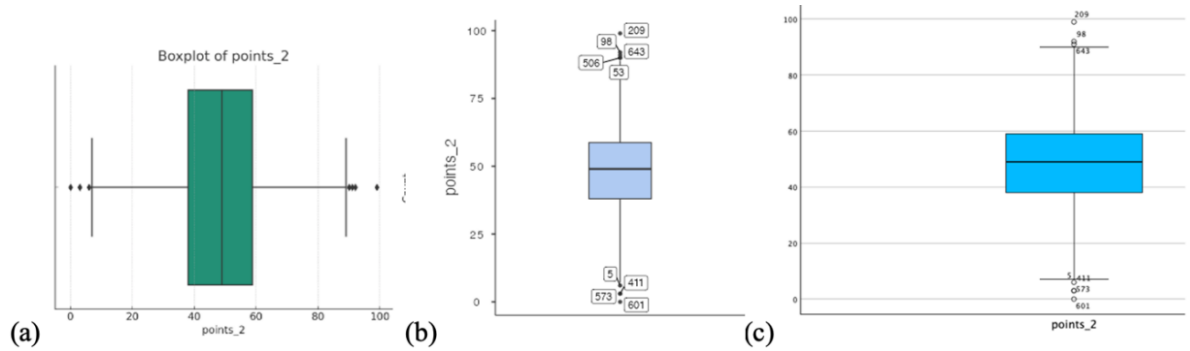
Boxplots and histograms ([a] Data Analyst GPT, [b] SPSS, and [c] JAMOVI) to verify their comparability when testing the normality of the distributions of the variables "point_1" and "point_2". The boxplots for the "point_1" variable can be seen in Graphic 2 and for the "point_2" variable, in Graphic 3. Additionally, histograms of the data for "point_1" are shown in Graphic 4 and for "point_2" in Graphic 5.

Graphic 2 – Boxplots obtained for the variable "point_1"



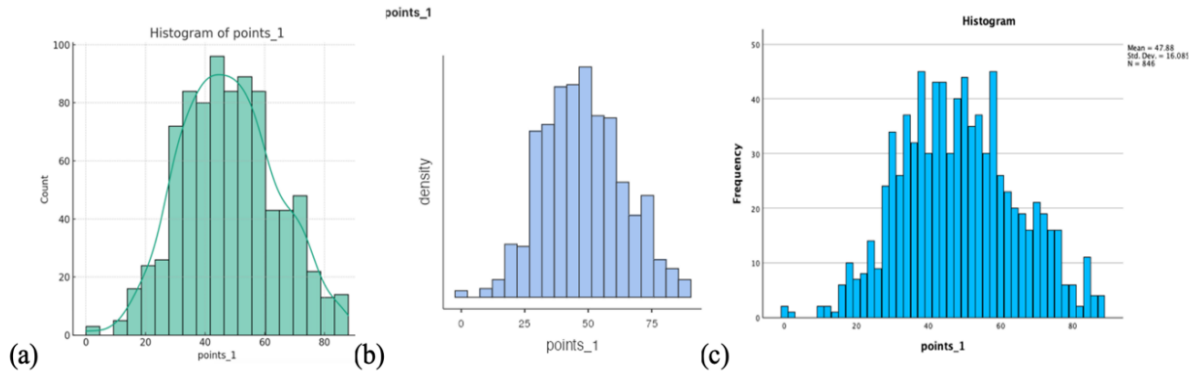
Source: Prepared by the author (2024)

Graphic 3 - Boxplots obtained for the variable "point_2"



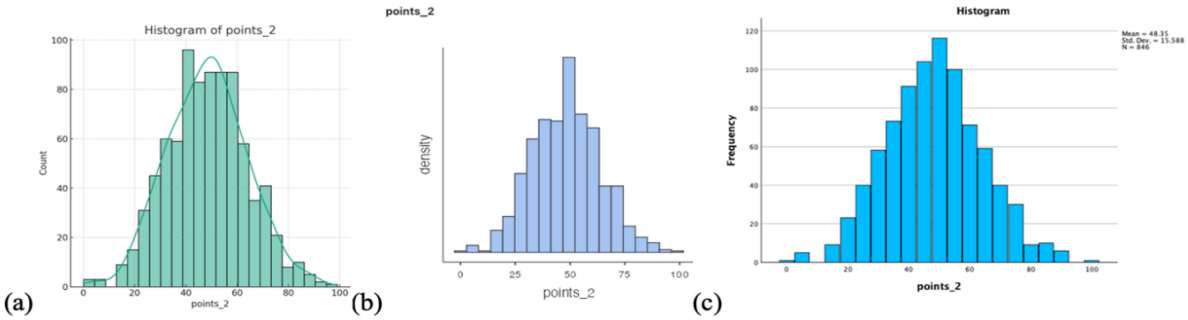
Source: Prepared by the author (2024)

Graphic 5 - Histograms obtained for the variable "point_2"



Source: Prepared by the author (2024)

Graphic 4 - Histograms obtained for the variable "point_1"



Source: Prepared by the author (2024)

The graphical analysis carried out using boxplots shows the reliability of Data Analyst GPT compared to SPSS and JAMOVI. It is important to emphasise that whereas Data Analyst GPT was only able to detect the presence of outliers, the SPSS and JAMOVI were also able to indicate which of them were outliers.

Correlation Analysis

Based on the results obtained by the normality tests, Spearman's coefficient was used for the non-parametric variable "point_1" ($p<0.05$), and Pearson's coefficient for the parametric variable "point_2" ($p>0.05$) with "number_of_devices", as can be observed in Table 2.

Table 2 – Spearman's ("point_1") and Pearson's coefficient ("point_2")

Software	Spearman's coefficient		Pearson's coefficient	
	ρ (rho)	***p-value***	**ρ (rho)**	***p-value***
Data Analyst GPT	0.228	<.001	-0.025	0.468
SPSS	0.228	<.001	-0.025	0.468
JAMOVI	0.228	<.001	-0.025	0.468

Source: Prepared by the author (2024)

The findings indicate the reliability of the Data Analyst GPT compared to SPSS and JAMOVI, both for Spearman's coefficient applied to data with a non-normal distribution and for Pearson's coefficient used for the normally distributed data.

Categorical Variables Analysis.

Chi-square tests were carried out on the categorical variables "points_2" and "situation", as shown in Table 3.

Table 3 - Chi-square

Software	statistic	df	*p-value*
Data Analyst GPT	50.767	15	<.001
SPSS	50.767	15	<.001
JAMOVI	50.8	15	<.001

Source: Prepared by the author (2024)

The findings indicate that Data Analyst GPT was able to provide consistent results when compared to SPSS and JAMOVI in calculating the categorical variables using the Chi-square test. It is important to note that, while JAMOVI reported the statistics to one decimal place, the other software displayed them in three decimal places. However, this does not impact the interpretation of the results.

Mean Comparison Tests

Factors with two groups

Based on the knowledge obtained about the normality of the data, the Mann-Whitney U test was used for the non-parametric variable "point_1" and the Test t for the parametric variable "point_2", as can be seen in Table 4.

Table 4 - Mann-Whitney U with "gender"

Software	Mann-Whitney U		Test t		
	statistic	*p-value*	statistic	df	*p-value*
Data Analyst GPT	92,001.5	0.390	0.810	844	0.418
SPSS	85,903.5	0.389	0.810	844	0.418
JAMOVI	85,904.0	0.390	0.810	844	0.418

Source: Prepared by the author

Regarding the p-value of the Mann-Whitney U test (Table 4), both Data Analyst GPT and JAMOVI presented similar results. SPSS presented slight variations in the third decimal place, likely due to rounding differences. A notable discrepancy was observed in the statistic value obtained by the Data Analyst GPT, whereas SPSS and JAMOVI presented similar results

with little variation, which can be attributed to decimal rounding. Again, such a difference does not affect the interpretation of the results.

More than two groups

Kruskal-Wallis tests were employed for non-parametric variables, such as "point_1" and "level_of_education," the ANOVA tests for parametric variables, such as "point_2," and the Levene tests, as can be seen in Table 5.

Table 5 – Kruskal-Wallis, ANOVA and Levene tests

Software	Kruskal-Wallis test			ANOVA test				Levene test	
	statistic	**df**	***p-value***	**statistic**	**df1**	**df2**	***p-value***	**F**	***p-value***
Data Analyst GPT	9.741	2	0.008	0.882	2	843	0.414	0.281	0.755
SPSS	9.741	2	0.008	0.882	2	843	0.414	0.281[a] 0.252[b]	0.755[a] 0.777[b]
JAMOVI	9.74	2	0.008	0.882	2	843	0.414	0.252	0.777

[a]Based on the median; [b]Based on the mean.
Source: Prepared by the author (2024)

The findings indicate that Data Analyst GPT was able to provide results comparable to the ones obtained by SPSS and JAMOVI, for both the Kruskal-Wallis test applied to data with a non-normal distribution and for the ANOVA test with normally distributed data.

Concerning Levene's test, the SPSS software offered two variants, one calculated based on the median and the other based on the mean, whereas both Data Analyst GPT and JAMOVI provide only one version of the result. The results suggest that Data Analyst GPT considered the median for the calculations, whereas JAMOVI used the mean. This correspondence demonstrates a strong consistency between both software.

Another ANOVA test was further carried out to analyse the reliability of the post-hoc test, now considering the "points_2" and "situation" variables. This was done to assess the software's reliability when there is a significant statistical difference between the means. The results for the ANOVA and Levene's tests are presented in Table 6, whereas the results for the Tukey post-hoc tests are presented in Table 7.

Table 6 – ANOVA and Levene tests

Software	ANOVA test				Levene test	
	statistic	df1	df2	*p-value*	F	*p-value*
Data Analyst GPT	3.372	3	842	0.018	0.108	0.956[a]
SPSS	3.372	3	842	0.018	0.108[a] 0.029[b]	0.956[a] 0.993[b]
JAMOVI	3.37	3	842	0.018	0.029[b]	0.993[b]

[a]Based on the median; [b]Based on the mean.
Source: Prepared by the author (2024)

Table 7 - Tukey post-hoc

Comparison		Data Analyst GPT	SPSS	JAMOVI
		statistic		
Employed	Retired	0.030	0.030	0.030
	Unemployed	0.900	0.946	0.946
	Student	0.484	0.483	0.483
Retired	Unemployed	0.177	0.177	0.177
	Student	0.030	0.030	0.030
Unemployed	Student	0.900	0.926	0.926

Source: Prepared by the author (2024)

The results demonstrate the reliability of Data Analyst GPT compared to SPSS and JAMOVI in the ANOVA tests when there is a statistically significant difference between the means, as well as in the Tukey post-hoc tests. It is important to note that the discrepancy observed in Levene's test has the same origin as the variability found in ANOVA, which occurs when there are no statistically significant differences between the means.

Discussion

This article aimed to assess the reliability of the Data Analyst GPT (ChatGPT) in quantitative data analysis by conducting a direct comparison with the results obtained from two classic statistical software packages, SPSS and JAMOVI. The tests selected for this direct comparison included a) tests for normality; b) correlation analysis using Pearson's coefficient for parametric samples and Spearman's for non-parametric ones; c) the Chi-square test for the analysis of categorical variables; and d) mean comparison tests, including the Test t and ANOVA for parametric samples, and the Mann-Whitney U and Kruskal-Wallis tests for non-parametric samples.

Two approaches were used to assess the reliability of the Data Analyst GPT in analysing normality: the Shapiro-Wilk statistical test for a quantitative assessment and graphical analysis with boxplots and histograms for a visual assessment.

Tests carried out with Data Analyst GPT and reference statistical software, namely SPSS and JAMOVI, require intermediate knowledge of statistics and data analysis (Huang *et al*., 2024), particularly for selecting the statistical tests to be carried out. In the case of Data Analyst GPT, execution is facilitated by a chat interace.

The Shapiro-Wilk test was applied to the "points_1" and "points_2" variables to verify the normality of the data, and the results were equivalent in terms of the test statistic and p-value. However, there is an apparent discrepancy in the Degrees of Freedom (df), with Data Analyst GPT registering 845, SPSS 846, and JAMOVI omitting this metric. It is important to clarify that, in the context of the Shapiro-Wilk test, the concept of degrees of freedom is not normally used, as this test focuses on assessing whether a sample comes from a normal distribution, without direct dependence on the degrees of freedom that usually apply to tests involving variations or standard deviations. Therefore, the mention of degrees of freedom in this context may not be essential, which may justify, at least in part, the absence of these values in the Shapiro-Wilk test performed by JAMOVI.

The boxplots generated to assess the data distribution were informative, as they illustrated their quartiles and highlighted the outliers. In the boxplots generated with SPSS (Graphic 2b) and JAMOVI (Graphic 2c), the outliers are indicated at the bottom; the boxplots produced by Data Analyst GPT (Graphic 2a) do not show such outliers; they only show their existence. Therefore, there is a limitation observed in the Data Analyst GPT regarding the visual representation of outliers, which can restrict a more in-depth analysis of extreme variations in the data. However, in cases where multiple outliers are present, visualising the outliers will prove challenging, regardless of the software used.

The histograms produced to assess the frequency distribution of the data provided a clear visualisation and were also informative. They allow a like-for-like comparison, even when their scales were automatically adjusted, and different data intervals were defined by the software. This could happen either on the X-axis (abscissa), which represents the frequency of each interval, with the highest bar indicating the highest frequency of values, or on the Y-axis (ordinate), with reflects to the numerical count of the corresponding occurrences on the X-axis is reflected. The histograms by Data Analyst GPT (Graphic 4a) were advantageous. The inclusion of a density curve or Kernel Density Estimate (KDE) provides an additional perspective on the overall distribution of the data, suggesting the shape of the underlying distribution in a more continuous and integrated way.

The results of the Spearman and Pearson correlation tests, as well as the Chi-square test, demonstrate Data Analyst GPT's comparable performance to SPSS and JAMOVI. This underscores the tool's reliability and accuracy in correlation analysis, affirming its capability to deliver robust analytical outcomes.

Regarding the Mann-Whitney U test for comparison of means, the results indicate the reliability of Data Analyst GPT compared to SPSS and JAMOVI in terms of p-value, even considering a slight difference in the third decimal attributable to rounding. However, Data Analyst GPT showed a significant difference in the values of the U statistic when compared to the other software.

The Mann-Whitney U test was initially devised by Frank Wilcoxon (Wilcoxon, 1945) to analyse measures of central tendency in samples of the same size. Later, Henry B. Mann and Donald R. Whitney (Mann & Whitney, 1947) extended their application to samples of different sizes. In this way, the Mann-Whitney U test statistical values can be derived through two distinct approaches: the Rank-based formulation (Wilcoxon, 1945) and the Direct comparison method (Mann & Whitney, 1947). The Rank-based formulation involves the combined ordering of all values from both groups, assigning ranks to each value, and using these ranks to calculate the U statistic, effectively adjusting for any ties (Wilcoxon, 1945). In contrast, the direct comparison method quantifies the number of times a value from one group exceeds that of the other, offering an intuitive approach that, despite its simplicity, becomes impractical for large sample analyses due to computational demands (Mann; Whitney, 1947).

The SPSS documentation (IBM Corporation, 2022) mentions the use of the Rank-based formulation, but equivalent documents for both Data Analyst GPT and JAMOVI were not found. The similarity of the U statistics results between SPSS and JAMOVI (85,903.5 and 85,904.0, respectively) might indicate that JAMOVI also employs the Rank-based method, whereas Data Analyst GPT adopts the Direct comparison method.

Considering the Mann-Whitney p-value, it can be concluded that the Data Analyst GPT provided reliable results compared to SPSS and JAMOVI.

The Kruskal-Wallis and ANOVA tests show that Data Analyst GPT reliability is comparable to SPSS and JAMOVI. This equivalence also extends to the results of the Tukey post-hoc tests, applied when ANOVA indicated the presence of statistically significant differences between groups, and to Levene's test to verify the homogeneity.

When faced with the processing limitation in Data Analyst GPT, even when using the paid version, ChatGPT Plus (GPT-4o), the main impact perceived was on confidence in its

availability. The OpenAI displays a message[3] indicating that the usage limit will be dynamically adjusted to prioritise access to GPT-4o by the greatest number of people according to demand and system performance. It also indicates a limit of 40 messages every 3 hours. This calls into question the availability of the Data Analyst GPT. This unexpected interruption and the need to pause for approximately two hours before resuming analyses highlight a significant concern: the lack of clarity and transparency regarding the current limitations of ChatGPT Plus (GPT-4o) at the time of subscription, especially for those who rely on the Data Analyst GPT tool to carry out continuous data analyses.

Another limiting factor that must be considered when adopting Data Analyst GPT is the lack of specification of the tool's version. Whereas in this article, the JAMOVI version 2.3.21.0 and SPSS version 29.0.2.0 were used and known, in Data Analyst GPT, the exact version was unknown, the only version was the GPT-4o model, given that as an artificial intelligence model, it has a learning capacity. It is, therefore, important that research carried out with Data Analyst GPT is accompanied by the execution prompt (European Commission, 2024).

Restoring and maintaining user confidence requires clear and comprehensive communication from the developers about all operational aspects, including possible usage limits. Such transparency at the time of subscription is essential to ensure that users can properly plan their use of Data Analyst GPT, avoiding unpleasant surprises and ensuring that availability expectations align with the tool's operational reality.

Other relevant aspects being aspects highlighted by several stakeholders are privacy, confidentiality, and intellectual property rights (European Commission, 2019, 2024; UNESCO, 2021). Unlike SPSS and JAMOVI, in which the dataset is stored in the software installed on the users' computers, Data Analyst GPT works in an online environment, and it is not very clear how this dataset is stored and how it will be used, if for intelligence training or to be included in a knowledge base. This jeopardises the use of Data Analyst GPT in research with confidential or sensitive data.

It is crucial to highlight specific variations of artificial intelligence technologies designed to ensure data privacy; an example includes ChatGPT Teams (Enterprise Privacy) and temporary chats (OpenAI, 2024), configured to ensure that dataset entered by users is not used for model training or inclusion in a knowledge base. Similarly, Microsoft's Copilot (Universidad de Granada, 2024), an AI-powered coding assistance tool, adheres to strict

[3] https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4

guidelines to protect user information, preventing the use of this dataset for enhancing machine learning algorithms.

In the European context, it is recommended that when using AI with sensitive or protected information, it is crucial to pay attention to privacy, confidentiality and intellectual property rights (European Commission, 2024). Researchers should protect unpublished or sensitive work by avoiding uploading it to online AI systems without warranty that the data will not be reused, such as when training future AI models or misusing the data. In addition, it is important not to provide third parties with personal data without the individual's explicit consent.

Despite these concerns, there is a significant advantage in analysing data provided by the Data Analyst GPT. Through the prompts, researchers and students have access to a tool that is easy to access and highly usable, making it easier to conduct data analysis with simple commands and direct language and process requests in natural language. That said, it has the potential to become an important ally in advancing research, stimulating innovation, and supporting the discovery of new insights in an intuitive and accessible way. Furthermore, it also can process large volumes of data without requiring advanced computing resources on the researcher's part. This is because the calculations are carried out on the artificial intelligence servers.

Conclusions

This article has demonstrated the potential of Data Analyst GPT in broadening the horizon of educational research, by showing its reliability in analysing quantitative data. Through a comparative analysis with standard statistical software, SPSS and JAMOVI, this article showed that Data Analyst GPT can be reliably employed as a statistical tool by educational researchers and students. Its user-friendly interface, which responds to simple commands and direct language, alleviates the need to master complex programming languages or have in-depth technical knowledge. This represents a significant advancement for conducting quantitative studies, making data analysis more accessible and less intimidating for education researchers.

A significant limitation of Data Analyst GPT is the lack of clear specification regarding the version of the tool being used, as it operates in a dynamic environment where the exact version may not be explicitly known, only the underlying model. Another important limitation

is the processing capacity of Data Analyst GPT, even in the paid version, ChatGPT Plus. The unexpected interruption and the need for a pause of approximately two hours before resuming analyses highlight a significant concern: the lack of clarity and transparency about usage limitations, especially for those who rely on the tool for continuous data analysis. Maintaining user trust requires clear and comprehensive communication from the developers regarding all operational aspects, including potential usage limits. This transparency is crucial to ensure that users can adequately plan their use of Data Analyst GPT, avoiding unpleasant surprises and aligning their availability expectations with the tool's operational reality. Additionally, the file upload limit of 50 MB, while sufficient for many quantitative data sets, could be a constraint in studies involving larger datasets.

The reliance on a tool that operates in an online environment also raises concerns about data privacy and confidentiality, as the details of how information is stored and used are not entirely transparent. Researchers should be cautious when using confidential or sensitive data with Data Analyst GPT, particularly in contexts where data security is critical.

The specific limitations of this study include the number of statistical tests conducted and the direct request approach for data analysis, specifying the desired tests. Additionally, the presentation of results generated by Data Analyst GPT may vary in future versions of the tool as improvements in the interface and visualization methods are implemented. However, the statistical results themselves, considering that Data Analyst GPT utilizes well-established libraries, should not undergo significant changes, ensuring the replicability and reliability of the results. It is important to note that this characteristic (use of libraries) is not exclusive to Data Analyst GPT; software like JAMOVI, which uses R libraries, also shares this consistency, though without the same graphical interface. In terms of user experience and data presentation, other traditional statistical software like SPSS and JAMOVI are also subject to updates that may impact these aspects.

REFERENCES

AL-GHONMEIN, A. M.; AL-MOGHRABI, K. G. The potential of ChatGPT technology in education : advantages , obstacles and future growth. **IAES International Journal of Artificial Intelligence (IJ-AI)**, Jacarta, v. 13, n. 2, p. 1206–1213, 2024. DOI: 10.11591/ijai.v13.i2.pp1206-1213.

ALGTHAMI, N. M. J.; HUSSIN, N. Meta-Analytic Evidence for Board Characteristics as Correlates of Firm Performance Among Saudi Arabian Businesses. **International Journal of Academic Research in Business and Social Sciences**, Islamabade, v. 12, n. 6, 4 jun. 2022. DOI: 10.6007/IJARBSS/v12-i6/13886.

ALILI, A.; KRSTEV, D. Using SPSS for research and Data Analysis. **Knowledge International Journal**, Escópia, v. 32, n. 3, p. 363–368, 26 jul. 2019. DOI: 10.35120/kij3203363a.

ALZUBAIDI, L.; ZHANG, J.; HUMAIDI, A. J.; AL-DUJAILI, A.; DUAN, Y.; AL-SHAMMA, O.; SANTAMARÍA, J.; FADHEL, M. A.; AL-AMIDIE, M.; FARHAN, L. Review of deep learning**:** concepts, CNN architectures, challenges, applications, future directions. **Springer International Publishing**, Berlin, v. 8, 2021. DOI: 10.1186/s40537-021-00444-8.

ATLAS, S. **ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI**. Kingston: University of Rhode Island, 2023. v. 1.

CHANG, D. T. Concept-Oriented Deep Learning with Large Language Models. **ArXiv**, [*S. l.*], 2023. DOI: 10.48550/arXiv.2306.17089.

CRAWFORD, J.; ALLEN, K.-A.; PANI, B.; COWLING, M. When artificial intelligence substitutes humans in higher education: the cost of loneliness, student success, and retention. **Studies in Higher Education**, London, v. 49, n. 5, p. 1–15, 2024. DOI: 10.1080/03075079.2024.2326956.

DANCEY, C. P.; REIDY, J. **Statistics without maths for psychology**. 8. ed. London: Prentice Hall, 2020.

DING, L.; LI, T.; JIANG, S.; GAPUD, A. Students' perceptions of using ChatGPT in a physics class as a virtual tutor. **International Journal of Educational Technology in Higher Education**, Barcelona, v. 20, n. 1, p. 1–18, 2023. DOI: 10.1186/s41239-023-00434-1.

EHLERS, U.-D.; LINDNER, M.; SOMMER, S.; RAUCH, E. AICOMP - Future Skills in a World Increasingly Shaped By AI. **Ubiquity Proceedings**, London, 2023. DOI: 10.5334/uproc.91.

EUROPEAN COMMISSION. Ethics guidelines for trustworthy AI. **European Commission**, Bruxelas, p. 1–39, 2019.

EUROPEAN COMMISSION. **Proposal for a Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial**

Intelligence Act) and amending certain union legislative acts. Brussels: European Commission, 2021.

EUROPEAN COMMISSION. **Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for Educators**. Brussels: European Commission, 2022.

EUROPEAN COMMISSION. **Living guidelines on the responsible use of generative AI in research**. Brussels: European Commission, 2024.

FAN, L.; LI, L.; MA, Z.; LEE, S.; YU, H.; HEMPHILL, L. A Bibliometric Review of Large Language Models Research from 2017 to 2023. ***ArXiv,*** [*S. l.*], p. 1–36, 2023. DOI: 10.48550/arXiv.2304.02020.

FIELD, A. **Discovering Statistics Using IBM SPSS Statistics**. 6. ed. London: SAGE Publications, 2024.

GIMPEL, H.; HALL, K.; DECKER, S.; LÄMMERMANN, L.; MÄDCHE, A.; RÖGLINGER, M.; RUINER, C.; SCHOCH, M.; SCHOOP, M.; URBACH, N.; VANDIRK, S. Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education. **Digital Annual Report**, Stuttgart, p. 1–54, 2023.

HUANG, Y.; WU, R.; HE, J.; XIANG, Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R. **Journal of global health**, New York, v. 14, n. 1088, p. 04070, 2024. DOI: 10.7189/jogh.14.04070.

IBM CORPORATION. **IBM SPSS Statistics Algorithms**. [*S. l: s. n.*], 2022.

JIA, X.-H.; TU, J.-C. Towards a New Conceptual Model of AI-Enhanced Learning for College Students: The Roles of Artificial Intelligence Capabilities, General Self-Efficacy, Learning Motivation, and Critical Thinking Awareness. **Systems**, [*S. l.*], v. 12, n. 3, p. 74, 2024.

LARY, D. J.; ALAVI, A. H.; GANDOMI, A. H.; WALKER, A. L. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, Beijing, v. 7, n. 1, p. 3–10, 2016. DOI: 10.1016/j.gsf.2015.07.003.

LI, J.; DADA, A.; PULADI, B.; KLEESIEK, J.; EGGER, J. ChatGPT in healthcare: A taxonomy and systematic review. **Computer Methods and Programs in Biomedicine**, Amsterdam, v. 245, p. 108013, 2024. DOI: 10.1016/j.cmpb.2024.108013.

MANN, H. B.; WHITNEY, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. **The Annals of Mathematical Statistics**, Washington, v. 18, n. 1, p. 50–60, mar. 1947. DOI: 10.1214/aoms/1177730491.

MAREK, J.; MAJ, E.; PRZYBYLA, O. K.; SKRZYNSKI, W.; PASICZ, K.; FABISZEWSKA, E.; PRUSZYNSKI, A.; ROWINSKI, O. The impact of studying on the hippocampal volume in medical students and its correlation with the results of the Final

Medical Examination: a single-centre, prospective observational cohort study. **Polish Journal of Radiology**, Warsaw, v. 88, p. 22–30, 16 jan. 2023. DOI: 10.5114/pjr.2023.124433.

MOHAMMADI, S. S.; NGUYEN, Q. D. A User-Friendly Approach for the Diagnosis of Diabetic Retinopathy Using ChatGPT and Automated Machine Learning. **Ophthalmology Science**, New York, v. 4, n. 4, p. 100495, 2024. DOI: 10.1016/j.xops.2024.100495.

MONTENEGRO-RUEDA, M.; LÓPEZ-MENESES, E.; FERNÁNDEZ-CERERO, J.; FERNÁNDEZ-BATANERO, J. M. Impact of the Implementation of ChatGPT in Education: A. **Computers**, Bern, v. 12, n. 153, p. 1–13, 2023. DOI: 10.3390/computers12080153.

NAVEED, H.; KHAN, A. U.; QIU, S.; SAQIB, M.; ANWAR, S.; USMAN, M.; AKHTAR, N.; BARNES, N.; MIAN, A. A Comprehensive Overview of Large Language Models. ***ArXiv***, p. 1–43, 12 jul. 2023. DOI: 10.48550/arXiv.2307.06435.

OPENAI. **Enterprise privacy at OpenAI**. Available at: https://openai.com/enterprise-privacy. Access: 25 Mar. 2024.

R CORE TEAM. **A Language and environment for statistical computing**. (Version 4.1) [Computer software], 2021.

SINGH-HARJIT; SINGH-AVNEET. ChatGPT: Systematic Review, Applications, and Agenda for Multidisciplinary Research. **Journal of Chinese Economic and Business Studies**, Washington, v. 21, n. 2, p. 193–212, 2023. DOI: 10.1080/14765284.2023.2210482.

STANFORD UNIVERSITY. **Artificial Intelligence Index Report 2021**. Stanford: Stanford University, 2021.

SUFI, F. Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. **Information**, Bern, v. 15, n. 2, p. 99, 2024. DOI: 10.3390/info15020099.

THE JAMOVI PROJECT. **Jamovi.** (Version 2.3) [Computer Software], 2024.

UNESCO. **Recommendation on the Ethics of Artificial Intelligence**. Paris: UNESCO. 2021.

UNESCO. **ChatGPT and Artificial Intelligence in Higher Education:** Quick start guide. Paris: UNESCO, 2023a.

UNESCO. **Harnessing the Era of Artificial Intelligence in Higher Education:** A Primer for Higher Education Stakeholders. Paris: UNESCO, 2023b.

UNIVERSIDAD DE GRANADA. **Inteligencia Artificial en la universidad**: Centro de Producción de Recursos para la Universidad Digital (CEPRUD). 2024. Available at: https://ceprud.ugr.es/formacion-tic/inteligencia-artificial. Access: 25 Mar. 2024.

WALTER, Y. Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. **International**

Journal of Educational Technology in Higher Education, Dublin, v. 21, n. 1, 2024. DOI: 10.1186/s41239-024-00448-3.

WILCOXON, F. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, Washington, v. 1, n. 6, p. 80, dez. 1945. DOI: 10.2307/3001968.

CRediT Author Statement

Acknowledgements: Acknowledgment to Professor Pedro Reis for his valuable contributions to the conceptualization of this article.

Funding: This work was supported by National Funds through FCT-Portuguese Foundation for Science and Technology, IP, under the scope of Unidade de Investigação e Desenvolvimento em Educação e Formação (UIDEF), UIDB/04107/2020, https://doi.org/10.54499/UIDB/04107/2020.

Conflicts of interest: The author declares no competing interests.

Ethical approval: This article does not require ethical approvals

Data and material availability: The anonymous datasets used and/or analyzed during the study and outputs of Data Analyst GTP, SPSS and JAMOVI are available in the Supplementary Information.

Authors' contributions: Sole authorship.

Processing and editing: Editora Ibero-Americana de Educação.
Proofreading, formatting, normalization and translation.