Chat GPT as a didactic tool to improve students' abilities in managing different semiotic representations: a first study

Giorgia Nieddu[1,2], Maria Cristina Carrisi[2], Maria Polo[2] and Antioco Luigi Zurru[2]

[1]Università degli Studi di Macerata, Italy; g.nieddu@unimc.it

[2]Università degli Studi di Cagliari, Italy; mariacri.carrisi@unica.it, mpolo@unica.it, antiocoluigi.zurru@unica.it

The large multimodal model GPT-4 with the chatbot ChatGPT can operate with text, images etc. With the further aim of understanding how to employ GPT-4 to design a classroom activity, we want to investigate if ChatGPT's multimodal functionalities allow it to effectively handle different (semiotic) representations while solving an assigned mathematical task with operations such as treatment and conversion of data, and provide potentially useful new representations for students. We conducted a qualitative test in which two tasks containing graphs were submitted to GPT-4, and we analysed its responses with a focus on used and produced representations.

Keywords: Artificial intelligence, multimodality, semiotic system, mathematics education.

Introduction

The Large Multimodal Model (LMM) of artificial intelligence GPT-4 (OpenAI, 2023) with the interface ChatGPT is capable of operating with – and generating – human natural language, programming languages, images, audio etc. interlaced together. In computer science field GPT-4 is considered a black box, as it is impossible to see all the processes that conducted to a precise answer; therefore it is a didactical black box, "a system whose inner-workings are (partly) unknown to the user" (Greubel & Siller, 2022, p.1), this representing a limit in education given the importance of argumentation and proper use of sources. Although it does not have didactical or mathematical purposes, its widespread use in mathematics classrooms by teachers and students has led to the emergence of various studies on its use in education. Yoon et al. (2024) showed how students could take authority in their reasoning through activities involving ChatGPT such writing a prompt, revising it based on the outcome, evaluating GPT's responses and deciding what to take from them, and finally write a proof. Barana et al. (2023) explained how ChatGPT supported and activated students' strategies of critical thinking and problem solving: as well as being helpful when they could not find a solution, students were able to find personal strategies such as dividing the problem into smaller tasks and analysing them with GPT or using it as a confirmation in case one or more solutions had already been found. Dilling et al. (2024) showed how ChatGPT can stimulate interaction among students in various ways such as identifying errors, discussing responses, comparing different solutions; these processes, as the authors affirm, recall and reflect the competences of mathematical communication and argumentation. Before its release, GPT-4 was tested in question answering with few mathematical tasks with diagrams to be interpreted; developers reported that its capabilities in questions with images are as effective as for text-only inputs (OpenAI, 2023); however, it is known that LMM are not reliable: the developers reported that "it "hallucinates" facts and makes reasoning errors" (OpenAI, 2023, p. 10), and it is affected by bias and disinformation, which can lead to several ethical issues. We point out that the model's

logical and mathematical performances are not solid when images are involved (Nieddu et al., 2024). It is possible that ChatGPT is subject to mathematical biases, which results in greater ability on certain topics than others (due to greater representation in the data used for training) or preference of some solving strategies over others; for example Getenet (2024) in his research involving pre-service teachers found that GPT primarily relies on algebraic equations and formulas, giving less importance to visual solutions. Variety of representations is a crucial focus in mathematics education, as the emergence of mathematical knowledge is linked to the production of different semiotic representations, for Duval (2006, p.103) "something that stands for something else". We address then the following research question: can ChatGPT handle different (semiotic) representations while solving an assigned mathematical task containing graphs and provide potentially useful new representations for students?

Our long-term goal is to design classroom activities in which students work collaboratively with different registers of semiotic representation and interact with multimodal ChatGPT; this paper constitutes a first study to inform a future activity design. After a description of the framework(s), we will analyse two tasks performed by ChatGPT with the aid of students' pre-existing data.

Theoretical Frameworks

As we focus on the design of collaborative activities with the use of ChatGPT and mobilization of representations, we employed the theory of registers of semiotic representation (Duval, 2006) and humans-with-media (Borba & Villareal, 2005); the former will help us to recognize the registers used and the relationships between them, the latter will be needed to understand the possibilities that ChatGPT offers in terms of sharing knowledge in a community with technologies.

Theory of registers of semiotic representation

According to Duval (2006), each representation can be generated automatically by a device or an organic system, as a memory image, a reflection or a photograph (non-semiotic) or produced with an organized system of signs as algebraic language, natural language etc. (semiotic). Systems which allow us to describe a representation with operations as treatment (modifying a given representation using the rules of the system in order) and conversion (translating the representation to one into another system) are called registers of semiotic representation. Although ChatGPT's representations are "automatically" generated, some of them could actually be expressed with precise registers.

Humans-with-media

The assumption that knowledge is produced by humans together with the media, whether it is writing, orality or a multimodal technology, is the basis of humans-with-media construct (Borba & Villarreal, 2005): computer technology rearranges human activity in a unique way as in activity theory (Tikhomirov, 1981), allows knowledge to be shared and a collective intelligence to be created (Lévy, 1993). This approach also focuses on the importance of representations for mathematics learning: Borba (2012, p. 803) states that "knowing can be understood as coordination of multiple representations" and

> "Based on a combination of images, other kinds of visual feedback, writing and sound, and possibilities for sharing ideas in multimedia environments, multimodal discourse becomes a new

means of expression when digital media are readily available in the production of knowledge" (Borba, 2012, p. 803).

Methodology and results

We selected two tasks from the Italian national INVALSI tests for 10th grade students containing text, tables, pie charts and bar charts. These representations were selected because discrete data is more manageable and effective compared to continuous data; tasks with other types of images, such as graphs or geometric figures, were the focus of a prior study (Nieddu et al., 2024). Mathematical contents are mean and frequency (absolute, cumulative, percentage). We submitted each task to ChatGPT as an image (see Figure 1) with the prompt (in Italian): "You are a 10th grade student. Solve the following task", as it performs better if provided with a context (Lee et al., 2023). We then analyzed GPT's answers and students' national results through the lens of Duval's representations, as in Ferretti et al. (2024), where student responses are interpreted by examining how they handle different registers.

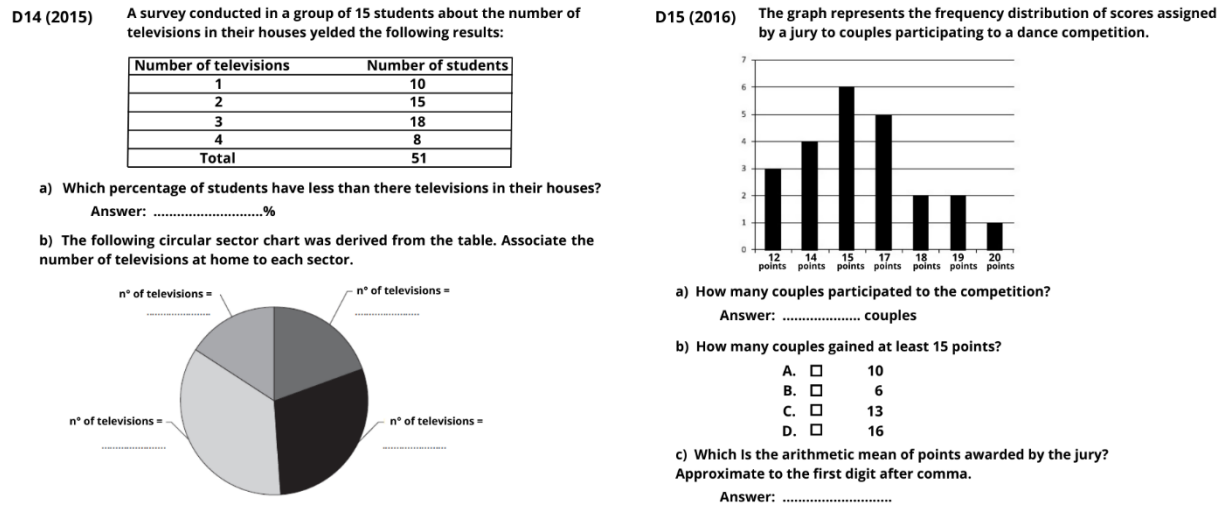


Figure 1: INVALSI tasks D14-2015, D15-2016 (https://gestinv.it/Index.aspx; authors' translation)

We employed the available pre-existing data (https://gestinv.it/Index.aspx) to understand whether the answers and produced representations could respond to students' potential difficulties. A description of GPT's answers and some transcripts follow.

Task D14-2015

ChatGPT correctly addresses part (a): "In order to calculate the percentage of students with less than three televisions at home, we sum the number of students with one or two televisions"; it calculates 10 +15 = 25 and then the percentage diving by 51 and multiplying for 100. The script for part b) follows:

ChatGPT: [..] we will need to determine the proportion of each category based on the data in the table. Remembering that the sum of the sectors in the graph must equal 100% and that each sector represents a proportion of the students with a certain number of televisions, we proceed to calculate the percentages for each category: [..] we need to associate each sector with the calculated percentages: The largest sector (about one-third of the graph) corresponds to 35.29%, representing students with 3 televisions. The next largest sector corresponds to 29.41%,

representing students with 2 televisions. The smallest sector corresponds to 15.69%, representing students with 4 televisions. Finally, the remaining sector represents 19.61%, which corresponds to students with 1 television.

It was sufficient to pick the higher frequency and so forth, without calculating percentages; we also wondered which one is for ChatGPT the largest area it refers to, and asked for more explanations:

Researcher: Are you sure that for part b you need to calculate the percentages corresponding to the various frequencies?

ChatGPT: You are right; [..] since the circular sector graph you provided should already represent the percentage distributions based on the data in the table. Therefore, my goal was more to confirm that the proportions displayed in the graph match the original data. [..] we can rely on the visual distribution to determine which sector corresponds to which category, based on the relative sizes of the sectors. [..] the largest sector should correspond to the category with the most students (in this case, those with 3 televisions), and so on.

Researcher: Can you complete the given pie chart? produce an image of the pie chart where the missing data are inserted.

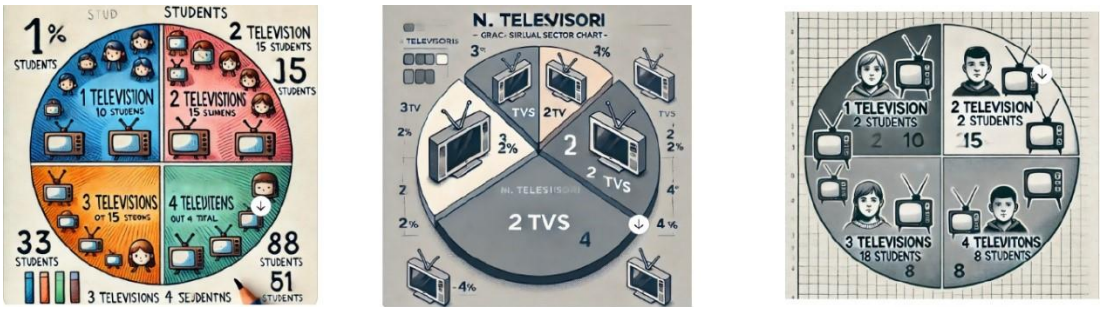


Figure 2: three attempts of ChatGPT to complete the pie chart, task D14-2015, part (b).

We suggested to associate the frequencies referring to sectors' greyscale as in the picture (we asked for the lightest sector, the second, the third and the darkest), but the answer still was meaningless.

Task D15, 2016

ChatGPT misunderstood the frequencies (it uses 1,4,6,6,5,3,1 instead of 3,4,6,5,2,2,1) in the chart.

ChatGPT: In order to determine the total number of participant couples, we will sum the frequencies of points in all columns in the graph: 12 points: 1 couple; 14 points: 4 couples; 15 points: 6 couples; 17 points: 6 couples; 18 points: 5 couples; 19 points: 3 couples; 20 points: 1 couple. Summing these values we obtain: 1+4+6+6+5+3+1=26. So, 26 couples participated to the competition.

For part b), it says: "We consider the couples which obtained 15 points or more", then proceeds to list the data (wrong as before) corresponding to 15, 17, 18, 19, 20 points and sums them. It also states that the result does not correspond to any option. In part (c), it explains "we have to multiply each score for its frequency and divide the total sum of points for the total number of couples". The calculation is addressed correctly (but with wrong data as before), and it also rounds the result as requested. We asked "are you sure you used the correct frequencies?" again obtaining wrong data (but different from the ones before).

Discussion

Using text and image input, one can effectively interact with ChatGPT and obtain new approaches when a mistake occurs. In task 14 GPT shows its understanding of the problem and its ability to

find solving strategies; it partially solved part (b) in another way when requested. It does not explicitly use the word 'frequency', this being an adaptation to the language used in the given task or a limitation in the natural language register, which remains colloquial and not specifically mathematical; it could extract necessary data from the table, showing some handling of the tabular register, a good conversion to natural language and symbolic registers, used explicitly to perform calculations. On the other hand, 51.7% of the students could not calculate the frequencies for less than 3 televisions, showing some difficulties with the cumulative frequencies (see Table 1). Part (b) was the one solved correctly by the largest number of students, we can assume that it was quite natural for them to proceed assigning the sectors in increasing or decreasing order; ChatGPT initially shows some understanding of the meaning of a pie chart, stating that the sum of the sectors must be equal to 100%; then, it maintains the link between a pie chart and percentages, associating them with the sectors. When we asked about this, it specifies that it is not necessary to calculate percentages. Finally, when asked to complete the graph, it changed the given one, 'forgetting' that the graph must visually represent the percentages (modalities) that it wrote down. It seems that priority is given to the numbers, that is the symbolic register, over the pie chart register. The answer was marked as unclear (Table 1) because it could describe the pie chart but it failed to show it, and representations are the primary focus of our analysis.

Dealing with task 15, we can affirm that GPT had a hallucination, resulting in a bad conversion from the bar chart into the natural language and algebraic register. We assume that GPT either misread the chart or randomly generated the data not knowing how to derive them. Since the study prioritizes representations, this mistake resulted in all three responses being marked as incorrect in Table 1. The natural language and symbolic registers are used in a consistent manner dealing with the concepts of frequency and mean; a weighted mean was not explicitly requested, but GPT was able to recognise the need for it. Furthermore, approximation is carried out correctly using the symbolic register. Students in general obtained better results in part (a); in part (b), it was found that most of the errors were for option B. (39.8% of total answers), corresponding to 6 couples. Being 6 the number of couples that gained exactly 15 points, we can suppose that the cumulative frequencies of 'at least 15 points' might have been mistaken for the frequency of the '15 points' modality; this shows a probable students' difficulty in decoding natural language and in transitioning from natural language register (at least 15 points) to the algebraic register (P>=15). Finally, only 24% of students performed correctly part (c); We can assume that it can largely be attributed to the presence of cumulative frequencies.

Table 1: National INVALSI results from https://gestinv.it/Index.aspx. Rows 2 and 3 refer to students' answers. A total number of 27207 students participated in 2015, and 33992 in 2016.

Task	D14-2015 (a)	D14-2015 (b)	D15-2016 (a)	D15-2016 (b)	D15-2016 (c)
Correct	48,3%	61,5%	65,3%	44,8%	24%
Wrong	44,8%	35,1%	29%	51,8%	52,6%
ChatGPT	Correct	Unclear	Wrong	Wrong	Wrong

Conclusion

Despite GPT-4 not being explicitly designed for mathematics, its performance in handling operations such as treatment and conversion between different registers has proven promising in certain contexts, such as processing natural language, tabular and symbolic data. However, limitations remain, as evidenced by its struggles in accurately converting pie charts and bar charts into meaningful representations, leading to mistakes, hallucinations and misinterpretation of data. Difficulties with pie chart register were both in replication and in analysis, because specific references to sectors in the image are never made, not even when suggested to indicate them using grayscale. It showed no difficulties in recognizing the questions' layout, not even options in question 15 (b). Dealing with the second part of our research question (production of useful representations) ChatGPT provided answers that may or may not be correct, but still well expressed, argued and explained in detail, including the use of bulleted lists, formulas and mathematical symbols used appropriately. This could provide students with useful examples of how a solution is argued and expressed using natural language and symbolic register. It is possible to edit a previously sent message, obtaining two different paths within the same conversation; this ensures the possibility of recalling previous elements or providing new representations to GPT. The model's ability to change its response once it receives a suggestion often turns out to be fictitious, as the second response will be a reworking of the first. Although the model cannot (and should not) replace human guidance in the classroom, it may enhance learning experiences by offering diverse representations and fostering critical thinking. When observing ChatGPT outputs, we can see how the model generates – types – its answers word by word, as each one is produced based on the previous one; this constitutes a huge difference between the knowledge produced, for example, by a book, or even a web page, as these media are produced before the use, and not during it. We argue that, as expressed by Borba and Villareal (2005) in their framework, the media in which a task takes place is relevant, and solving problems with ChatGPT (or other chatbots) is different to do it with every previously known media. It allows students to access information produced, expressed in a certain way, with certain modalities, exclusively for them. The clarity of GPT's argumentations and the need for precise questioning to obtain relevant responses (as seen in our pie chart task) lead us to hypothesize, based on findings from the studies mentioned in the introduction as well, that by interacting with ChatGPT, arguing and refining their queries, students could benefit of new signs to interpret the representations; students can be encouraged to conjecture solutions, evaluate and check AI solutions and discuss with peers, for example with the organization of a final discussion guided by the teacher, focusing on discerning and comparing personal contributions from GPT's ones. However, the lack of control on generated contents needs to be addressed, as well as ethical issues, mistakes, hallucinations and other limitations; it is necessary to provide some basic information about how chatbots work, such as the need of asking questions (prompt engineering) in a purposeful way and specifying the context; indeed, as found by Yoon et al. (2024), teachers can influence students' conceptions of genAI and stimulate constructive use by providing examples of errors and inaccuracies of artificial intelligence and emphasizing its use as a support and not as a solver. As expressed by Arzarello and Robutti (2010), to use effectively the new technologies available in education it is necessary to introduce new methodologies and not simply repeat traditional activities

with the new devices. For this reason, we believe it is necessary to further investigate the issue, starting from established best practices that can guide its integration. Multimodal ChatGPT could be integrated in a Concrete-Pictorial-Abstract approach to reinforce the transition between levels of abstraction; in particular, it can prove a support in the pictorial phase for images and virtual drawings, but also in the abstract phase, since, as we have seen, it can consistently handle symbolic register. The practice of scaffolding can also be applied with ChatGPT, progressively reducing its support, encouraging autonomous responses and using it only at a final stage. Future research should investigate its use in real classroom settings, applying more refined strategies for multimodal learning.

References

Arzarello, F. & Robutti, O. (2010). Multimodality in multi-representational environments. *ZDM, 42*(7), 715–731. https://doi.org/10.1007/s11858-010-0288-z

Barana, A., Marchisio, M. & Roman, F. (2023). Fostering problem solving and critical thinking in mathematics through generative artificial intelligence. In D.G. Sampso, D. Ifenthaler, & P. Isaías (Eds.), *Proceedings of the International Conference on Cognition and Exploratory Learning in the Digital Age* (pp. 377–385). Curran Associates. https://doi.org/10.33965/CELDA2023_202306L046

Borba, M. C., & Villarreal, M. E. (2005). *Humans-with-Media and the Reorganization of Mathematical Thinking: Information and Communication Technologies, Modeling, Visualization and Experimentation.* Springer US (Mathematics Education Library, Vol. 39). https://doi.org/10.1007/b105001

Borba, M. C. (2012). Humans-with-media and continuing education for mathematics teachers in online environments. *ZDM, 44*(6), 801–814. https://doi.org/10.1007/s11858-012-0436-8

Borba, M. D. C., & Balbino Junior, V. R. (2023). ChatGPT e educação matemática. *Educação Matemática Pesquisa Revista Do Programa de Estudos Pós-Graduados Em Educação Matemática, 25*(3), 142–156. https://doi.org/10.23925/1983-3156.2023v25i3p142-156

Dilling, F., Herrmann, M., Müller, J., Pielsticker, F., & Witzke, I. (2024). Initiating interaction with and about ChatGPT – an exploratory study on the angle sum in triangles. In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.G. Weigand (Eds.), *Proceedings of MEDA4* (pp. 145–152), Università di Bari Aldo Moro.

Duval, R. (2006). A Cognitive Analysis of Problems of Comprehension in a Learning of Mathematics. *Educational Studies in Mathematics, 61*(1–2), 103–131. https://doi.org/10.1007/s10649-006-0400-z

Ferretti, F., Gambini, A., & Spagnolo, C. (2024). Management of semiotic representations in mathematics: Quantifications and new characterizations. *European Journal of Science and Mathematics Education, 12*(1), 11-20. https://doi.org/10.30935/scimath/13827

Getenet, S. (2024). Pre-service teachers and ChatGPT in multistrategy problem-solving: Implications for mathematics teaching in primary schools. *International Electronic Journal of Mathematics Education, 19*(1), em0766. https://doi.org/10.29333/iejme/14141

Greubel, A., & Siller, H. S. (2022). Learning about black-boxes: A mathematical-technological model. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of CERME12* (pp. 2741–2748). Free University of Bozen-Bolzano and ERME. https://hal.science/hal-03748388v1

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies, 29,* 11483–11515. https://doi.org/10.1007/s10639-023-12249-8

Levy, P. (1993). *Les Technologies de l'intelligence. L'avenir de la pensée à l'ère informatique* [Technologies of Intelligence: The future of thinking in the informatics era]. La Découverte, Paris. https://doi.org/10.7202/1002157ar

Nieddu, G., Carrisi M.C., Polo, M. & Zurru A.L. (2024, September 2–6). Analysis of GPT-4's responses: an unconventional student. In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.G. Weigand (Eds.), *Proceedings of MEDA4* (pp. 440–441), Università di Bari Aldo Moro.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). GPT-4 Technical Report (arXiv:2303.08774). https://doi.org/10.48550/arXiv.2303.0874

Tikhomirov, O. K. (1981). The psychological consequences of the computerization. In J. Werstch (Ed.), *The concept of activity in soviet psychology* (pp. 256–278). New York: Sharpe Inc. https://doi.org/10.4324/9781003575429

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM, 56,* 1531–1551. https://doi.org/10.1007/s11858-024-01629-0