Can Generative AI Solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish

Verónica Parra[1,3], Patricia Sureda[1,3], Ana Corica[1,3], Silvia Schiaffino[2,3], Daniela Godoy[2,3] *

[1] Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ciencias Exactas, NIEM, Tandil, Buenos Aires (Argentina)
[2] Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ciencias Exactas, ISISTAN, Tandil, Buenos Aires (Argentina)
[3] CONICET, Buenos Aires (Argentina)

Received 18 November 2023 | Accepted 15 February 2024 | Published 28 February 2024

Abstract

Generative Artificial Intelligence (AI) has emerged as a disruptive technology that is challenging traditional teaching and learning practices. Question-answering in natural language fosters the use of chatbots, such as ChatGPT, Bard and others, that generate text based on pre-trained Large Language Models (LLMs). The performance of these models in certain areas, like Math problem solving is receiving a crescent attention as it directly impacts on its potential use in educational settings. Most of these evaluations, however, concentrate on the construction and use of benchmarks comprising diverse Math problems in English. In this work, we discuss the capabilities of most used LLMs within the subfield of Geometry, in view of the relevance of this subject in high-school curricula and the difficulties exhibited by even most advanced multimodal LLMs to deal with geometric notions. This work focuses on Spanish, which is additionally a less resourced language. The answers of three major chatbots, based on different LLMs, were analyzed not only to determine their capacity to provide correct solutions, but also to categorize the errors found in the reasoning processes described. Understanding LLMs strengths and weaknesses in a field like Geometry can be a first step towards the design of more informed methodological proposals to include these technologies in classrooms as well as the development of more powerful automatic assistance tools based on generative AI.

Keywords

Chatbots, Generative AI, Geometry, LLMs, Math Problem-Solving.

DOI: 10.9781/ijimai.2024.02.009

I. Introduction

The emergence and fast adoption of natural-language chatbots, such as OpenAI ChatGPT[1], or Google Bard[2], leveraging Large Language Models (LLMs) to question-answering, is a phenomenon having a growing impact in several daily activities. Education is among the most heavily impacted areas by the irruption of these tools as the interaction between generative AI with both students and teachers allows to envision promising applications in pedagogical scenarios, but also unveils potential risks.

[1] https://chat.openai.com/

[2] https://bard.google.com/

* Corresponding author.

E-mail addresses: vparra@niem.exa.unicen.edu.ar (V. Parra), psureda@niem.exa.unicen.edu.ar (P. Sureda), acorica@niem.exa.unicen.edu.ar (A. Corica), silvia.schiaffino@isistan.unicen.edu.ar (S. Schiaffino), daniela.godoy@isistan.unicen.edu.ar (D. Godoy).

Mathematics is a valuable testbed for evaluating problem-solving capabilities of LLMs as it involves the ability to analyze and comprehend the problem stated, select viable heuristics from a potentially large set of strategies, and combine them into a chain-of-thought leading to a solution. Each of these high-level abilities poses complex challenges for AI-based technologies, in general, and generative AI models, in particular.

The incorporation of generative AI in educational settings requires a deep understanding of both the capabilities and limitations of LLMs to provide solutions to Math problems as well as step-by-step explanations at different levels. Novel AI-based techniques can be built upon this knowledge and exploit LLMs potential for the development of more powerful tools, including Math teaching assistants interacting with students during their learning process and potentially offering individualized instruction.

Studies oriented to evaluate the performance of LLMs on mathematical reasoning have been mostly concerned with the construction of appropriate benchmarks and the quantitative analysis of a given model results with respect to them [1]–[5]. Although their findings can provide an overall view of LLMs performance in the Math domain, there is still a lack of understanding of their strengths and weaknesses in general terms and in specific Math areas, such as Geometry.

Finding solutions for Geometry problems might result in a specially challenging task for generative AI based on multimodal LLMs as it not only involves the knowledge of fundamental concepts (theorems) and its correct application, but specially the use of spatial reasoning skills. At the same time, Geometry has a preeminent place in high-school curricula in many countries. Because of this, it becomes essential to better understand the potential and pitfalls of chatbots in solving Geometry problems as an essential step towards the construction of more powerful teaching assistance tools as well as pedagogical strategies integrating available general-purpose chatbots.

In addition, current studies are concentrated on English texts, while the performance of LLMs in less represented languages, such as Spanish, remains to be investigated. The quality of answers of models for different languages is directly related to the amount of training data available for each language, performing better for languages with larger representation like English and exhibiting an inferior performance for languages like Spanish.

This work presents an study tending to shed some light on the abilities of chatbots to provide accurate solutions to Geometry problems in Spanish. We carried out an analysis of the answers provided by three available chatbots, namely OpenAI ChatGPT, Microsoft Bing Chat (BingChat)[3], and Google Bard, using a case study of Geometry high-school problem. The three major chatbots covered, leveraging versions of GPT-3.5 [6], GPT-4 [7] and PalM-2 [4] models, were chosen because they are accessible and currently being used by students in everyday activities and schools. The problem analyzed corresponds to an Iberoamerican Math competition[4] oriented to high school students, and it is targeted to students under 13 years old. As a result of this study, we propose a categorization of errors made by chatbots in Geometry reasoning that can be used as input towards the construction of methodological proposals fostering the use of generative AI for learning and skill acquisition.

The structure of this document is as follows: section II discusses related works in the area, section III introduces the material and methods used in this study, section IV discusses the the results obtained and, finally, section V presents the conclusions and devises promising avenues for further research.

II. Background & Related Works

In this section we first summarize some aspects regarding the use of LLMs in education (subsection A), then we discuss research on the performance of these models in Math problem-solving (subsection B) and finally we introduce some context and background concepts related to Geometry teaching (subsection C).

A. LLMs in Education

Since the launching of ChatGPT by OpenAI in 2022, there has been an intensive discussion about the integration of generative AI in several fields, particularly in education [8],[9], as well as about the ethical aspects of using artificial intelligence (AI) systems in educational contexts [10], [11]. ChatGPT was trained on a large volume of text data, using the Generative Pre-trained Transformer (GPT) deep learning architecture. Immediately, the friendly, human-like responses in natural language conversations lead ChatGPT to be one of the technologies of fastest adoption.

The irruption of generative AI and the widespread adoption of ChatGPT opened the discussion on both challenges and concerns regarding its use in educational settings. On one side, there is a pressing need of harnessing the power of these tools for enhancing teaching and learning practices. Among other benefits, LLMs can be used in the development of personalized learning tutors for students and being of assistance to teachers in the creation of educational resources (e.g. syllabus and class planning, course material and exercises) as well as the assessment of students capabilities (e.g. generating tests and evaluation scenarios), among many other applications. On the other side, LLMs potential uses rise concerns in relation to their accuracy and reliability as well as other threats such as misuses, plagiarism, the presence of biases and hallucinations and other ethical considerations. In [12] it had even found that risks also encompass the potential to limit critical thinking and creativity and impede a deep understanding of subject matter, and foster passivity.

General purpose chatbots, such as ChatGPT or Bard, are trained for dealing with question-answering in diverse domains as they are trained with large portions of the Web. However, recent studies have shown that chatbots perform differently in different subject areas including finance, coding, maths, and general public queries [13]. In [14], for example, it was found that ChatGPT performance varied across subject domains, ranging from outstanding (e.g., economics) and satisfactory (e.g., programming) to unsatisfactory (e.g., mathematics). Fine-tuning LLMs in specific domains to build educational applications upon these trained models can circumvent this issue, examples include ChemBERTa [15] or MathChat [16]. However, training for downstream tasks requires specialized data corpora and the final product is tied to the language of such data. Understanding the capabilities of most accessed, general-purpose chatbots is relevant to both introduce them as a pedagogical tool in classrooms, but also counteract inaccuracies students and teachers are exposed to while interacting with generative AI.

B. LLMs in Math and Geometry Problem-Solving

Although the entire scholar curricula is affected, the presence of AI impacts differently according to the competences and skills to be acquired by students, depending on whether they involve, for example, language abilities, communication, problem-solving capabilities, researching factual information or critical thinking.

Given its current level of adoption by students, it becomes increasingly important to evaluate LLMs performance on specific tasks, such as in this case Geometry problem-solving. It is worth noticing that, as pointed out by [17], autoregressive language models are trained for predicting the next word given a previous sequence of words. The mismatch between the problem the model was developed to solve and the task that is being given, can have significant consequences. In fact, the authors highlight the importance of viewing LLMs not as a “Math problem solver” but rather as a “statistical next-word prediction system" being used to solve Math problems. Then, failures can be understood directly in terms of a conflict between next-word prediction and the task at hand.

Different LLMs have been tested on multiple mathematical reasoning datasets showing how these models struggle to solve problems even at the level of a graduate student. In [1] a new natural-language dataset, named GHOSTS[5], was introduced. This dataset that covers graduate-level Mathematics and was curated by researchers working in Mathematics includes a subset, named Olympiad-Problem-Solving, consisting of a selection of exercises often used to prepare for Mathematics competitions. The study over this dataset concluded that ChatGPT cannot get through a university Math class, but for undergraduate Mathematics, GPT-4 can offer sufficient (but not perfect) performance. In a quantitative comparison of GPT versions in different subsets of GHOSTS it was shown that Olympiad problem solving was the subset proving to be the more difficult for these models, obtaining lower scores in such problems than even for symbolic integration.

[3] https://www.bing.com/chat

[4] https://www.oma.org.ar/internacional/may.htm

[5] https://github.com/xyfrieder/science-GHOSTS

GPT-2 and GPT-3 were tested in the Mathematics Aptitude Test of Heuristics (MATH) dataset [2] consisting of problems from high school Math competitions classified in different subjects and levels. GPT-2 accuracy reached an average of 6.9%, being better at problems of Pre-Calculus and Geometry and worse for problems related to Number Theory. GPT-3, in turn, reaches an average accuracy of 5.2%, being better at pre-Algebra and worse at Geometry. In [3], an study on the performance of ChatGPT on Math word problems (MWPs) from the dataset DRAW-1K[6] found that it changes dramatically if it is asked to provide explanations of the answer instead of simply being asked for the answer without further text. PaLM [4] version of 540-billion parameters reported to solve 58% of the problems in GSM8K[7], a benchmark of thousands of challenging grade school level Math questions, with 8-shot chain-of-thought prompting in combination with an external calculator. In turn, this result outperforms the prior top score of 55% achieved by fine-tuning the GPT-3 175B model with a training set of 7500 problems and combining it with an external calculator and verifier [5].

A few studies can be found comparing multiple available chatbots answers for Math problems. In [18] an evaluation of the Mathematics performance of Google Bard in solving Mathematics problems commonly found in the Vietnamese curricula was presented. The work findings indicate that in this regard Google Bard's performance falls behind its counterparts (Bing Chat and ChatGPT). For these experiments, a Vietnamese dataset was translated into English since Bard lacks support for Vietnamese at the moment the study was carried out. A comparison between three chatbots like ChatGPT-3.5, ChatGPT-4 and Google Bard was presented in [19], focusing on their ability to give correct answers to Mathematics and Logic problems. For a set of 30 questions, it was found that for straightforward arithmetic, algebraic expressions, or basic logic puzzles, chatbots may provide accurate solutions, although not in every attempt. For more complex Mathematics problems or advanced logic tasks, their answers were unreliable.

Mechanisms to improve the ability of LLMs to complex reasoning are based on generating a chain of thought, i.e. a series of intermediate reasoning steps. Chain-of-thought prompting (CoT) [20] leverages intermediate natural language rationales as prompts to enable LLMs to first generate reasoning chains and then predict an answer for an input question. On the GSM8K benchmark of Math word problems, for example, chain-of-thought prompting with PaLM 540B outperforms standard prompting by a large margin and achieves new state-of-the-art performance, surpassing even finetuned GPT-3 with a verifier. In the same direction, an evaluation on difficult high school competition problems from the MATH dataset was presented in [16] and MathChat, a conversational problem-solving framework was proposed. It simulates a mock conversation between an LLM assistant using GPT-4 and a user proxy agent working together to solve the Math problem. On the problem with the highest level of difficulty from MATH, MathCat improves the accuracy from 28% of GPT-4 to 44% and has competitive performance across all the categories of problems.

Multimodal LLMs (MLLMs) seem to be the most appropriate option to complement reasoning capabilities with the spatial thinking needed to Geometry problem-solving. However, even the most advanced MLLMs still exhibit limitations in addressing geometric problems due to challenges in accurately comprehending geometric figures [21]. Specifically, the model struggles with understanding the relationships between fundamental elements like points and lines, and in accurately interpreting elements such as the degree of an angle. It has been argued [21] that the inaccurate descriptions for geometric shapes produced by models such as GPT4-V (GPT4 with vision) reside on the fact that the model struggles with understanding the relationships between fundamental elements like points and lines, and in accurately interpreting elements such as the degree of an angle. Current solutions like G-LLaVA [21], built upon LLaVA (Large Language and Vision Assistant) model [22], involve enriching the training data and creating augmented datasets (Geo170K) for improving model training. As mentioned before, the resulting models are less accessible than general-purpose ones and available a mainstream language as English.

With large language models rapidly evolving, there is a pressing need to understand their capabilities and limitations in the context of mathematical reasoning and, particularly, in specific fields like Geometry. Current studies have been centered on measuring the performance of LLMs on benchmarks of broad sets of Mathematical problems in English. To the best of our knowledge, this is the first work focusing on understanding question-answering capabilities of the widely available chatbots regarding Geometry in Spanish language.

C. Geometry in the Classroom

Geometry is one of the basic subjects of Mathematics. For analyzing Geometry in the context of Argentine educational system, in which the present study takes place, three edges need to be considered: curricular design, actual work in classrooms and the Argentine Mathematics Olympiads (OMA[8]). In the first case, one of the four priority learning blocks proposed by the Argentine Ministry of Education is Geometry [23]. Thus, the vast majority of the curricular designs of each jurisdiction prescribe studying Geometry throughout the secondary education (both in the basic and higher levels). The curricular relevance of Geometry derives from its close relationship with various fields, including Natural and Social Sciences, as well as everyday life [24]–[26]. However, even though Geometry continues to be present in secondary school curricular designs, various researchers highlight the absence of Geometry in the classroom [24],[27]. The third edge corresponds to a competition that has been taking place in Argentina for more than 30 years: the Argentine Mathematics Olympiads [28]. The fundamental objective of these Olympiads is to stimulate mathematical activity among young people and develop the ability to solve problems (OMA, regulations, art 2.). The OMA proposes the resolution of problems, which can be grouped into two large types: arithmetic-algebraic and geometric.

In summary, the official curricular guidelines propose studying Geometry in secondary school, however, this guideline is not materialized in the classrooms (or it is, but weakly). Moreover, Geometry is one of the two types of problems that are used to assess mathematical skills of the students who participate in the OMA. We highlight the importance given to OMA because it is not only promoted by educational centers, but also by provincial governments (as it can be seen in their official site), motivating students to participate actively. In this work we explore how various resources from generative AI can be used to study geometric problems.

III. Materials and Methods

The goal of the analysis carried out in this work is to explore the performance of chatbots when dealing with a problem involving Geometry notions at the level of second and third year of high-school curricular design. The assessment of chatbots capacity of providing accurate answers and, or in the case of failure, the common mistakes and deficiencies found in the described solutions, can serve as basis for the creation of more efficient teaching methodologies involving generative AI.

For the purpose of this study, an Olympiad problem was selected, as described in section A, and the answers of three chatbots, enumerated in section B, to its formulation were collected. The methodology used for analyzing these answers is described in section C.

[6] https://paperswithcode.com/dataset/draw-1k

[7] https://paperswithcode.com/dataset/gsm8k

[8] https://www.oma.org.ar/

TABLE I. Summary of Errors Found in the Answers of Chatbots

Error type	ChatGPT 3.5				Bing Chat				Bard			
	#1	#2	#3	Total	Precise	Balanced	Creative	Total	#1	#2	#3	Total
Construction	2	0	2	4	-	0	3	3	3	0	1	4
Conceptual	2	3	0	5	-	3	0	3	0	2	0	2
Contradiction	0	0	1	1	-	0	1	1	0	0	0	0
Total	4	3	3	10	-*	3	4	7	3	2	1	6

* This is a case in which the chatbot did not provide a solution to the problem.

A. Geometry Problem

The problem used in this work belongs to the May Olympiads, an Iberoamerican Mathematics contest. This competition has 2 levels, the first level is for students who, in the year previous to the contest, are under 13 years old at December 31st, and the second level is for students under 15 years old at December 31st. In each level the test is unique, and it consists of 5 problems that students must solve within 3 hours. From these problems, a Geometry problem of level 1 proposed at May Olympiads in 2018[9] [29] was considered.

The problem selected is characterized by not having an immediate and unique solution. In fact, reaching a solution requires knowledge about regular polygons and their properties, circumference and its properties, similarity between polygons, the Pythagorean theorem, trigonometric ratios, among other concepts. Therefore, it is necessary to know and understand a variety of geometrical notions to decide which is the most appropriate to reach a solution.

The geometric problem was selected in such a way that both the mathematical concepts involved and the procedures for its resolution correspond to what is indicated in the official curricular design for Argentine secondary schools [23]. In these designs, the Ministry of Education proposes the minimum knowledge that must be taught in each discipline for each year of the Argentine secondary level. In particular, in Mathematics and in the Geometry area, for students aged 12–13 years old, the study of figures is proposed, arguing about the analysis of properties. In correspondence with the selected problem, students are encouraged to: determine points that meet conditions related to distances and construct circumferences, circles, bisectors and perpendicular bisectors as geometric spaces; explore different constructions of triangles and argue about necessary and sufficient conditions for their congruence; construct similar figures from different information and identify necessary and sufficient conditions of similarity between triangles; analyze claims about properties of figures and argue about their validity, recognizing the limits of empirical evidence; formulate conjectures about properties of figures (in relation to interior angles, bisectors, diagonals, among others) and produce arguments that allow them to be validated. Therefore, the problem analyzed in this work, although it may not be a typical high-school task, involves the concepts that should be addressed at school according to what is prescribed by the Argentinian curricular design.

The problem statement is as follows:

Problem Statement

Sea ABCDEFGHIJ un polígono regular de 10 lados que tiene todos sus vértices en una circunferencia de centro O y radio 5. Las diagonales AD y BE se cortan en P y las diagonales AH y BI se cortan en Q. Calcular la medida del segmento PQ.

English translation: Let ABCDEFGHIJ be a regular 10-sided polygon that has all its vertices in a circumference with center O and radius 5. The diagonals AD and BE intersect at P and the diagonals AH and BI intersect at Q. Calculate the length of segment PQ.

[9] https://www.oma.org.ar/enunciados/enunciados_Mayo2018.pdf

The solution proposed by the OMA [29] is based on the graphic representation of the decagon and the identification of the segment that needs to be calculated (PQ). The suggested strategy for reaching the solution consists in drawing segments that join the vertices of the decagon with its center and diagonals. The analysis of the triangles and trapezoids that result from the constructions allows to infer that the triangles are isosceles. From this analysis it is concluded that the requested segment has the same length as the radius of the circumference in which the decagon is inscribed. This resolution enables to find the exact value of the length of the segment PQ, which is 5 cm.

B. Chatbots and LLMs

The three major, freely accessible chatbots available at the time of this article were used for collecting answers for the previous problem. Each of these chatbots rely on its own large language model, an AI model designed to understand and generate human-like text based on deep learning techniques, learned on different corpus using also different learning strategies. LLMs have a large number of parameters and are trained over a massive amount of text data from different sources to capture complex language patterns and relationships. Specifically, the chatbots used for this study were:

ChatGPT: ChatGPT (September 25 version) trained over GPT-3.5 language model is the original chatbot launched by OpenAI in November, 2022.

Bing Chat: the chatbot accessible through Microsoft Bing search engine and running on GPT-4. This chat offers answers in three modes: (1) More Creative: responses are original and imaginative, creating surprise and entertainment; (2) More Precise: responses are factual and concise, prioritizing accuracy and relevancy; and (3) More Balanced: responses are reasonable and coherent, balancing accuracy and creativity in conversation.

Bard: the chatbot developed by Google AI and powered by PaLM-2 large language model.

For this analysis, zero-shot learning was employed. This is, LLMs were asked to answer the question directly, without any prior data or example questions. The prompt was the problem statement in Spanish exactly as in the original text of the Olympiad competition. For each model, 3 answers were obtained by regenerating the responses in order to account for the randomness in text generation.

C. Methodology

Beyond the correctness of the solution itself, the answers provided by chatbots were scanned for identifying reasoning mistakes and inaccuracies in the generated chain-of-thought, individual steps and operations. Basically, it was checked if the appropriate notions were recalled and correctly applied and if the chatbot was able to generate a coherent answer with an accurate solution.

In the process of analyzing the answers of chatbots to the stated Geometry problem, several mistakes of different types were identified. After grouping these mistakes according to their nature, we propose a general categorization of errors. Mistakes made in solving the problem were classified into three main types or categories:

- **Construction**: in this category we find errors originated on the representation made on the plane of the geometric elements indicated in the text answer given by a chatbot. In other words, a construction error is a mismatch between the textual response and the actual geometric figures and their graphical representation. For example, the chatbot ensures that a central angle has 72° when the actual amplitude according to the description given of the figure's elements is necessarily a different one.

  Construction errors denote a lack of comprehension of the LLMs of the spatial relationships among elements like points, lines and angles. As the description of the geometric problem reasoning advances, it starts to lose correlation with the actual graph that materializes such description. More likely, there errors stem from the inability of generative AI to understand the semantics behind these geometric notions at the level required for geometric reasoning.
- **Conceptual**: errors in this category relate to incorrect definitions, the application of properties without guaranteeing the necessary conditions or mixing measurement units (e.g. units of length with those of amplitude). An example of conceptual error can be applying the Pythagorean theorem to a not right-angled triangle. The possible causes of these mistakes can be varied. Language generation tools based on AI are capable of producing text using geometric vocabulary, which allows them, for example, to give a reasonable explanation of the Pythagorean theorem. However, as a consequence of an inadequate knowledge and representation of geometric shapes, they are also likely to offer solutions that apply the theorem incorrectly or make inaccurate calculations. LLMs can also suffer from a deficient context description, which in a next-word mechanism is the previous sequence of words. Then, the omission of relevant information reduces the precision in text prediction. The deficient description of the context includes simply missing some piece of information (e.g. the amplitude of a given angle), but also well-known properties (e.g. that the angles of a triangle must sum to 180 degrees) and common assumptions. Furthermore, LLMs are data-driven models trained on data that might include generalized mistakes and misconceptions. Due to their probabilistic nature, LLMs are then prone to reproduce them.
- **Contradiction**: in a number of reasoning steps, contradictions arise as an inconsistency between a deduction and either information involved in the following reasoning steps or the representation on the plane. In other words, the chain-of-thought contains contradictory knowledge, which invalidates the whole reasoning. For example, a contradiction can be inferring that an angle is acute while the graphical representation built starting from this deduction depicts a straight angle.

The mentioned categories groups a number of mistakes found in the solutions provided by chatbots. In a single answer, one or more of these mistakes were identified, leading to a conjunction of errors that ended up in a wrong answer to the problem. This general classification of mistakes found in the collected answers enables to reach a better comprehension about the failures on geometric reasoning of LLM generated texts.

IV. Results & Discussion

In order to compare the performance of chatbots according to the provided responses, which due to space limitations are not detailed here, Table 1 summarizes the total number of errors found within each category. For ChatGPT 3.5 three responses were generated, Bard also offers three versions of the answer through its interface, and Bing Chat provides three answers in the form of the more precise, the more balanced and the more creative one.

From the 9 answers (3 for each model) extracted from ChatGPT 3.5, Bing Chat and Bard, only one of them indicated the correct value of the PQ segment length, i.e. only one provided the correct solution to the stated problem, this corresponds to the Bard response #2. However, the model arrived at the result through a method having conceptual errors, thereby it cannot be considered a satisfactory solution either. In addition, there was a case in which the chatbot did not provide a solution at all, this is the case of Bard when it is asked for the More Precise answer to the question. The answer pointed out some decagon properties, but ends up saying (translated from Spanish): "*However, this calculation can be quite complicated and would require in-depth knowledge of the Geometry of the decagon. I would recommend that you consult a Geometry textbook or online resource for a detailed explanation of how to perform these calculations*".

Overall, the general performance of LLMs in generating a text for answering the Geometry problem stated was disappointing, completely failing at providing an accurate answer to the problem at hand and making a considerable number of mistakes of different types along the reasoning process. This is a concerning finding, considering that the problem presented is a high-school level one, designed for students under 13 years old, which are likely to access chatbots looking for help and would receive not only unreliable answers, but possible introducing or reaffirming Geometry misconceptions.

Considering the type of errors made by each chatbot, ChatGPT 3.5 and Bard were the ones exhibiting more errors belonging to the *Construction* type. Additionally, ChatGPT 3.5 contains a greater number of errors of the *Conceptual* category. Less frequent in all answers are the errors in the "Contradiction" category, accounting for one error of ChatGPT 3.5 and one of Bing Chat, but none in Bard.

For illustrating the different types of errors found in the analyzed answers, Tables II, III and IV provide examples of each type of the errors existing in the actual answers from the model. The tables include a fragment of the response (2nd column) generated by a chatbot (indicated in the 1st column) based on the corresponding LLM when queried using the problem statement and a description of the mistake made (3rd column). In the last column, observations related to the error detected are commented accompanied by a graph, made by the authors of this paper, based on the indications provided in the response.

In the first of them, Table II, the errors refer to the construction of angles (ChatGPT3.5), the construction of right triangles (Bing Chat) and supplementary angles (Bard). Then, in Table III, the errors that are exemplified refer to units of length and amplitude (ChatGPT3.5), to lengths of diagonals of the decagon (Bing Chat) and to heights of triangles (Bard). Table IV contains prototypes of statements about the equality of segments of different lengths (ChatGPT3.5), and mismatch between exterior and interior angles (Bing Chat). In this case, Bard does not contain errors of the *Contradiction* type.

From a qualitative point of view, responses of the different chatbots powered by the corresponding LLM, exhibit some distinctive characteristics.

In the case of the ChatGPT 3.5, some observations can be made about the provided answers. First, the generated text do not justify the mathematical procedures. Justification and reasoning about solutions are important goals in mathematical work in general, and for students who solve math problems in particular. Research shows that this is an issue that needs attention as there are important differences in the way math problems are approached and how justification supports the proposed solutions [30], [31]. Second, the notions used are not explicitly identified and mentioned as denoted in the field. Third, the response does not offer definitions and/or clarifications that could contribute to the resolution.

TABLE II. Examples of *Construction* Errors Produced by the Models

<table>
<tr><th>Chatbot</th><th>Fragment of the response</th><th>Description</th></tr>
<tr>
<td>ChatGPT 3.5</td>
<td>
Ahora, consideremos el triángulo OPQ formado por las diagonales AD y BE que se cruzan en P y las diagonales AH y BI que se cruzan en Q. Este triángulo tiene un ángulo en el centro O que subtiende el arco AB (36 grados), y los otros dos ángulos en O son iguales a α (los ángulos en los vértices A y B).

Usaremos la propiedad de la suma de los ángulos de un triángulo, que nos dice que la suma de los ángulos en un triángulo es siempre igual a 180 grados. Entonces:

36° + α + α = 180°
2α + 36° = 180°
2α = 180° - 36°

English translation:

Now, consider the triangle OPQ formed by the diagonals AD and BE that intersect at P and the diagonals AH and BI that intersect at Q. This triangle has a central angle O that subtends the arc AB (36 degrees), and the other two angles at O are equal to α (the angles at vertices A and B).

We will use the property of the sum of the angles of a triangle, which tells us that the sum of the angles in a triangle is always equal to 180 degrees. Then:

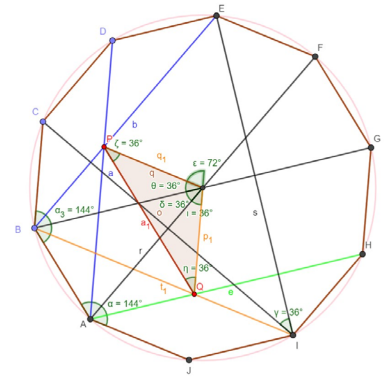
$36° + \alpha + \alpha = 180°$
$2\alpha + 36° = 180°$
$2\alpha = 180° - 36°$
</td>
<td>


Fig. 1. Graphical representation in GeoGebra of ChatGPT 3.5 response.

Considering the OPQ triangle, the angle with center O is 108°(3α with α=36°) as can be observed in Fig. 1. The model states erroneously that the angle with center O is 72°(2α with α=36°)
</td>
</tr>
<tr>
<td>Bing Chat (More creative)</td>
<td>
3. Aplicando el teorema del ángulo exterior, hallé que el ángulo BPD mide 72° + 36° = 108° y el ángulo AQH mide 72° + 72° = 144°.
4. Usando el teorema de los ángulos opuestos por el vértice, deduje que el ángulo APD también mide 108° y el ángulo BQH también mide 144°.
5. Por último, aplicando el teorema de Pitágoras en el triángulo rectángulo OPQ, encontré que la medida del segmento PQ es:
PQ^2 = OP^2 + OQ^2 - 2OPOQ*cos(PÔQ)
PQ^2 = 5^2 + 5^2 - 255*cos(108° + 144°)
PQ^2 = 50 - 50*cos(252°)
PQ^2 = 50 + 50*0.9511
PQ^2 = 97.555
PQ = √97.555

English translation:

3. Applying the exterior angle theorem, I found that angle BPD measures 72°+ 36° = 108°and angle AQH measures 72°+ 72°= 144°.

4. Using the vertex opposite angles theorem, I deduced that angle APD also measures 108°and angle BQH also measures 144°.

5. Finally, applying the Pythagorean theorem to the right triangle OPQ, I found that the measure of the segment PQ is:

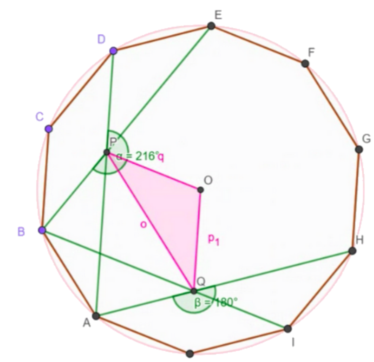
$PQ^2 = OP^2 + OQ^2 - 2 * OP * OQ * cos(P\hat{O}Q)$
$PQ^2 = 5^2 + 5^2 - 255 * cos(108∘ + 144∘)$
$PQ^2 = 50 - 50 * cos(252∘)$
$PQ^2 = 50 + 50 * 0.9511$
$PQ^2 = 97{,}555$
$PQ = \sqrt{97{,}555}$
</td>
<td>


Fig. 2. Graphical representation in GeoGebra of Bing Chat response.

The model states that the OPQ triangle is a right one, pink triangle in Fig. 2, but it is not. The POQ angle is 108°because it is the sum of 3 central angles corresponding to the 3 sides of the decagon involved and each of them is 36°. Thus, each of the other 2 angles is (180°-108°)/2= 36°.
</td>
</tr>
<tr>
<td>Bard</td>
<td>
Solución:

Dado que el polígono es regular, todos los ángulos interiores son iguales. En un polígono de 10 lados, la suma de los ángulos interiores es $(10-2)\cdot 180° = 1440°$. Como cada ángulo interior es igual, cada uno mide $1440°/10 = 144°$.

Los ángulos $\angle APD$ y $\angle BQE$ son suplementarios, por lo que $\angle APD + \angle BQE = 180°$. Sustituyendo los valores de los ángulos, obtenemos $180° = 144° + \angle BQE$. De esta ecuación, se deduce que $\angle BQE = 36°$.

Los ángulos $\angle AOB$ y $\angle BOC$ son iguales, ya que son opuestos por el vértice O. Por lo tanto, $\angle AOB = \angle BOC = \frac{180°-144°}{2} = 18°$.

English translation:

Since the polygon is regular, all interior angles are equal. In a 10-sided polygon, the sum of the interior angles is (10-2)*180°=1440°. Since each interior angle is equal, each one measures 1440°/10=144°.

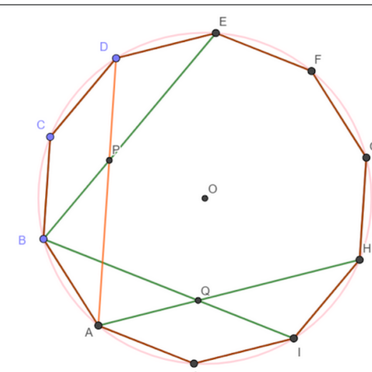
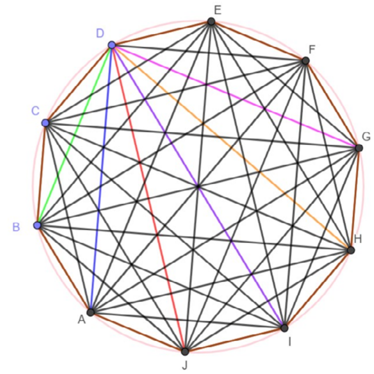
The angles ∠APD and ∠BQE are supplementary, so ∠APD + ∠BQE = 180°. Substituting the values of the angles, we obtain 180° = 144° + ∠BQE. From this equation, it follows that ∠BQE = 36°.
The angles ∠AOB and ∠BOC are equal, since they are opposite by the vertex O. Therefore, $\angle AOB = \angle BOC = \frac{180° - 144°}{2} = 18°$
</td>
<td>


Fig. 3. Graphical representation in GeoGebra of Bard response.

The model considers that the APD angle is supplementary to BQE, but APD is already 180°as can be observed in Fig. 3.
</td>
</tr>
</table>

TABLE III. Examples of *Conceptual* Errors Produced by the Models

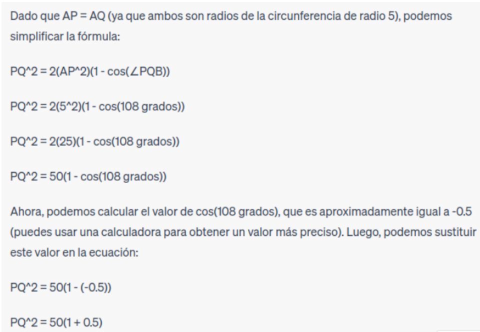
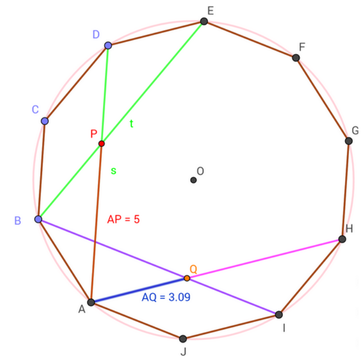
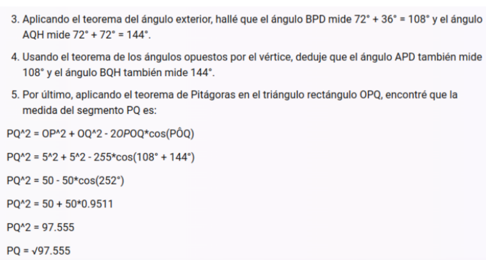
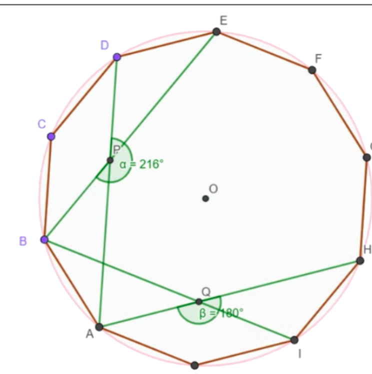
Chatbot	Fragment of the response	Description
ChatGPT 3.5	La longitud del segmento PQ es igual a la longitud del arco PQ, que es (5/4)π unidades. **English translation**: The length of the segment PQ is equal to the length of the arc PQ, which is (5/4) π units.	A radian is a unit of measurement for angles. The model gives the length of the segment PQ in radians, therefore, it confuses the units of length with those of amplitude.
Bing Chat (More creative)	Este problema se puede resolver utilizando la fórmula de la longitud de la diagonal de un polígono regular. La longitud de la diagonal de un polígono regular de n lados es igual a: $d = \frac{s2(n-2)}{2}$ **English translation**: This problem can be solved using the formula for the length of the diagonal of a regular polygon. The length of the diagonal of a regular polygon with n sides is equal to: $d = \frac{s2(n-2)}{2}$	 Fig. 4. Graphical representation in GeoGebra of Bing Chat response. The model proposes a formula to calculate the length of the diagonals of the decagon. It assumes that all diagonals that can be drawn have the same length, which, as shown in Fig. 4, is incorrect.
Bard	Por lo tanto, el segmento $\overline{PQ}$ es la altura del triángulo isósceles $\triangle POQ$. Si x es la medida del segmento $\overline{PQ}$, entonces $QO = 5 - x$. Por el teorema de Pitágoras en $\triangle POQ$, tenemos que $x^2 = (5-x)^2 + 5^2$ $x^2 = 25 - 10x + x^2 + 25$ $2x^2 - 10x = 0$ $x(2x-10) = 0$ $x = 0 \text{ o } x = \frac{10}{2} = 5$ Como x es la medida de un segmento, $x > 0$. Por lo tanto, la medida del segmento $\overline{PQ}$ es $\boxed{5}$. **English translation**: Therefore, the segment $\overline{PQ}$ is the height of the isosceles triangle $\triangle POQ$. If x is the measure of the segment $\overline{PQ}$, then $QO = 5 - x$. By the Pythagorean theorem in $\triangle POQ$, we have that $x^2 = (5-x)^2 + 5^2$ $x^2 = 25 - 10x + x^2 + ^{25}$ $2x^2 - 10x = 0$ $x(2x-10) = 0$ $x = 0 \text{ or } x = \frac{10}{2} = 5$ Since x is the measure of a segment, $x > 0$. Therefore, the measure of segment $\overline{PQ}$ is 5.	The answer considers that the triangle POQ is isosceles. It also considers PQ to be the height of the triangle, but this is not the case. Based on this, using the Pythagorean theorem, it proposes to calculate the segment PQ. It is not noticed that the triangle POQ is not a right angle one, because the angle POQ measures 108° as explained above. The correct value of the segment PQ is found in this response, but the procedure is incorrect.

In terms of this general characterization of responses, in the first response, Bing Chat explains the characteristics of the decagons, the properties of the angles and the sides, but it does not solve the problem at all. Instead, the chatbot limits itself to suggest consulting a Geometry book or online resources. On the contrary, in the second and third answers, it uses an equation editor (instead of equation written in a textual manner) and suggests resorting to external websites (which are linked) either to reference figures or schemes included in the answer (answer #2) or to refer the reader to more examples of solved Geometry problems (answer #3). Both characteristics, using proper notation to better display equations and linking to external sources, are beneficial for students as they can resort to some extra help beyond the generated text. Finally, Bard mentions the decagons, the properties of angles and sides, and tries to solve the problem, but the justification of the procedure is incorrect. For answer #2, a justification is developed and, as previously mentioned, the chatbot arrives at the correct answer, but through a procedure containing *Conceptual* errors.

V. Conclusions

In this work, we have presented an analysis and comparison of resolutions formulated by three major chatbots such as ChatGPT 3.5, Bing Chat and Bard, to a Geometry problem extracted from the first level of the May Olympiads competition (for students under 13). The

TABLE IV. Examples of *Contradiction* Errors Produced by the Models

Chatbot	Fragment of the response	Description
ChatGPT 3.5	 Dado que AP = AQ (ya que ambos son radios de la circunferencia de radio 5), podemos simplificar la fórmula: PQ^2 = 2(AP^2)(1 - cos(∠PQB)) PQ^2 = 2(5^2)(1 - cos(108 grados)) PQ^2 = 2(25)(1 - cos(108 grados)) PQ^2 = 50(1 - cos(108 grados)) Ahora, podemos calcular el valor de cos(108 grados), que es aproximadamente igual a -0.5 (puedes usar una calculadora para obtener un valor más preciso). Luego, podemos sustituir este valor en la ecuación: PQ^2 = 50(1 - (-0.5)) PQ^2 = 50(1 + 0.5) **English translation**: Since AP = AQ (since both are radii of the circle of radius 5), we can simplify the formula: $PQ^2 = 2(AP^2)\ (1 - cos(\angle PQB))$ $PQ^2 = 2(5^2)\ (1 - cos(108°))$ $PQ^2 = 2(25)\ (1 - cos(108°))$ $PQ^2 = 50\ (1 - cos(108°))$ Now, we can calculate the value of $cos(108°)$, which is approximately equal to -0.5 (you can use a calculator to get a more precise value). Then, we can substitute this value into the equation: $PQ^2 = 50\ (1 -(-0.5))$ $PQ^2 = 50\ (1 + 0.5)$	 Fig. 5. Graphical representation in GeoGebra of ChatGPT 3.5 response. It states that AP (5 u) is equal to AQ (3,09 u), both are depicted in Fig. 5.
Bing Chat (More creative)	 3. Aplicando el teorema del ángulo exterior, hallé que el ángulo BPD mide 72° + 36° = 108° y el ángulo AQH mide 72° + 72° = 144°. 4. Usando el teorema de los ángulos opuestos por el vértice, deduje que el ángulo APD también mide 108° y el ángulo BQH también mide 144°. 5. Por último, aplicando el teorema de Pitágoras en el triángulo rectángulo OPQ, encontré que la medida del segmento PQ es: PQ^2 = OP^2 + OQ^2 - 2*OP*OQ*cos(PÔQ) PQ^2 = 5^2 + 5^2 - 2*5*5*cos(108° + 144°) PQ^2 = 50 - 50*cos(252°) PQ^2 = 50 + 50*0.9511 PQ^2 = 97.555 PQ = √97.555 **English translation**: Idem Table II, row 2.	 Fig. 6. Graphical representation in GeoGebra of Bing Chat response. It proposes to calculate the angle BPD by identifying an exterior angle and proposes that the angle measures 108° and AQH=144°, this cannot be right because, as the points are aligned, the angle is straight (180°) as can be observed in Fig. 6.
Bard	No *Contradiction* errors were identified in this model answers	

three chatbots leverage different LLMs, namely GPT-3.5, GPT-4 and PaLM-2, to generate textual responses to natural language queries. In particular, the problem statement as originally presented to students in Spanish was used as a prompt for the chatbots so that three answers were collected from each in order to account for the random components of content generation.

In terms of correctness of the obtained solutions, chatbots had a disappointing performance. Only one answer, provided by Bard, reached the number that was expected ($\overline{PQ} = 5$). However, even when it arrives to the right answer, the described reasoning contains conceptual errors. On the other side, the first response given by Bing Chat does not offer a solution, it only refers the user to consult a Geometry book or some online resource.

In a more detailed analysis of the answers, we found that all of the responses given by the different chatbots contained several types of errors. In a further inspection of these different errors we were able to define a classification encompassing three main categories: construction, conceptual and contradiction. Construction errors correspond to a mismatch between the text description and its geometric representation, conceptual errors involve the incorrect use of geometric concepts and misconceptions, while the last type of error refers to contradictions appearing within the textual description or with respect to the graphical representation.

According to the proposed categorization of errors, ChatGPT 3.5 and Bard made most mistakes within the Construction category. This is an issue related specifically to Geometry as it has to do with the translation of a geometric specification given in text to a graphical representation. Additionally, ChatGPT 3.5 responses contain a greater number of errors in the Conceptual category, this is, in the application of geometric notions. The Contradiction category is the less frequent one, appearing once in ChatGPT 3.5 answers and once in the ones from Bing Chat, but never in Bard answers.

Most failures observed in the answers to the proposed problem are related to two common criticisms of LLMs [32], the lack of symbolic structure and the lack of grounding. Both questions their capacity to provide human language representation and understanding in spite of their human-like language abilities. The lack of symbolic structure prevents the model to perform formal reasoning and verify reasoning steps, whereas the lack of grounding leads to the misinterpretation of geometric notions and their visual representations. to In other words, the fact of being language models poses some limitations for solving more formal problems, such as Geometry ones.

The proposed classification contributes to a better understanding of the failures of LLMs in math-problem solving and, more specifically, those related to spatial representations involved in Geometry problems (e.g. construction errors refers to the relation between the text and its graphical interpretation). The knowledge and recognition of these issues represent also an opportunity to see errors as a valuable educational tool [33]. This categorization can serve as the basis for the construction of methodologies that include the interaction with chatbots in the classroom leveraging on errors to foster their identification, critical thinking of reasoning steps and operations, and reflection on alternative problem solutions.

Although the disappointing results provided by chatbots cannot be directly attributed to the language used, training data in Spanish is known to be smaller than in English. Consequently, next-word prediction performed by LLMs can be assumed to be less precise, thereby the generated lower-quality content. In fact, the reported evaluations of LLMs on different benchmarks including Geometry problems in English, as discussed in section II, showed a better performance than the one achieved with this particular problem. Even tough an example is clearly not sufficient to draw conclusions, the language can be considered a source of additional difficulties for LLMs.

Findings of the analysis carried out in this work are specially concerning, considering that the problem presented is a high-school level one, designed for students under 13 years old (although being an Olympiad problem may be beyond the capabilities of a typical of student of that age), which have easy access and are likely to resort to chatbots looking for help to solve similar problems. In this context, they not only will receive unreliable answers in terms of the correctness of the solution to a stated problem, but what is even more serious, they will be also exposed to inaccurate applications of mathematical notions, possibly introducing new misconceptions or reaffirming existing ones. This is also a warning sign for teachers using chatbots to generate course material or exam questions, as they can inadvertently introduce some mistakes.

According to the results obtained in solving the problem stated and taking into account the general characterization of the interface of these tools, it can be concluded that the use of chatbots (and the models behind them) for solving Geometry problems is not appropriate without a critical analysis from teachers as well as the students. The inclusion of these technologies in the classroom must follow a careful methodological approach. Potentially valuable applications of these models in the classroom could be the critically enhanced analysis, supported by teachers, of the responses obtained by chatbots, such as the one presented in this work. This would allow students to discuss and learn Geometry concepts (properties, characteristics, constructions in the plane, etc.) in a practical way. For example, it would be useful to distinguish when it is possible (or not) to apply a theorem (lemma, corollary, etc.).

In view of the current wide adoption of chatbot technologies in the classroom and by students of different ages, future work is envisioned to expand the categorization of errors in Geometry problems through the analysis of more problems in different levels. The analysis of a wider variety of problems would likely allow a finer-grained categorization of errors and the emergence of more types, less frequent types of mistakes. Ultimately, systematic evaluations of LLMs performance as the one carried out in this work contributes to the ongoing development of more advanced, capable AI chatbot systems that can be fully integrated in teaching practices to enhance learning processes.

References

[1] S. Frieder, L. Pinchetti, A. Chevalier, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, J. Berner, "Mathematical capabilities of ChatGPT," 2023.

[2] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, "Measuring mathematical problem solving with the MATH dataset," in *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[3] P. Shakarian, A. Koyyalamudi, N. Ngu, L. Mareedu, "An independent evaluation of ChatGPT on mathematical word problems (MWP)," in *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023)*, 2023.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, "PaLM: Scaling language modeling with pathways," 2022.

[5] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, Vancouver, BC, Canada, 2020.

[7] OpenAI, "GPT-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023.

[8] F. J. García Pen¨ alvo, F. Llorens-Largo, J. Vidal, "La nueva realidad de la educación ante los avances de la inteligencia artificial generativa," *RIED-Revista Iberoamericana de Educación a Distancia*, vol. 27, p. 9–39, ene. 2024, doi: 10.5944/ried.27.1.37716.

[9] B. Memarian, T. Doleck, "ChatGPT in education: Methods, potentials, and limitations," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, p. 100022, 2023, doi: 10.1016/j.chbah.2023.100022.

[10] B. Han, S. Nawaz, G. Buchanan, D. McKay, "Ethical and pedagogical impacts of AI in education," in *Artificial Intelligence in Education*, Tokyo, Japan, 2023, pp. 667–673.

[11] J. Flores-Vivar, F. García-Pen¨ alvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)," *Comunicar*, 2023, doi: 10.3916/C74-2023-03.

[12] R. Hadi Mogavi, C. Deng, J. Juho Kim, P. Zhou, Y. D. Kwon, A. Hosny Saleh Metwally, A. Tlili, S. Bassanelli, A. Bucchiarone, S. Gujar, L. E. Nacke, P. Hui, "ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100027, 2024, doi: 10.1016/j.chbah.2023.100027.

[13] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, R. Buyya, "Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 19–23, 2024, doi: 10.1016/j.iotcps.2023.06.002.

[14] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Education Sciences*, vol. 13, no. 4, 2023, doi: 10.3390/educsci13040410.

[15] S. Chithrananda, G. Grand, B. Ramsundar, "ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction," *ArXiv*, vol. abs/2010.09885, 2020.

[16] Y. Wu, F. Jia, S. Zhang, H. Li, E. Zhu, Y. Wang, Y. T. Lee, R. Peng, Q. Wu, C. Wang, "An empirical study on challenging math problem solving with GPT-4," 2023.

[17] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, T. L. Griffiths, "Embers of autoregression: Understanding large language models through the problem they are trained to solve," 2023.

[18] P. Nguyen, P. Nguyen, Bruneau, L. Cao, Wang, H. Truong, "Evaluation of mathematics performance of Google Bard on the mathematics test of the vietnamese national high school graduation examination," 07 2023. doi: 10.36227/techrxiv.23691876.v1.

[19] V. Plevris, G. Papazafeiropoulos, A. Jiménez Rios, "Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT- 3.5, ChatGPT-4, and Google Bard," 2023.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[21] J. Gao, R. Pi, J. Zhang, J. Ye, W. Zhong, Y. Wang, L. Hong, J. Han, H. Xu, Z. Li, L. Kong, "G-LLaVA: Solving geometric problem with multi-modal large language model," 2023.

[22] H. Liu, C. Li, Q. Wu, Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[23] Ministerio de Educación, Argentina, *Núcleos de Aprendizajes Prioritarios. Matemática. Ciclo Básico Educación Secundaria 1° y 2° / 2° y 3° An¨ os.* 2006.

[24] R. S. Abrate, G. I. Delgado, M. D. Pochulu, "Caracterización de las actividades de geometría que proponen los textos de matemática," *Revista Iberoamericana de Educación*, vol. 39, pp. 1–9, jun. 2006, doi: 10.35362/rie3912598.

[25] M. B. López, I. B. Fernández, "Tendencias actuales de la ensenanza-aprendizaje de la geometría en educación secundaria," *Revista Internacional de Investigación en Ciencias Sociales*, vol. 8, no. 1, pp. 25–42, 2012.

[26] A. M. Bressan, K. Crego, B. Bogisic, *Razones para ensenar geometría en la educación básica: mirar, construir, decir y pensar (1a. ed.).* Novedades educativas, 2000.

[27] C. R. Suárez, T. Ángel Sierra Delgado, "Spatial problems: An alternative proposal to teach geometry in compulsory secondary education," *Educaçao Matemática Pesquisa*, vol. 22, ago. 2021, doi: 10.23925/1983-3156.2020v22i4p593-602.

[28] L. Santalo, "Olimpíadas matemáticas," *Revista de Educación Matemática*, vol. 6, ago. 2021, doi: 10.33044/revem.11101.

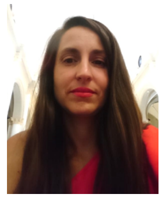
[29] P. Fauring, F. Gutierrez Eds., *Olimpiadas de Mayo - XVII a XXIV.* Buenos Aires, Argentina: Red Olimpica, 2020.

[30] B. Glass, C. Maher, "Students problem solving and justification," in *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*, vol. 2, 2004, pp. 463–470.

[31] Y. S. Eko, S. Prabawanto, A. Jupri, "The role of writing justification in mathematics concept: the case of trigonometry," *Journal of Physics: Conference Series*, vol. 1097, p. 012146, sep 2018, doi: 10.1088/1742-6596/1097/1/012146.

[32] E. Pavlick, "Symbols and grounding in large language models," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 381, no. 2251, p. 20220041, 2023, doi: 10.1098/rsta.2022.0041.

[33] G. M. Zunzarren, "The error as a problem or as teaching strategy," *Procedia - Social and Behavioral Sciences*, vol. 46, pp. 3209–3214, 2012, doi: 10.1016/j.sbspro.2012.06.038.



Verónica Parra

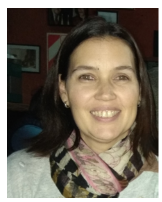
PhD in Mathematics Education from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2012. Associate professor in the Teacher Training Department at UNCPBA, member of NIEM Research Institute and Associate researcher at CONICET. Her research interests include mathematics teaching and use of resources for teaching.

Patricia Sureda

PhD in Mathematics Education from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2012. Associate professor in the Teacher Training Department at UNCPBA, member of NIEM Research Institute and Assistant researcher at CONICET. Her research interests include mathematics teaching and use of resources for teaching.

Ana Corica

PhD in Education Science from the Universidad Nacional de Córdoba (UNC), 2010. Associate professor in the Teacher Training Department at UNCPBA, director of NIEM Research Institute and Associate researcher at CONICET. Her research interests include mathematics teaching and use of resources for teaching.



Silvia Schiaffino

PhD in Computer Science from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2004. Full-time associate professor in the Computer Science Department at UNCPBA, member of ISISTAN Research Institute and Principal researcher at CONICET. Her research interests include recommender systems, user profiling and personalization.

Daniela Godoy

PhD in Computer Science from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2005. Full-time associate professor in the Computer Science Department at UNCPBA, member of ISISTAN Research Institute and Principal researcher at CONICET. Her research interests include recommender systems, social networks and text mining.