International Electronic Journal of Mathematics Education
2024, 19(4), em0795
e-ISSN: 1306-3030
https://www.iejme.com

MODESTUM

Research Article

OPEN ACCESS

ChatGPT's performance in university admissions tests in mathematics

Angel Udias [1,2*], Antonio Alonso-Ayuso [1], César Alfaro [1], María Jesús Algar [1], Marina Cuesta [1], Alberto Fernández-Isabel [1], Javier Gómez [1], Carmen Lancho [1], Emilio L. Cano [1], Isaac Martín de Diego [1], Felipe Ortega [1]

[1] DSLAB Research group, Universidad Rey Juan Carlos, Móstoles, Madrid, SPAIN
[2] European Commission, Joint Research Center, Ispra (VA), ITALY
***Corresponding Author:** angelluis.udias@urjc.es

Citation: Udias, A., Alonso-Ayuso, A., Alfaro, C., Algar, M. J., Cuesta, M., Fernández-Isabel, A., Gómez, J., Lancho, C., Cano, E. L., Martín de Diego, I., & Ortega, F. (2024). ChatGPT's performance in university admissions tests in mathematics. *International Electronic Journal of Mathematics Education, 19*(4), em0795. https://doi.org/10.29333/iejme/15517

ARTICLE INFO

Received: 01 Jan. 2024

Accepted: 26 Sep. 2024

ABSTRACT

This study comprehensively analyses the performance of the artificial intelligence (AI)-based language model, ChatGPT 4.0, in solving Spanish university admission tests in *applied mathematics in social sciences*. Using exams taken at public universities in Madrid, we have analysed ChatGPT's answers and concluded that its performance varies significantly across different areas of mathematics, excelling in probability and statistics exercises, but performing significantly worse in algebra and calculus. When compared with students, ChatGPT clearly outperforms them in all areas except algebra. Despite the model's limitations in interpreting complex mathematical ideas, in some cases its responses are positively surprising, indicating its potential as a valuable tool in certain mathematical problem-solving scenarios. Our results suggest significant potential for the introduction of these AI-based systems into the classroom. Despite the progress made, much remains to be explored regarding the efficient integration of chatbots into course development and the subsequent impact on education.

Keywords: ChatGPT 4.0, large language models, maths problems, university access exams, artificial intelligence in education

INTRODUCTION

The field of teaching has undergone significant changes in recent years with a shift towards digitalization and the use of computer tools (Hofer et al., 2021). This digital pivot, accelerated by the global pandemic, has led to the increased adoption of technology in education, from the development of audio-visual materials to automated grading exercises and videoconferencing (Giesbers et al., 2013). It is critical to note that while these developments have enhanced the learning experience, the importance of knowledge consolidation exercises, particularly in mathematics and statistics, cannot be overstated (Freeman et al., 2014).

The field of artificial intelligence (AI) integration in mathematics education is an emerging area of research, with a wide range of applications being explored, from robotics to more comprehensive approaches (Guan et al., 2020; Matzakos et al., 2023; Mohamed et al., 2022). The importance of deep understanding and reasoning in mathematical learning is being emphasized from a cognitive science perspective (Zhang et al., 2023). Similarly, Lu et al. (2022) conducted a survey on deep learning techniques for mathematical reasoning, highlighting the potential of these techniques in this field. Davis (2024) has recently discussed the intersection of mathematics, word problems, and AI. His work underscores the unique challenges and complexities present in this space, pointing to the need for further research and development. This discussion brings attention to the intricate relationship between these areas and the potential for AI to contribute to advancements in mathematics education.

Large language models (LLM), such as generative pre-trained transformers (GPT) introduced by OpenAI in 2018 (Radford et al., 2018), are among the AI tools that have garnered much attention. These models employ transformer architectures (Vaswani et al., 2017) that process natural language input and follow instructions through reinforcement learning based on human feedback (Ouyang et al., 2022). The evolution of these models has led to language models such as ChatGPT, which can engage in interactive text-based conversations with users, providing applications in education (Floridi & Chiriatti, 2020; Wollny et al., 2021).

ChatGPT has shown promising potential in various educational contexts, particularly in the field of mathematics. An analysis by Xuan-Quy et al. (2023) on the Vietnamese national high school graduation examination revealed that ChatGPT performed well on knowledge-based questions but faced difficulties with more intricate topics such as derivatives and spatial geometry. Frieder et al. (2023) conducted a comprehensive investigation into the mathematical capabilities of ChatGPT-3.5 and its successor, ChatGPT-4.0. Their findings indicated that these models could be valuable tools for querying facts and acting as mathematical search engines. However, they also underscored that the models' overall performance in advanced mathematics is still not on par with that of a graduate student.

In addition, Wardat et al. (2023) explored the perspectives of different stakeholders on the use of AI in mathematics education. They found a general enthusiasm for the potential of ChatGPT in teaching mathematics, but also uncovered some of its current limitations, such as a shallow understanding of geometry and a dependence on equation complexity.

In a comparative study, Plevris et al. (2023) found that while ChatGPT 4.0 outperformed ChatGPT 3.5 in basic arithmetic and logic puzzles, both models exhibited consistency issues, often providing conflicting answers to the same questions. These findings align with Wardat et al. (2023) observation of ChatGPT's limitations in handling complex mathematical problems. Chen et al. (2023) noted further variations in the performance of these models over time, emphasizing the need for their continuous monitoring. This point was also corroborated by Lo (2023), who highlighted variations in ChatGPT's performance across different domains and its potential issue of generating incorrect information. Udias et al.'s (2023) study on probability learning found that while ChatGPT 3.5 outperformed average students in problem formulation and logical reasoning, it struggled with basic numerical operations, paralleling the findings of Xuan-Quy et al. (2023) and Frieder et al. (2023).

Collectively, these studies highlight the potential of AI-based models such as ChatGPT in enhancing mathematics education, while simultaneously emphasizing the need for careful integration of these tools into educational settings, considering their various limitations and the need for continuous monitoring and updates. Nevertheless, the integration of AI models such as ChatGPT into education must be approached with caution, considering potential inaccuracies and ethical concerns (Holmes & Tuomi, 2022; Kasneci et al., 2023).

In addition, as these models become more integrated into educational settings, it is crucial to upgrade competencies to improve teachers' and students' proficiency with these technologies (Holmes et al., 2022).

In the realm of Spanish education, the university admission tests (EVAU, *evaluación para el acceso a la universidad*) hold significant importance. These non-compulsory exams serve as a gateway for students after completing secondary school, as they are a prerequisite for admission into universities. The EVAU is held in two sessions, ordinary and extraordinary (the latter for those who either did not reach the appropriate level in secondary education to take the ordinary exam, or who took the ordinary exam but did not obtain the minimum mark to pass it) and consists of several 90-minute written tests on different subjects over three days. These exams encompass both common and specific subjects studied during the last two years at the secondary school. It is worth noting that the EVAU tests ultimately give students access to the Spanish university system. The responsibility of the admission lies with the public universities of each autonomous community (region). The regional administration government acts as a co-ordinator

In accordance with national regulations, each autonomous community independently formulates the examination questions to be presented to students within its jurisdiction. These questions are carefully prepared and selected each year by committees of university professors and secondary school teachers. The rigorous examination process includes thorough review by multiple committee members to ensure accuracy and reliability. Strict security measures are maintained to ensure the confidentiality of the exams.

The aim of this research is to evaluate the capability of ChatGPT 4.0 to solve mathematics exercises included in various EVAU in Spain. It also seeks to compare the answers generated by the model with those given by a set of students in real situations. The second objective is to provide a qualitative assessment of the results obtained and to discuss the potential application of LLM in education at the university level. This investigation into ChatGPT's performance on university entrance mathematics tests is a step forward in understanding the potential and limitations of AI application in education, as well as its prospects for future research.

The rest of the paper is structured, as follows: We first provide a detailed description of the experiment, including the methodology used. We then explain the analysis of the results, and finally we present the conclusions and suggestions for future research.

METHODS

For this study, we have included exams administered in the community of Madrid. We have conducted a comprehensive comparative analysis of student performance on EVAU tests and the innovative AI-based language model, ChatGPT. ChatGPT 4.0 was chosen for this study, having just been released at the time of the research (April 2023), considering that its predecessor was the most widely utilized, and initial analyses indicated that the new version surpassed it in performance (Katz et al., 2024). Specifically, our focus is on the subject *applied mathematics in social sciences*, which is taken by students interested in studying social science and economics degrees. To carry out this research, we have selected and carefully examined the results of 5 past exams, which were held between 2009 and 2012, in both ordinary and extraordinary sessions.

The extraordinary session is exclusively for students who either didn't achieve the minimum grade in their final year of secondary school but succeeded later, or those who failed the regular exam despite attempting it.

All the exams in *applied mathematics in social sciences* have the same structure: four exercises, one from algebra (which accounts for 30% of the mark), one from calculus (30%), one from probability (20%) and one from statistics (20%) (see **Appendix A** for an example). In addition, each student is offered two different exams (option A and oprion B) with the same structure and must choose one of them. It is not possible to choose exercises from different options.

Table 1. Summary of question distribution across subject areas and the number of students who took the exam

Exam	Year	Session	Option	Number of questions				Number of students
				Algebra	Calculus	Statistics	Probability	
Ex1A	2009	Ordinary	A	3	3	2	2	56
Ex1B	2009	Ordinary	B	1	2	2	2	85
Ex2A	2010	Ordinary	A	1	3	2	2	35
Ex2B	2010	Ordinary	B	3	3	2	2	104
Ex3A	2011	Ordinary	A	3	3	2	2	83
Ex3B	2011	Ordinary	B	3	3	2	2	31
Ex4A	2011	Extraordinary	A	2	3	2	2	35
Ex4B	2011	Extraordinary	B	2	2	2	2	22
Ex5A	2012	Ordinary	A	3	2	2	2	120
Ex5B	2012	Ordinary	B	1	3	4	2	37
TOTAL				22	27	22	20	608

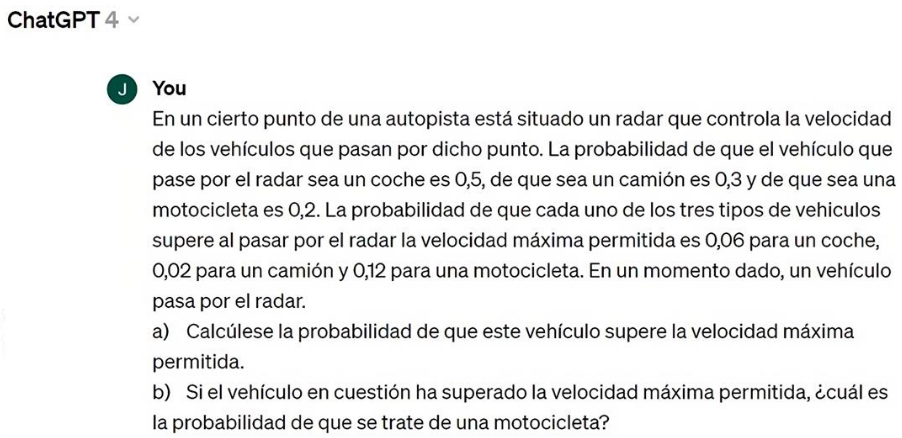


Figure 1. Verbal expression of the probability question from the June 2011 exam B, as an example of such a question as it was presented to ChatGPT (Source: The author's statement of the question, using OpenAI, 2023)

Within each area, there are different types of exercises. In algebra, these include linear optimization, systems of equations, and matrices. In calculus, most of the exercises focus on functions and their properties (continuity, derivability, optima, etc.) and on defined integrals, but they can also consist of constructing a function that represents a realistic situation and finding its optima. As for statistics, exercises focus on confidence intervals and the distribution of means, while probability exercises deal with the Laplace rule, Bayes' theorem, conditional probability, and events. Each exercise is composed of a variable number of questions, whose weights within the assessment were originally defined. The 40 exercises considered in this study contain a total of 91 questions. **Table 1** displays the number of questions in the corresponding exercise for each area across the ten different exams.

Data Collection Proces

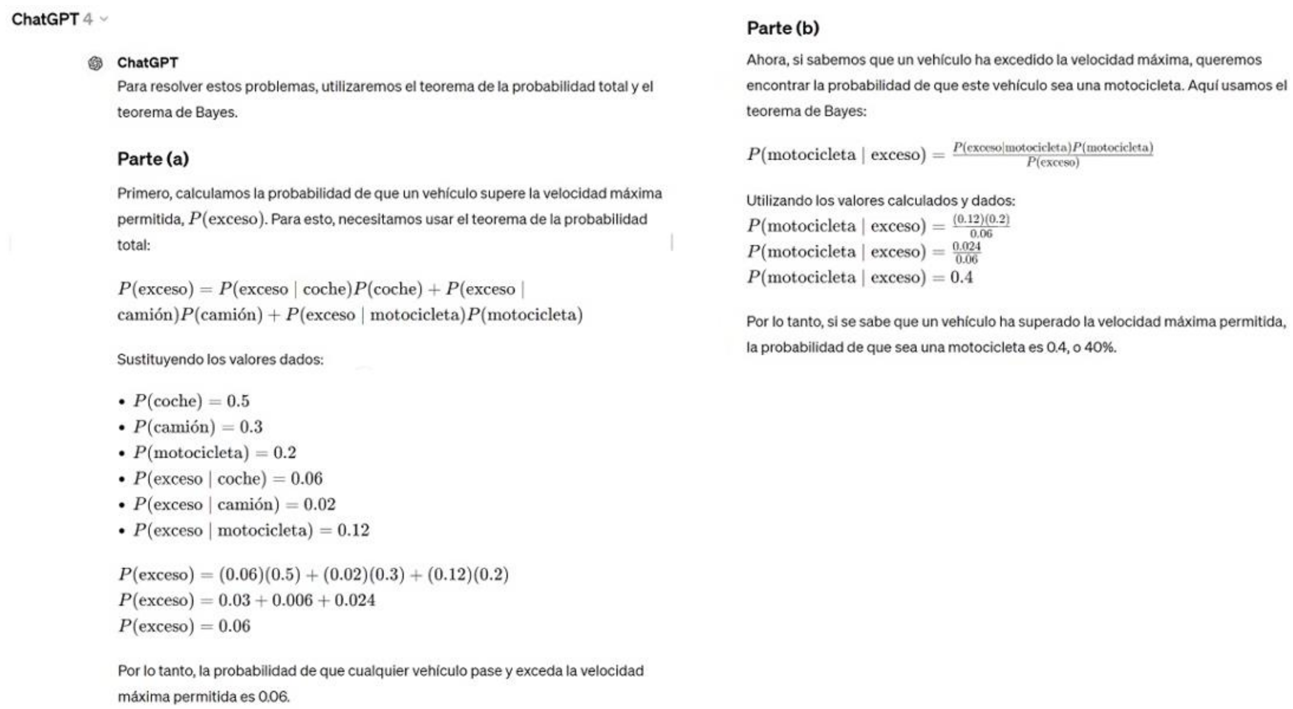
For these 5 exams we have information about the answers given by a set of students: which option (A or B) they chose in the exam and what grade they obtained for each of the questions. The last column of **Table 1** shows the number of students for whom we have information. As can be seen, we have data from 608 students and the number of student in each group varies between 22 (Ex4B exam) and 120 (Ex5B exam), being sufficiently high in all cases.

The 40 exercises from the 10 exams were presented to ChatGPT 4.0 (March version) via the web interface (https://chat.openai.com/chat), and the answers were recorded. The exercises were provided to the GPT in their original form (see **Appendix A**), as written by the lecturers in Spanish, without any modifications or clarifications and without the addition of any prompts. The answers given by the ChatGPT were also in Spanish.

The exercises were presented to ChatGPT exactly as they were originally formulated in Spanish (example in **Figure 1**), and the wording (in Spanish) for all the exercises can be found in the supplementary materials.

It should be noted that ChatGPT does not always generate the same response to a given question. To reduce the impact of this variability on the response, all exercises were presented to ChatGPT five times on different days between 5 and 15 of May 2023, always by the same person using the same account. At the end of this phase, 455 (91 questions × 5 repetitions) answers were available. **Figure 2** shows one of the answers given by the ChatGPT to the question proposed in **Figure 1**.

The next phase of the data collection process was to evaluate all the ChatGPT responses. Nine university teachers, all with extensive experience in teaching mathematics (some of whom were even part of the committees that evaluated the exercises in the community of Madrid), participated in this phase. Each exercise underwent independent evaluation by three distinct educators, seeking an unbiased grading of the answers. The teachers were provided with the grading rubric that was used to evaluate the students' entrance exams in order to standardize the grading process. This rubric encompassed various facets, including exemplar solutions for each exercise type and delineated the relative weight assigned to every question within an exam (see **Appendix A** for an example). Consequently, this approach aimed to mitigate biases and enhance impartiality in the



ChatGPT 4

ChatGPT

Para resolver estos problemas, utilizaremos el teorema de la probabilidad total y el teorema de Bayes.

Parte (a)

Primero, calculamos la probabilidad de que un vehículo supere la velocidad máxima permitida, $P(\text{exceso})$. Para esto, necesitamos usar el teorema de la probabilidad total:

$$P(\text{exceso}) = P(\text{exceso} \mid \text{coche})P(\text{coche}) + P(\text{exceso} \mid \text{camión})P(\text{camión}) + P(\text{exceso} \mid \text{motocicleta})P(\text{motocicleta})$$

Sustituyendo los valores dados:

- $P(\text{coche}) = 0.5$
- $P(\text{camión}) = 0.3$
- $P(\text{motocicleta}) = 0.2$
- $P(\text{exceso} \mid \text{coche}) = 0.06$
- $P(\text{exceso} \mid \text{camión}) = 0.02$
- $P(\text{exceso} \mid \text{motocicleta}) = 0.12$

$$P(\text{exceso}) = (0.06)(0.5) + (0.02)(0.3) + (0.12)(0.2)$$
$$P(\text{exceso}) = 0.03 + 0.006 + 0.024$$
$$P(\text{exceso}) = 0.06$$

Por lo tanto, la probabilidad de que cualquier vehículo pase y exceda la velocidad máxima permitida es 0.06.

Parte (b)

Ahora, si sabemos que un vehículo ha excedido la velocidad máxima, queremos encontrar la probabilidad de que este vehículo sea una motocicleta. Aquí usamos el teorema de Bayes:

$$P(\text{motocicleta} \mid \text{exceso}) = \frac{P(\text{exceso}|\text{motocicleta})P(\text{motocicleta})}{P(\text{exceso})}$$

Utilizando los valores calculados y dados:

$$P(\text{motocicleta} \mid \text{exceso}) = \frac{(0.12)(0.2)}{0.06}$$
$$P(\text{motocicleta} \mid \text{exceso}) = \frac{0.024}{0.06}$$
$$P(\text{motocicleta} \mid \text{exceso}) = 0.4$$

Por lo tanto, si se sabe que un vehículo ha superado la velocidad máxima permitida, la probabilidad de que sea una motocicleta es 0.4, o 40%.

Figure 2. Screenshot with one of the answers given by the ChatGPT to the proposed question in **Figure 1** (Source: ChatGPT personal communication, OpenAI, 2023)

Table 2. ANOVA and Kruskal Wallis results for assessing agreement among three professors' evaluations of the five responses to each exam question

Exam berto	Num Q	p-values	
		ANOVA	Kruskal Wallis
Ex1A	4 × 5	0.709	0.4425
Ex1B	4 × 5	0.993	0.9022
Ex2A	4 × 5	0.829	0.8130
Ex2B	4 × 5	0.496	0.4986
Ex3A	4 × 5	0.393	0.7955
Ex3B	4 × 5	0.177	0.2928
Ex4A	4 × 5	0.083	0.1209
Ex4B	4 × 5	0.232	0.4990
Ex5A	4 × 5	0.900	0.9757
Ex5B	4 × 5	0.846	0.9138

assessment process among the diverse evaluators of ChatGPT's responses. Furthermore, it facilitated a more equitable comparison with contemporaneous evaluations of student responses, a crucial consideration in this study's analysis.

In addition to assigning a final mark for the exercises, they were also asked to provide additional comments. These comments included the reason for the main penalty they assigned to the answer (e.g., identification of the problem, explanation of the reasoning, calculations), their perception of whether the answer was given by a human or not, and, in the case of multiple question exercises, which question was penalised the most.

Finally, numerical analyses were conducted on the scores obtained by ChatGPT in each examination and area, juxtaposed with the scores attained by students in corresponding exercises. These analyses aimed to establish comparative insights between the AI-generated responses and those provided by students within identical exercises, contributing to a comprehensive evaluation of ChatGPT's performance across distinct academic domains.

Reliability of the Evaluation Process

The first step involved assessing the consistency among the different professors in evaluating ChatGPT's responses to the exercises. The aim of this statistical analysis was to ascertain agreement or disagreement in the teachers' evaluations, thus validating the reliability of the evaluation process. An analysis of variance (ANOVA) and a Kruskal-Wallis test (Ugarte et al., 2015) were performed to compare the scores given by three professors for each of the five repetitions of each exercise within a given exam. For each one of the 10 exams, 60 marks are available (3 teachers per exam, 4 exercises per exam, and 5 answers per exercise) and the teacher is used as factor. The results are presented in **Table 2**, where the p-value for both tests are reported

The significant difference of means per teacher in the grading is reflected in the p-values obtained. Lower p-values denoted more significant differences between the gradings. Both, the ANOVA and the Kruskall-Wallis test results showed a high level of agreement between the teachers' gradings in all the exams, confirming the reliability and credibility of the grading procedure.

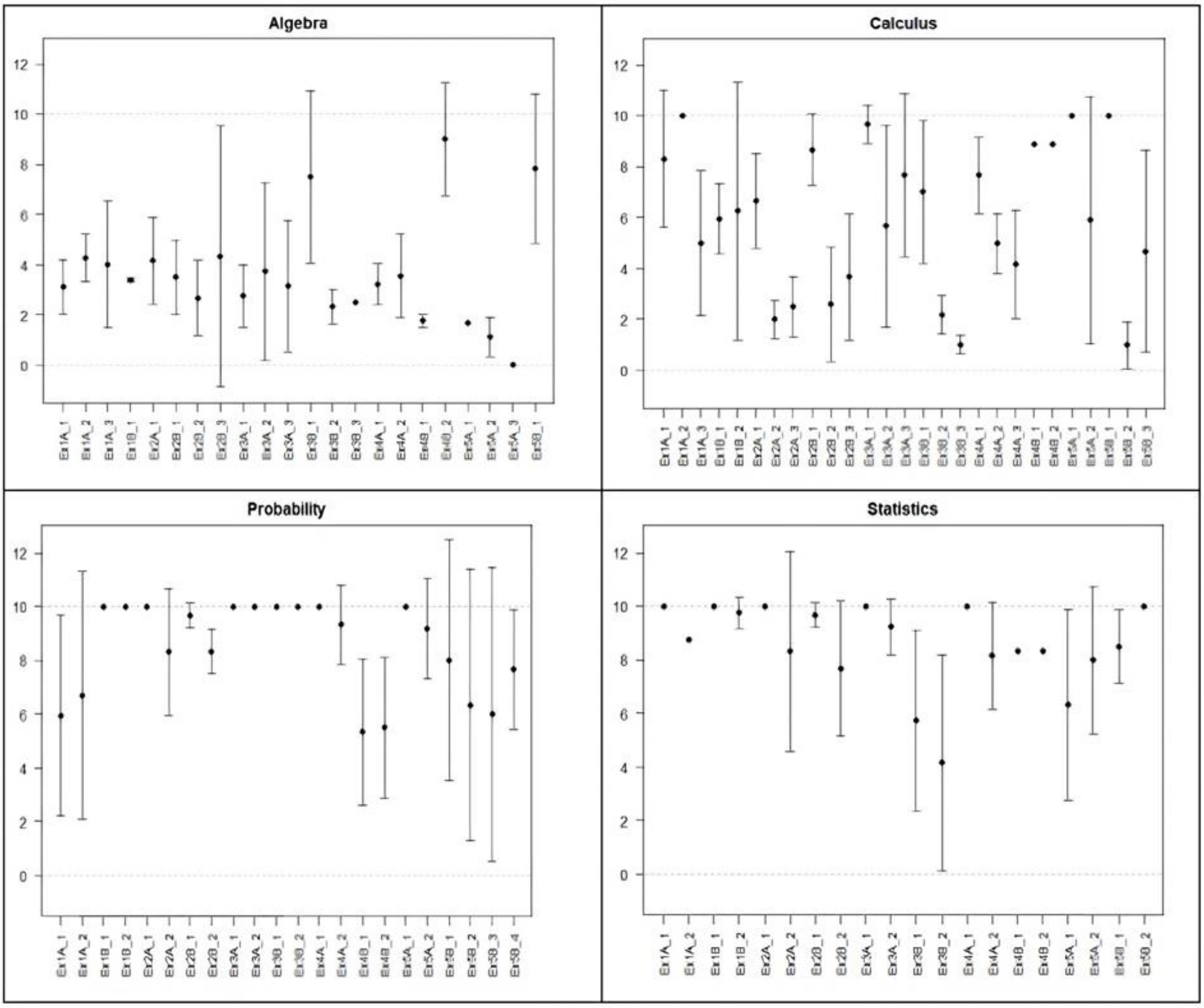


Figure 3. Graphical analysis examining ChatGPT's consistency across five responses per posed question (Source: Authors' own elaboration)

Only Ex4A showed a p-value nearing the critical value for hypothesis test rejection (α = 0.05), indicating a potential point of ambiguity. However, such results might be due to inherent variations among evaluators rather than a systemic issue (Cohen, 1992; McDonald, 2014).

Upon finding no significant score differences among the three teachers, the mean score given by them for each question's answer was adopted as ChatGPT's grade. This consistent approach formed the basis for subsequent analyses, facilitating a comprehensive assessment of ChatGPT's performance across diverse exercises.

DATA ANALYSIS

ChatGPT's Performance and Consistency Across Responses

In the subsequent analysis phase, the consistency of the five answers provided by ChatGPT for the 91 questions was examined. **Figure 3** shows the results grouped by subject area: algebra, calculus, probability, and statistics. In each of the four graphs, each mark corresponds to a question and shows the mean of the marks given by teachers to the five ChatGPT answers, accompanied by an error bar representing the standard deviation as an indicator of the variability between them

Figure 3 shows that the similarity between the five ChatGPT answers for the same question is higher in the area of statistics (the five answers scored the same in 9 out of 20 questions, 45%) and in probability (in 9 out of 22 questions, 41%). Conversely, responses related to algebra and calculus displayed lesser consistency (only in 3 out of 22 answers scored the same in algebra, 14%, and 5 out of 26 in calculus 19%). Moreover, in calculus, probability, and statistics, almost all of these identically-scored answers are the maximum score, while at the other extreme are the equally-scored answers in algebra, where in the 3 cases the score was below 3. Furthermore, it can already be seen that the quality of the ChatGPT answers is not the same in all areas: in algebra the average scores are mostly below 5, in calculus there is a lot of variability, with about half of the average scores below

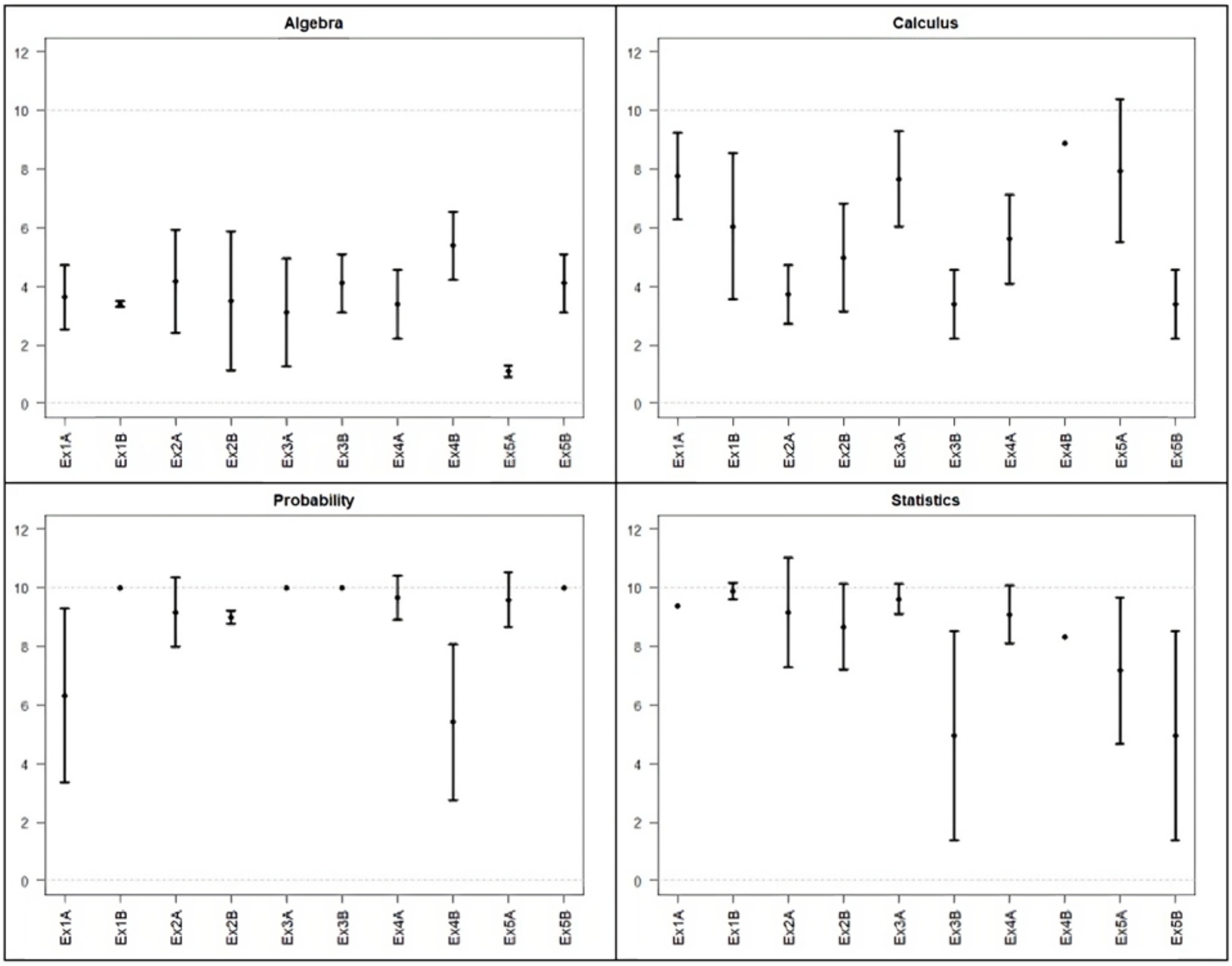


Figure 4. Graphical analysis examining ChatGPT's consistency across five responses per posed exercise (Source: Authors' own elaboration)

Table 3. Mean, median, and mean of the standard deviation (SD) of the scores obtained by the five ChatGPT answers for all questions and exercise in subject area

	Questions			Exercise		
	Mean	Median	Mean (SD)	Mean	Median	Mean (SD)
Algebra	3.99	3.22	1.65	3.96	3.57	1.36
Calculus	5,84	6.40	2.00	6.12	5.83	1.52
Probability	8.41	9.50	1.82	8.62	9.37	1.14
Statistics	8.74	10.00	1.24	8.55	9.12	1.85

5 and the other half above 10. However, in probability the average score is above 5 for all questions and in statistics for all but one question. **Figure 4** shows the same data for the exercises, also grouped by subject area.

Analyzing **Figure 4**, it can be observed that in the area of statistics, ChatGPT demonstrates consistently high average scores (exceeding 9 points) with minimal variability in 8 out of the 10 exams. However, in Ex3B (which corresponds to the option B of the ordinary exam in 2011), its performance is notably lower, obtaining a score of 5. The situation is similar in the domain of probability, with 7 out of 10 questions receiving scores higher than 9 and showing little variability. On the contrary, in the area of algebra, ChatGPT's scores are lower, with 9 out of 10 exams receiving scores below 5 points, also exhibiting relatively contained variability.

The consistency between the scores of the five ChatGPT answers is quite dependent on each area and on whether the study is done considering the exercises total score (by exercise in **Table 3**) or with the detail of the questions of each exercise (by question exam in **Table 3**). This is especially evident in the area of statistics where the mean deviations per question exams and per exercise are the lowest and highest, respectively in the cross-area comparison (see **Table 3**).

Student's Performance

The number of answers we have for each exam varies because the number of students taking the EVAU (and choosing mathematics applied to social sciences) varies from year to year; moreover, the choice of option (A or B) is free and is made by the student during the exam. Therefore, there are exams for which we have only 22 answers and others for which we have up to 120

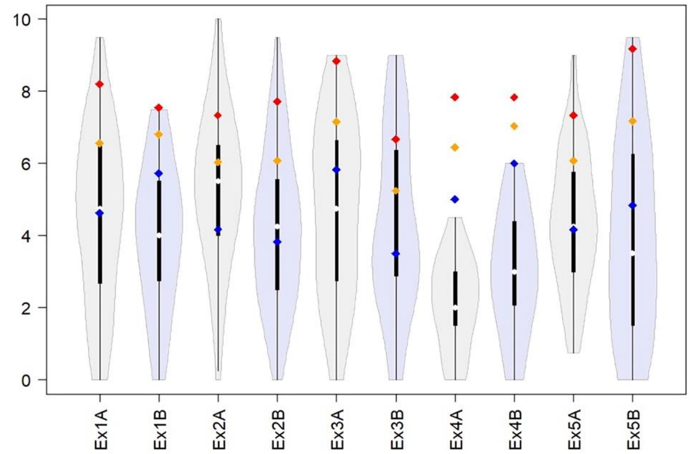


Figure 5. Violin plots with the distribution of students' scores per exam (in addition, the GPT scores are included; orange diamond: mean of responses; red/blue diamonds: highest/lowest scores out of 5 responses per question) (Source: Authors' own elaboration)

answers (as shown in **Table 1**). In any case, the sample is sufficiently representative and homogeneous (they are students from the same geographical area, with similar socio-economic and academic levels). It is therefore expected that the results for each problem reflect the difficulty of the exercise.

Next, the results obtained by ChatGPT are compared with the results obtained by the groups of students who took the EVAU at that time. **Figure 5** shows the distribution of students' scores for each exercise in the form of a violin plot (with a box plot inside). Violin plots provide a more detailed representation than traditional box plots of the distribution of student scores when the distributions are multimodal and have many observations at the extremes (Hintze & Nelson, 1998). The white dot corresponds to the median of the distribution, the thick black segment represents the interquartile range, and the thin segment represents the range. In addition to the distribution of student scores, the mean of the ChatGPT answer scores is presented as an orange diamond. The red and blue diamonds represent the highest and lowest scores from the five responses for each question, respectively.

The overall distribution of students' scores across the exams does not show any remarkable differences, except for the exams corresponding to the extraordinary sessions in 2011 (Ex4A and Ex4B), where students achieve significantly lower scores. However, this result is logical because ChatGPT shows the same behaviour in all the exams, whereas the students who take the extraordinary exam are students who either did not reach the appropriate level in secondary education to take the ordinary exam, or who took the ordinary exam but did not obtain the minimum mark to pass it. For the ordinary exams, median scores are consistently in the range of 4 to 6 points, with slight variations in the distribution (**Figure 5**). In particular, the average ChatGPT score (represented by the orange diamond in **Figure 5**) is consistently higher than the median student score, and even the lowest ChatGPT score (minimum of the sums of 5 answers for each exercise, blue diamond) is higher than the median student score in half of the exams.

Figure 6 shows the distribution of students' scores for each of the 40 questions, grouped by subject area. It can be seen that there are large differences between the subject areas. Clearly, the worst results are obtained in calculus, where the difference with ChatGPT is more pronounced, while in algebra, where ChatGPT shows lower performance, students obtain comparatively higher scores.

In the field of statistics, both the students and ChatGPT show consistent behavior, with similar performances. An exception can be observed in the exercise corresponding to the extraordinary call of 2011, where ChatGPT obtains a very good result, while the students do not. This is because, unlike other tests of that year, the statistics question did not require calculating a confidence interval, but rather focused on the distribution of the sample mean.

In probability, students obtain the lowest results in exercises that require applying the Laplace rule and calculating some conditional probabilities. However, they achieve better results in seemingly more complex exercises that involve applying Bayes' theorem. This may be because these latter problems are more common, and students prepare for them intensively. Furthermore, they usually have similar structured statements.

Based on probability exercise results, ChatGPT 4.0 significantly outperforms version 3.5, as studied by Udias et al. (2023). Both this work and Udias et al.'s (2023) included Spanish-written university level probability exercises. Unlike Udias et al.'s (2023) study, which reported numerical operation issues in version 3.5, this study found no such problems. In fact, ChatGPT 4.0 consistently delivered excellent solutions for most probability exercises, demonstrating enhanced proficiency in numerical operations. In calculus, students achieve the worst result in exercises 2009 ordinary B, 2011 extraordinary B, and 2012 ordinary A. In the last two exams, the proposed exercises had a different nature than usual: instead of providing an explicit function and asking for analysis (optimum, asymptotes, continuity, etc.), students had to construct the function representing the given situation and find the maximum. The 2009 ordinary B exam was of this type, but the function depended on a parameter and required a more detailed analysis than what was requested.

In algebra exercises, the highest level of homogeneity is observed among the responses of the students, ChatGPT, and between the students and ChatGPT.

This disparity is visually evident in the scatter plot presented in **Figure 7**, where the scores are plotted by subject area for both ChatGPT and students. Points below the diagonal correspond to exercises where students score better than ChatGPT, the reverse

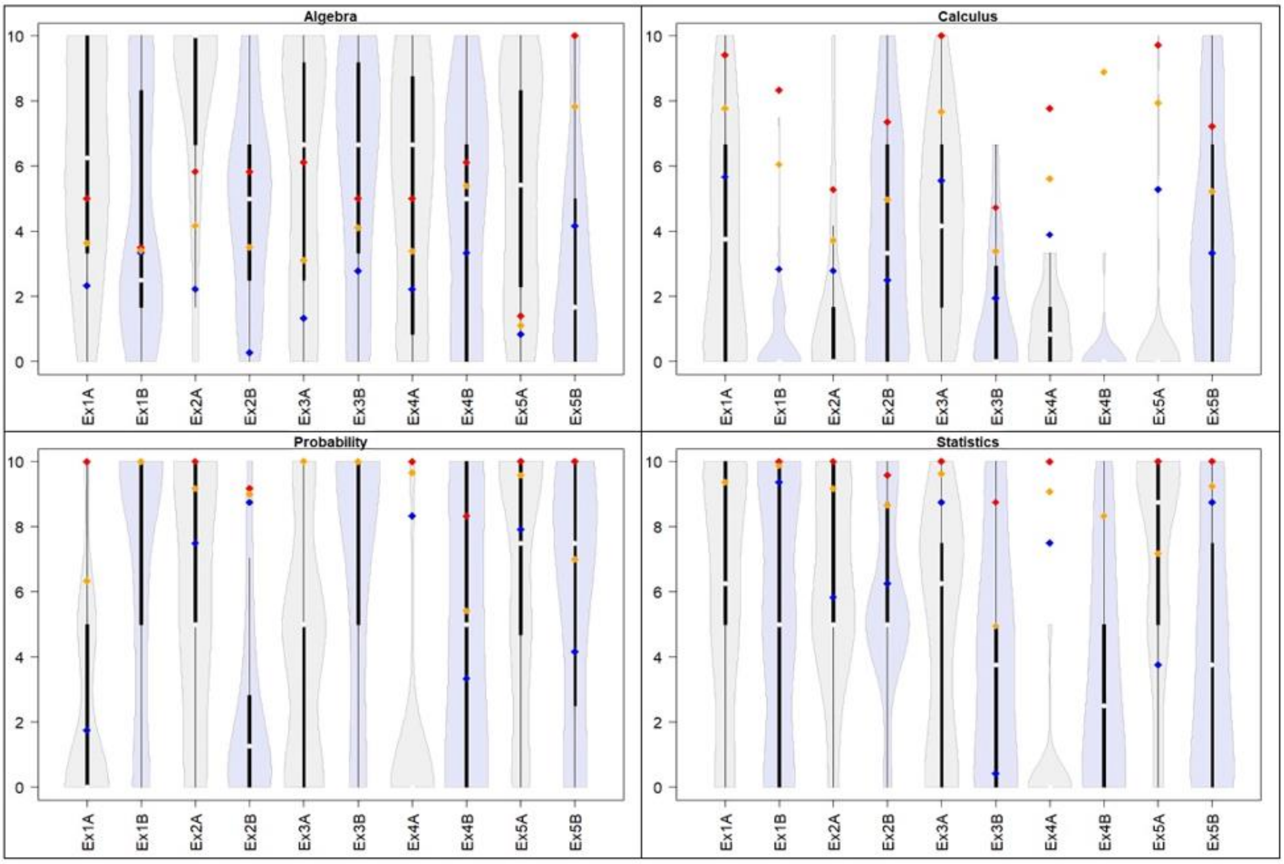


Figure 6. Violin plots with the distribution of student marks per exercise (in addition, the GPT scores are included; orange diamond: mean of responses; red/blue diamonds: highest/lowest scores out of 5 responses per question) (Source: Authors' own elaboration)

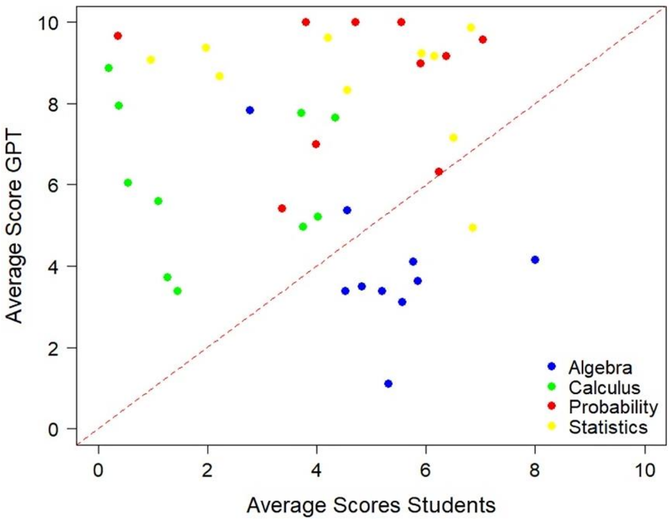


Figure 7. Scatter plot comparing student and ChatGPT scores for 40 exam questions (Source: Authors' own elaboration)

being true for points above the diagonal. It is evident that students outperform ChatGPT only in the algebra area, while in the other subject areas, ChatGPT consistently achieves higher scores.

Table 4 summarises the results of the paired t-tests between the ChatGPT and Student groups, including individual comparisons for each of the 10 exam score areas and an overall score across all exams. **Table 4** shows the statistical results, including corresponding p-values, confidence intervals and mean differences. These results highlight differences not only in the four specific subject areas, but also in the overall exam scores. While the comparison in algebra shows a non-significant difference (p-value = 0.1624), indicating a slightly worse performance in ChatGPT scores, substantial differences significantly favour the ChatGPT in calculus, probability, and statistics. In addition, the overall assessment of the total examination results emphasises a significant advantage for the ChatGPT results.

Table 4. Summary of paired t-test results comparing ChatGPT and students' scores per subject areas and exam

Area	p-value	Interval	Mean difference
Algebra	0.16242	(-3.15, 0.62)	-1.27
Calculus	0.00077	(2.20, 5.90)	4.05
Probability	0.00097	(2.06, 5.73)	3.89
Statistics	0.00287	(1.74, 6.13)	3.94
Total	0.00013	(1.55, 3.25)	2.40

These findings demonstrate that ChatGPT consistently outperforms the median student grades across various exam scenarios. The performance gap between ChatGPT and students varies across subject areas, with students excelling in algebra, an area where ChatGPT shows comparatively lower performance. Overall, the results shed light on the contrasting performance patterns between students and ChatGPT in different exam contexts, providing valuable insights into the capabilities and limitations of the AI-based system in educational assessment.

In addition to analysing the mathematical results, the formal quality of the ChatGPT answers was also assessed. To do this, the teachers who evaluated the tests were required to answer two questions: firstly, whether the answer appeared to have been written by a human, and secondly, whether the level of explanation was superior or inferior to that which a student would provide. In both cases, the general opinion of the teachers is that ChatGPT's explanations are much more detailed than those usually given by students (it generally justifies and develops in detail all the steps it takes and correctly establishes the notation used). Its style can be considered to resemble that of a human author; however, the level of development is higher than that demonstrated by the students, making it now relatively straightforward to distinguish ChatGPT's responses from those typically produced by students.

DISCUSSION AND CONCLUSIONS

This study evaluated the AI-based language model, ChatGPT 4.0, focusing on its potential as an educational resource for solving *applied mathematics in social sciences* for University Admission Tests in Spain. The model's performance was compared with that of students who took the same exams, providing valuable insights into its capabilities and areas for improvement.

When comparing the performance of ChatGPT with student groups who took the same exams, ChatGPT's mean scores consistently surpassed the median student grades. Our findings reveal that ChatGPT consistently outperforms the average student grades across most exams, with notable competency in the areas of probability and statistics. However, its performance in algebra and calculus was comparatively lower, highlighting areas where enhancements are needed in the model's ability to interpret and solve complex mathematical problems. This variability underscores the need for ongoing improvements in the model's ability to interpret and solve more complex mathematical problems.

ChatGPT's responses exhibited consistency in statistics and probability, while demonstrating greater variability in algebra and calculus. This suggests that the model's effectiveness may vary depending on the specific area of mathematics, and whilst it generally provides accurate mathematical responses, there is a risk of generating incorrect or misleading information.

Notice that Korkmaz Guler et al. (2023) conclude that ChatGPT gives poor results in probability exercises. However, the design of the experiment, the type of problems and the language to be used make the results not very comparable with ours. Both, this discrepancy and the variability in responses shown in the previous sections make it clear that ChatGPT 4.0 is not flawless in solving mathematical problems.

The introduction of advanced language models such as ChatGPT into educational settings brings both opportunities and challenges. A prominent concern is the possibility of inaccuracies (Borji, 2023), wherein the models might inadvertently present compelling yet false information to users. Additionally, biases present in the model's output could perpetuate and even exacerbate existing societal inequalities, raising ethical concerns (Barikeri et al., 2021; Shahriar & Hayawi, 2022).

The transition highlights the need for both educators and learners to develop the necessary competencies and literacies that enable them to comprehend the technology and its limitations. They also stressed the importance of a clear pedagogical approach with a strong focus on critical thinking and fact-checking strategies, as well as the need for continuous human oversight and awareness of potential misuse of LLM (Bender & Friedman, 2020). Despite these challenges, LLM can offer insights and opportunities in educational scenarios. In addition to the mathematical results, we also assessed the formal quality of ChatGPT's answers. Teachers who evaluated the tests generally believed that ChatGPT's explanations were much more detailed than those typically given by students. They also stated that its style could easily pass for that of a human. However, they noted that the level of development in the responses was higher than that given by most students, making it relatively easy to distinguish ChatGPT's answers from those given by students.

In conclusion, this study provides valuable insights into the capabilities and limitations of AI-based systems such as ChatGPT in educational assessment. Despite promising results, the model needs further exploration, particularly in areas requiring visual reasoning or graphical data interpretation. As AI systems continue to evolve, future studies should aim to address these limitations and improve these models' problem-solving capacity in mathematics. Moreover, specifically in probability exercises, ChatGPT 4.0 demonstrated substantial improvement over version 3.5, consistently providing excellent solutions without any difficulty with numerical operations. The potential of ChatGPT to enhance mathematics education is noteworthy, but its integration into education should be approached with caution, considering potential inaccuracies, ethical concerns, and the need for continuous human oversight.

Finally, it is important to note that this study's findings are based on Spanish-language data from a specific time period, and results may vary with different prompts or languages.

Author contributions: AU & AA-A: conceptualization, methodology, visualization, writing – original draft; **CA:** data curation; **MC:** visualization; **JG:** data curation; **ELC:** Validation. All authors are involved in investigation and writing – review & editing. All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that the study does not require any ethical approval. The authors further stated that the information used comes from freely available data in the public domain (university websites), and the analysis of datasets, are either open source, from the data collection or obtained from other researchers.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021). RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1941-1955). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.151

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6,* 587-604. https://doi.org/10.1162/tacl_a_00041

Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv*. https://doi.org/10.48550/arXiv.2302.03494

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? *arXiv*. https://doi.org/10.48550/arXiv.2307.09009

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Davis, E. (2024). Mathematics, word problems, common sense, and artificial intelligence. *Bulletin of the American Mathematical Society, 61*(3), 287-303. https://doi.org/10.1090/bull/1828

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111

Frieder, S., Pinchetti, L., , Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems 36* (pp. 27699-27744). Curran Associates, Inc.

Giesbers, B., Rienties, B., Tempelaar, D., & Gijselaers, W. (2013). Investigating the relations between motivation, tool use, participation, and performance in an e-learning course using web-videoconferencing. *Computers in Human Behavior, 29*(1), 285-292. https://doi.org/10.1016/j.chb.2012.09.005

Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *International Journal of Innovation Studies, 4*(4), 134-147. https://doi.org/10.1016/j.ijis.2020.09.001

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician, 52*(2), 181-184. https://doi.org/10.2307/2685478

Hofer, S. I., Nistor, N., & Scheibenzuber, C. (2021). Online teaching and learning in higher education: Lessons learned in crisis situations. *Computers in Human Behavior, 121*, Article 106789. https://doi.org/10.1016/j.chb.2021.106789

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57*(4), 542-570. https://doi.org/10.1111/ejed.12533

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S., Santos, O., Rodrigo, M., Cukurova, M., Bittencourt, I., & Koedinger, K. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*, 504-526. https://doi.org/10.1007/s40593-021-00239-1

Kasneci, E., Seßler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., …, & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, 382*(2270), Article 20230254. https://doi.org/10.1098/rsta.2023.0254

Korkmaz Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research, 9*(2), Article em0188. https://doi.org/10.29333/pr/14145

Lo, C. K. (2023). What is the impact of chatgpt on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Lu, P., Qiu, L., Yu, W., Welleck, S., & Chang, K. W. (2022). A survey of deep learning for mathematical reasoning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14605-14631). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.817

Matzakos, N., Doukakis, S., & Moundridou, M. (2023). Learning mathematics with large language models: A comparative study with computer algebra systems and other tools. *International Journal of Emerging Technologies in Learning, 18*(20), 51-71. https://doi.org/10.3991/ijet.v18i20.42979

McDonald, J. H. (2014). *Handbook of biological statistics* (3rd ed.). Sparky House Publishing.

Mohamed, M. Z. B., Hidayat, R., Suhaizi, N. N. B., Sabri, N. B. M., Mahmud, M. K. H. B., & Baharuddin, S. N. B. (2022). Artificial intelligence in mathematics education: A systematic literature review. *International Electronic Journal of Mathematics Education, 17*(3), Article em0694. https://doi.org/10.29333/iejme/12132

OpenAI. (2023). *ChatGPT (Mar 14 version)* [Large language model]. https://chat.openai.com/chat

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems 35* (pp. 27730-27744). Curran Associates, Inc.

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI, 4*(4), 949-969. https://doi.org/10.3390/ai4040048

Radford, A., Narasimhan, K., Salimans, T.,Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Shahriar, S., & Hayawi, K. (2023). Let's have a chat! A conversation with ChatGPT: Technology, applications, and limitations. *Artificial Intelligence and Applications, 2(1),* 11-20. https://doi.org/10.47852/bonviewAIA3202939

Udias, A., Alonso-Ayuso, A., Sanchez, I., Hernandez, S., Castellanos, M. E., Diez, R. M., & Cano, E. L. (2023). The potential of large language models for improving probability learning: A study on ChatGPT3.5 and first-year computer engineering students. *arXiv*. https://doi.org/10.48550/arXiv.2310.05686

Ugarte, M. D., Militino, A. F., & Arhholt A. T. (2015). *Probability and statistics with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18682

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.) *Advances in neural information processing systems 30*. Curran Associates, Inc.

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7), Article em2286. https://doi.org/10.29333/ejmste/13272

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence, 4*, Article 654924. https://doi.org/10.3389/frai.2021.654924

Xuan-Quy, D., Ngoc-Bich, L., The-Duy, V., Xuan-Dung, P., Bac-Bien, N., Van-Tien, N., Thi-My-Thanh, N., & Hong-Phuoc, N. (2023). VNHSGE: VietNamese high school graduation examination dataset for large language models. *arXiv*. https://doi.org/10.48550/arXiv.2305.12199

Zhang, C. E., Collins, K. M., Weller, A., & Tenenbaum, J. B. (2023). AI for mathematics: A cognitive science perspective. *arXiv*. https://doi.org/10.48550/arXiv.2310.13021

APPENDIX A: EXAMPLE OF ONE EXAM (EX5A)

Questions

OPCIÓN A

Ejercicio A1

a)

$$\left(\begin{array}{ccc|c} 1 & a & -7 & 4a-1 \\ 1 & 1+a & -(a+6) & 3a+1 \\ 0 & a & -6 & 3a-2 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & a & -7 & 4a-1 \\ 0 & 1 & 1-a & 2-a \\ 0 & 0 & (a+2)(a-3) & (a+2)(a-1) \end{array}\right)$$

1. $a \neq -2, 3$. El sistema es COMPATIBLE DETERMINADO.
2. $a = 3$. El sistema es INCOMPATIBLE.
3. $a = -2$. SISTEMA COMPATIBLE INDETERMINADO.

b) Para $a = -2$ el sistema que debemos resolver es:

$$\begin{cases} x - 2y - 7z = -9 \\ y + 3z = 4 \end{cases} \to x = \lambda - 1, \quad y = 4 - 3\lambda, \quad z = \lambda \quad \text{con } \lambda \in \mathbb{R}.$$

c) Para $a = -3$ el sistema que debemos resolver es:

$$\begin{cases} x - 3y - 7z = -13 \\ x - 2y - 3z = -8 \\ -3y - 6z = -11 \end{cases} \to x = -\frac{4}{3}, \quad y = \frac{7}{3}, \quad z = \frac{2}{3}.$$

Ejercicio A2

Sea x el número de cepas que se deben añadir. La producción será $P(x) = (1200 + x) \cdot (16 - 0,01x)$
$P'(x) = 16 - 0,01x - 0,01 \cdot (1200 + x) = 4 - 0,02x$
$P'(x) = 0 \Longrightarrow 4 - 0,02x = 0 \Longrightarrow x = 200$.
Deben añadirse 200 cepas.

Ejercicio A3

Consideremos los sucesos:
A: "El alumno es del colegio A"
B: "El alumno es del colegio B"
C: "El alumno es del colegio C"
S: "El alumno ha superado la prueba"

$P(A) = 0,4$ $P(S|A) = 0,8$ $P(B) = 0,35$ $P(S|B) = 0,9$ $P(C) = 0,25$ $P(S|C) = 0,82$

a) $P(S) = P(A)P(S|A) + P(B)P(S|B) + P(C)P(S|C) = 0,84$.

b) $P(B|\bar{S}) = \dfrac{P(\bar{S}|B)P(B)}{P(\bar{S})}$; $\quad P(\bar{S}|B) = 0,1; \quad P(\bar{S}) = 0,16; \quad P(B|\bar{S}) = \dfrac{0,35 \cdot 0,1}{0,16} = 0,21875.$

Figure A1. Questions (https://www.urjc.es/estudiar-en-la-urjc/admision/407-estudiantes-de-bachillerato-y-formacion-profesional-lomce#examenes-de-convocatorias-anteriores)

Rubric

CRITERIOS ESPECÍFICOS DE CORRECCIÓN Y CALIFICACIÓN

ATENCIÓN: La calificación debe hacerse en múltiplos de 0,25 puntos

OPCIÓN A

Ejercicio 1. (Puntuación máxima: 3 puntos).

Apartado (a)	Obtención de los valores críticos	0,75 puntos
	Discusión del sistema para cada caso (3x0,25)	0,75 puntos
	Total apartado (a)	1,50 puntos
Apartado (b)	Planteamiento del sistema de 2 ecuaciones	0,25 puntos
	Resolución del sistema	0,50 puntos
	Total apartado (b)	0,75 puntos
Apartado (c)	Planteamiento del sistema de ecuaciones	0,25 puntos
	Resolución del sistema	0,50 puntos
	Total apartado (c)	0,75 puntos

Ejercicio 2. (Puntuación máxima: 3 puntos).

	Expresión correcta de la función que se debe optimizar	1,50 puntos
	Obtención del nº de cepas que optimizan el problema	1,50 puntos

Ejercicio 3. (Puntuación máxima: 2 puntos).

Apartado (a)	Planteamiento correcto	0,50 puntos
	Cálculo correcto de la probabilidad pedida	0,50 puntos
	Total apartado (a)	1,00 punto
Apartado (b)	Planteamiento correcto	0,50 puntos
	Cálculo correcto de la probabilidad pedida	0,50 puntos
	Total apartado (b)	1,00 punto

Ejercicio 4. (Puntuación máxima: 2 puntos).

Apartado (a)	Cálculo correcto de $z_{\alpha/2}$	0,25 puntos
	Expresión correcta de la fórmula del intervalo de confianza	0,25 puntos
	Obtención correcta del intervalo de confianza	0,50 puntos
	Total apartado (a)	1,00 punto
Apartado (b)	Cálculo correcto de $z_{\alpha/2}$	0,25 puntos
	Planteamiento correcto	0,25 puntos
	Cálculo correcto del tamaño muestral	0,50 puntos
	Total apartado (b)	1,00 punto

NOTA: La resolución de ejercicios por cualquier otro procedimiento correcto, diferente al propuesto por los coordinadores, ha de valorarse con los criterios convenientemente adaptados.

Figure A2. Rubric (https://www.urjc.es/estudiar-en-la-urjc/admision/407-estudiantes-de-bachillerato-y-formacion-profesional-lomce#examenes-de-convocatorias-anteriores)

Solution

OPCIÓN A

$\boxed{\text{Ejercicio A1}}$

a)

$$\left(\begin{array}{ccc|c} 1 & a & -7 & 4a-1 \\ 1 & 1+a & -(a+6) & 3a+1 \\ 0 & a & -6 & 3a-2 \end{array}\right) \rightarrow \left(\begin{array}{ccc|c} 1 & a & -7 & 4a-1 \\ 0 & 1 & 1-a & 2-a \\ 0 & 0 & (a+2)(a-3) & (a+2)(a-1) \end{array}\right)$$

1. $a \neq -2, 3$. El sistema es COMPATIBLE DETERMINADO.
2. $a = 3$. El sistema es INCOMPATIBLE.
3. $a = -2$. SISTEMA COMPATIBLE INDETERMINADO.

b) Para $a = -2$ el sistema que debemos resolver es:

$$\begin{cases} x - 2y - 7z = -9 \\ y + 3z = 4 \end{cases} \rightarrow x = \lambda - 1, \quad y = 4 - 3\lambda, \quad z = \lambda \quad \text{con } \lambda \in \mathbb{R}.$$

c) Para $a = -3$ el sistema que debemos resolver es:

$$\begin{cases} x - 3y - 7z = -13 \\ x - 2y - 3z = -8 \\ -3y - 6z = -11 \end{cases} \rightarrow x = -\frac{4}{3}, \quad y = \frac{7}{3}, \quad z = \frac{2}{3}.$$

$\boxed{\text{Ejercicio A2}}$

Sea x el número de cepas que se deben añadir. La producción será $P(x) = (1200 + x) \cdot (16 - 0,01x)$
$P'(x) = 16 - 0,01x - 0,01 \cdot (1200 + x) = 4 - 0,02x$
$P'(x) = 0 \Longrightarrow 4 - 0,02x = 0 \Longrightarrow x = 200.$
Deben añadirse 200 cepas.

$\boxed{\text{Ejercicio A3}}$

Consideremos los sucesos:
A: "El alumno es del colegio A"
B: "El alumno es del colegio B"
C: "El alumno es del colegio C"
S: "El alumno ha superado la prueba"

$P(A) = 0,4$ $P(S|A) = 0,8$ $P(B) = 0,35$ $P(S|B) = 0,9$ $P(C) = 0,25$ $P(S|C) = 0,82$

a) $P(S) = P(A)P(S|A) + P(B)P(S|B) + P(C)P(S|C) = 0,84.$

b) $P(B|\bar{S}) = \dfrac{P(\bar{S}|B)P(B)}{P(\bar{S})}$; $\quad P(\bar{S}|B) = 0,1$; $\quad P(\bar{S}) = 0,16$; $\quad P(B|\bar{S}) = \dfrac{0,35 \cdot 0,1}{0,16} = 0,21875.$

$\boxed{\text{Ejercicio A4}}$

a) $\bar{x} = 29$; $z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}} = 1,645 \cdot \dfrac{2,8}{\sqrt{8}} = 1,63.$
El intervalo de confianza es $(29 - 1,63\ ;\ 29 + 1,63) = (27,37\ ;\ 30,63)$

b) $z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}} \leq E \Longrightarrow 2,17 \cdot \dfrac{2,8}{\sqrt{n}} \leq 0,9 \Longrightarrow \sqrt{n} \geq 6,75 \Longrightarrow n \geq 45,56$
El mínimo tamaño muestral es de 46 alumnos.

Figure A3. Solution (https://www.urjc.es/estudiar-en-la-urjc/admision/407-estudiantes-de-bachillerato-y-formacion-profesional-lomce#examenes-de-convocatorias-anteriores)