On the Use of Large Language Models at Solving Math Problems: A Comparison Between GPT-4, LlaMA-2 and Gemini

Alejandro L. García Navarro, Nataliia Koneva, José Alberto Hernández, Alfonso Sánchez-Macián *

Universidad Carlos III de Madrid (Spain)

* Corresponding author: agnavarr@pa.uc3m.es (A. L. García Navarro), nkoneva@pa.uc3m.es (N. Koneva), jahgutie@it.uc3m.es (J. A. Hernández), alfonsan@it.uc3m.es (A. Sánchez-Macián).

Received 19 April 2024 | Accepted 20 December 2024 | Early Access 4 March 2025

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In November 2022, ChatGPT v3.5 was announced to the world. Since then, Generative Artificial Intelligence (GAI) has appeared in the news almost daily, showing impressive capabilities at solving multiple tasks that have surprised the research community and the world in general. Indeed the number of tasks that ChatGPT and other Large Language Models (LLMs) can do are unimaginable, especially when dealing with natural text. Text generation, summarisation, translation, and transformation (into poems, songs, or other styles) are some of its strengths. However, when it comes to reasoning or mathematical calculations, ChatGPT finds difficulties. In this work, we compare different flavors of ChatGPT (v3.5, v4, and Wolfram GPT) at solving 20 mathematical tasks, from high school and first-year engineering courses. We show that GPT-4 is far more powerful than ChatGPT-3.5, and further that the use of Wolfram GPT can even slightly improve the results obtained with GPT-4 at these mathematical tasks.

Keywords

ChatGPT, Generative AI, Mathematical Problems, Wolfram Mathematica.

DOI: 10.9781/ijimai.2025.03.001

I. Introduction

The development of large language models (LLMs) represents a significant milestone in Artificial Intelligence (AI) and Natural Language Processing (NLP), built upon decades of research and advancements in machine learning (ML), data availability, and computational power [1].

The genesis of ChatGPT[1] and other LLMs can be traced back to the 2010s, a pivotal decade marked by significant research breakthroughs in neural networks and NLP. This period witnessed the emergence of advanced deep neural network architectures, notably Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [2], [3]. These architectures were instrumental in enhancing sequential data processing, a critical component for effectively handling language-related tasks. Furthermore, the development and adoption of word embeddings, exemplified by innovations like Word2Vec and GloVe, revolutionized the approach to linguistic data handling in models [4], [5]. These embeddings facilitated the creation of more nuanced and semantically rich representations of words, significantly advancing the capabilities of language models.

Later, attention mechanisms and the Transformer architecture, introduced in papers like "Attention Is All You Need" (2017) [6], led to significant improvements in language understanding and generation. Finally, the introduction of models like GPT (Generative Pretrained Transformer) by OpenAI [7] and BERT (Bidirectional Encoder Representations from Transformers) by Google [8] marked a new era in language models. These models, pre-trained on vast amounts of text, demonstrated remarkable language understanding and generation capabilities.

In the 2020s decade, OpenAI's GPT-3 and its successors, including ChatGPT, showcased the power of large language models in generating human-like text. These models are characterized by their deep learning architecture, massive scale (billions of parameters), and ability to perform a wide range of language tasks without task-specific training [9],[10].

After ChatGPT v3.5 was introduced in November 2022 and rapidly adopted worldwide, the research community was surprised again with the release of GPT-4 in March 2023, showing enhanced reasoning, improved capabilities at generating images from text and understanding images as input [11]–[15]. In addition, GPT-4 was enhanced with multiple plugins to allow better specialization at

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

Please cite this article as: A. L. García Navarro, N. Koneva, J. A. Hernández, A. Sánchez-Macián. On the use of Large Language Models at Solving Math Problems: A Comparison Between GPT-4, LlaMA-2 and Gemini, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 5, pp. 40-50, 2025, http://dx.doi.org/10.9781/ijimai.2025.03.001

certain tasks. One of them was the Wolfram Mathematica Plugin[2] which allows interacting with Wolfram libraries and functions for better mathematical reasoning, processing and visualization from natural text as input. In particular, logic, reasoning, and math-solving were found to be one of the main weaknesses of GPT models, since they were mainly trained to work with text. Remark that **Wolfram Mathematica**[3] is a computing environment used for mathematical computation, algorithm development, data visualization, and symbolic manipulation, widely used in scientific, engineering, mathematical, and computing fields. These advancements not only signify a rise in AI's problem-solving abilities but also underscore the transformative potential of AI in educational settings [16]–[18]. Integrating AI tools in educational procedures proposes a change towards more interactive, adaptive, and personalized learning experiences, showing the role AI could play in promoting both teaching and learning outcomes [14].

Nonetheless, as of April 9, 2024 conversations with plugins could no longer be continued. Instead, GPTs were created based on feedback from users [19]. GPTs are custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills [20]. Most plugins have been transformed into GPTs, including the previously mentioned Wolfram Mathematica plugin, now referred to as Wolfram GPT [21].

The goal of this article is to evaluate the ability of the most popular free-to-use LLMs to understand and solve mathematical problems with varying levels of specificity, determining whether it requires detailed background context for effective problem-solving or if they can efficiently derive solutions from minimal information. We show that GPT-4 is far more powerful than its earlier version GPT-3.5, and that Wolfram GPT may slightly enhance its performance. Other LLMs like LLaMa-v2 and Gemini show similar performance to GPT-v3.5.

The remainder of this study is organized as follows: Section II reviews the most recent and relevant studies regarding the ability of LLMs to solve mathematical problems. Next, Section III outlines the problem statement and methodology conducted aimed at understanding the ability of the most-popular free-to-use LLMs at solving 20 mathematical tasks, which are outlined in Section IV along with the performance of the LLMs. Finally, Section V concludes this study with a summary of its main findings, conclusions, and future work.

II. Related Work

The capabilities of LLMs in problem-solving, particularly in mathematical reasoning, have been a subject of growing interest in recent research. As these models continue to evolve, there is an increasing need to evaluate their performance across various domains and compare them to human abilities.

Authors in the paper [22] (2023) provide a valuable comparison between general-purpose language models like ChatGPT and GPT-4 and models specifically trained for single mathematical tasks. Their findings indicate that while specialized models outperform ChatGPT and GPT-4 in specific areas, they lack the flexibility and universal applicability of these general-purpose models. This insight contributes to the ongoing discussion about the trade-offs between specialized and general AI systems in the context of mathematical problem-solving.

A significant contribution to this field comes from Cherian et al. (2024) [23], where it provided valuable insights into the mathematical reasoning abilities of Large Vision and Language Models (LVLMs). By creating the SMART-840 dataset, comprising 840 problems from the Mathematical Kangaroo Olympiad, the authors offer a systematic approach to comparing AI performance with that of children across different age groups. LVLMs struggle with problems designed for younger children, indicating a lack of foundational knowledge, but demonstrate increasingly powerful reasoning skills when solving problems designed for higher grades. This observation suggests that current machine-learning approaches may not be capturing the fundamental competencies that underlie human reasoning. The recent survey by Ahn et al. (2024) [24] provided a comprehensive survey of LLMs in mathematical reasoning, covering various problem types, datasets, and techniques using GPT-4. Their work also highlights persistent challenges, such as the need for more robust foundational understanding and human-centric approaches in math education. These studies collectively demonstrate the evolving landscape of AI in mathematical reasoning, showcasing both the significant advancements and the remaining challenges in developing LLMs capable of human-like mathematical problem-solving.

Recent research has also focused on enhancing LLMs' mathematical reasoning capabilities through various techniques. Imani et al. (2023) [25] introduced 'MathPrompter', which uses Zero-shot chain-of-thought prompting to generate multiple solutions for the same problem, improving performance and confidence in results. Similarly, Zhou et al. [26] proposed a code-based self-verification method for GPT-4 Code Interpreter, significantly boosting its performance on challenging math datasets.

Davis and Aaronson (2023) [27] conducted a comprehensive test of GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on 105 original science and math problems at high school and college levels. Their findings suggest that while these plug-ins significantly enhance GPT-4's problem-solving abilities, interface failures remain a central challenge, particularly in formulating problems to elicit useful responses from the plug-ins.

Building upon these foundational studies, our work contributes significantly to the field by offering a comprehensive comparison of multiple state-of-the-art LLMs, including GPT-4, LlaMA-2, and Gemini, as well as Wolfram GPT, in the domain of mathematical problem-solving, identifying ChatGPT-4 as the best candidate. While previous research has focused on either specialized tasks or broader cognitive assessments, our study bridges this gap by evaluating these models on a carefully curated set of mathematical problems using both Zero-shot and Zero-shot Chain of Thought techniques, showing that the latter technique behaves better in most situations. Finally, we have shown that, in general, Wolfram GPT does not provide a competitive advantage when solving the tested mathematical problems with respect to GPT-4.

III. Problem Definition and Methodology

A. Problem Definition

In a nutshell, the problem we aim to address is whether existing free-to-use popular Large Language Models can be effectively used to solve mathematical problems. To explore this, we are conducting a thorough validation process, testing several models on a variety of mathematical problems using different prompting techniques.

Our objective is to identify which models demonstrate superior problem-solving capabilities and determine the reliability of these models in providing accurate solutions. This analysis will help us understand the strengths and limitations of current LLMs in the context of mathematical problem-solving, guiding future developments and applications in this field.

B. Methodology

In the rapidly evolving field of artificial intelligence, particularly in the domain of language models, "prompting" [28] emerges as a pivotal

[2] https://www.wolfram.com/wolfram-plugin-chatgpt/

[3] https://www.wolfram.com/mathematica/

concept. This term refers to the method of interacting with an AI language model by providing it with specific inputs (prompts), which guide the model in generating a desired output.

As previously stated in the introduction, we intend to understand how LLMs interpret and respond to prompts with varying levels of specificity. This section delves into prompting, the main method used to carry out this study.

C. What Is a Prompt?

A prompt is essentially an input statement or question given to an AI model. It acts as a catalyst that initiates the model's generation process, leading to a variety of potential outputs. Prompts can vary significantly in complexity, ranging from simple questions to detailed instructions or scenarios.

The development of PromptGen, a model that automates prompts generation by transforming input sentences into prompts, showcases the complexity and importance of prompt engineering in AI interactions [29]. Further research into prompt patterns for conversational LLMs, such as ChatGPT, categorizes prompts into several types—Output Customization, Error Identification, Prompt Improvement, Interaction, and Context Control—demonstrating the depth of prompt engineering's role in enhancing AI model responsiveness and interaction quality [30].

D. Prompting Techniques

In the course of this paper, we focus particularly on two advanced prompting techniques [31]: **zero-shot learning** and **zero-shot chain of thought** (zero-shot-CoT).

- Zero-shot learning: It is a technique where the AI model responds to prompts without any prior specific training or examples related to that task [32]. It relies on the model's pre-trained knowledge and its ability to generalize from that knowledge to new scenarios.

  Let us take an example from the prompts shown in Appendix A.

  A prompt like "Find the minimum of the function f(x)=(x^2 − x − 2)/(x^2 − 6x+9)" would be handled without the model having seen this exact sentence before.

- Zero-shot-CoT: It constitutes a nuanced extension of the zero-shot technique. In this approach, the model is prompted to articulate its reasoning process step by step, leading to the final answer. This method not only sheds light on the model's decision-making process but also enhances the clarity and interpretability of its responses [33]. To do this, the sentence "Let's think step by step" is added at the end of the prompt.

  For instance, consider an example from Appendix B.

  A prompt like "Find the minimum of the function f(x) = (x^2−x−2) / (x^2−6x+9). Let's think step by step" would encourage the model to break down the calculation process step by step.

E. Available Resources

As of right now, we found out that there are no specific texts on the most efficient way of prompting Wolfram GPT. However, there is an interesting introduction to using the GPT in the official Wolfram webpage: **Wolfram GPT**. It provides an installation guide and some applications of it and, at the end of the page, there is a link about **getting to know the Wolfram GPT**, where several prompts are tested.

F. Using Wolfram GPT

Combining ChatGPT and Wolfram Mathematica can be a powerful way to leverage the strengths of both platforms. ChatGPT is proficient in natural language processing and can handle a wide range of queries and tasks, while Wolfram Mathematica excels in computational mathematics, data analysis, and visualization. Here are a few ways to integrate them:

1. **Automating Mathematica Scripts**: ChatGPT can be used to create a user-friendly interface for creating Mathematica scripts. Users can describe in natural language what they want to compute or analyze, and then ChatGPT will translate this into a Mathematica script and execute it (e.g. [34]).
2. **Data Analysis and Visualization**: ChatGPT can be used to interpret and structure data analysis queries. After passing these structured queries, Mathematica can then perform complex data analysis and generate visualizations.
3. **Algorithm Design and Problem optimization**: Combining ChatGPT and Mathematica for algorithm development can be achieved by using ChatGPT for initial brainstorming and pseudocode generation, and then translating these ideas into Mathematica's powerful computational language for detailed analysis and visualization, or even solving optimization problems.

The GPT is only one of the many available for ChatGPT Plus. The following lists the instructions necessary to use it:

1. Select **Explore GPTs** in the side bar.
2. Once selected, **search for Wolfram** in the search tab.
3. Click on **Start Chat**.

IV. Evaluation

To carry out our goal, we provided some mathematical problems with varying difficulty. To ensure our comparison extends beyond just ChatGPT, we conducted our experiments on Gemini [35] and LLaMA [36] as well. This approach allows us to provide a more comprehensive evaluation by including multiple AI models in our analysis, thereby avoiding any bias that may arise from focusing solely on ChatGPT. By examining the performance and capabilities of Gemini and LLaMA alongside ChatGPT, we aim to achieve a balanced and wide-ranging understanding of the current landscape of conversational AI technologies.

There will be five different sections of evaluations, each corresponding to a different model: ChatGPT-3.5, Wolfram GPT, ChatGPT-4, Gemini, and LLaMA. Note that ChatGPT-4 solves mathematical problems using Python typically, and related libraries like `sympy`, `numpy`, or even `math`.

Each model was accessed on its dedicated page via a web browser, and different sessions were created for each type of prompting to ensure consistency and isolation of experiments. Inside each evaluation section, we applied two different types of prompting: Zero-shot and Zero-shot-CoT. For each one, we tested their performance by being very specific about what we wanted to do and by not being specific about what we wanted to do.

That is, for example, instead of asking to solve an integral in a range (specific instructions), we could ask to obtain the area and check if it knows that an integral must be computed (general instructions). With this, we aim to validate the possibility of using existing LLMs to automatically solve mathematical problems and see to what extent they are reliable in providing solutions.

A. Mathematical Problems

In the following lines, we will be introducing the mathematical problems, which happen to be classified into **Specific Instructions** and those with **General Instructions**, each containing 10 problems:

- Specific Instructions

1. **Level 1: Inverting a matrix**

   Given the matrix A =

   $$\begin{bmatrix} 1 & -2 & 1 \\ -2 & 3 & 1 \\ 5 & -7 & -3 \end{bmatrix}$$

   find the inverse (A^{-1}).

2. **Level 2: Generating normally distributed variables with a specific mean and variance**

   Generate several normally distributed variables with mean = 45.6 and variance = 13.84.

3. **Level 3: Plotting a function**

   Plot $-3\,(x-2)^2 - 5$.

4. **Level 4: Finding the minimum of a one-variable function**

   Find the minimum of the function $f(x) = \frac{x^2-x-2}{x^2-6x+9}$.

5. **Level 5: Intersecting two functions**

   Find the intersection points of the functions

   $f(x) = |x-5|$ and $g(x) = \log x$

6. **Level 6: Differentiating a function**

   What is the derivative of $f(x) = \frac{5}{\sqrt{3x-1}}$?

7. **Level 7: Integrating a function**

   What is the integral of $sin(2x)\ cos(2x)$?

8. **Level 8: Mathematical series**

   Study the convergence of $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$

9. **Level 9: Fourier transform**

   Compute the Fourier transform of $f(t) = cos(w_0\,t)$.

10. **Level 10: Doing regression in data**

    Table I shows information regarding the sales of daily press in the year 1998, as the number of daily copies sold per thousand inhabitants in 8 autonomous Spanish regions. The sales are assumed to be related to economic activity levels as measured by the Gross Domestic Product (GDP) per inhabitant in thousands of euros (Source: INE. Anuario Estadistico). Use least squares to estimate a simple regression model that explains the number of copies sold as a function of GDP per capita.

TABLE I. GDP and Copies Sold in 1998

GDP	Copies Sold
8.3	57.4
9.7	106.8
10.7	104.4
11.7	131.9
12.4	144.6
15.4	146.4
16.3	177.4
17.2	186.9

- General Instructions

11. **Level 11: Finding the area (requires knowing that an integral must be used)**

    Find the area bounded by the curve $y = x^2 + x + 4$, the x-axis and the ordinates $x = 1$ and $x = 3$.

12. **Level 12: Predicting the range of a projectile (requires knowing that kinematic equations must be used)**

    A soccer player kicks a ball at an angle of 30 degrees to the horizontal. The initial speed of the ball is 20 meters per second. Assuming no air resistance and that the ball is kicked from ground level, predict how far the ball will travel horizontally before hitting the ground. Use the acceleration due to gravity as 9.8m/s^2.

13. **Level 13: Getting the best path from one node to another (requires knowing that Dijkstra's algorithm must be used)**

    Imagine I have a graph (with nodes s, s1, s2, s3, s4, s5, t), whose connections (with weights) are represented with the following adjacency matrix:

    $$\begin{bmatrix} 0 & 1 & 3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 8 & 4 & 0 & 0 \\ 3 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 & 1 & 7 \\ 0 & 4 & 3 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 10 \\ 0 & 0 & 0 & 7 & 0 & 10 & 0 \end{bmatrix}$$

    Obtain the shortest path from *s* to *t*.

14. **Level 14: Predicting the next number in a complex sequence (requires recognizing and applying the underlying pattern or mathematical rule governing the sequence)**

    Consider the following sequence of numbers: 3, 8, 15, 24, 35, 48, ... Your task is to predict the next number in this sequence.

15. **Level 15: Calculating the future value of an investment (requires understanding and applying the compound interest formula)**

    If you invested $5,000 for 10 years at 9% compounded quarterly, how much money would you have? What is the interest earned during the term?

16. **Level 16: Deciphering a phrase knowing each letter was shifted by a certain number (requires knowing that we are talking about Caesar's cipher decoding)**

    Decode this phrase "SERR CVMMN VA GUR PNSRGREVN" knowing that each letter was shifted by 13.

17. **Level 17: Calculating the time it takes for an object to cool down to a certain temperature (requires the use of Newton's Law of Cooling, which often involves solving differential equations)**

    A freshly baked pie is taken out of the oven and left to cool in a room. The temperature of the oven was 200°C, and the room temperature was a constant 25°C. When the pie is first taken out, its temperature is 180°C. After 20 minutes, the temperature of the pie drops to 100°C. Calculate the time it takes for the pie to cool down to 50°C.

18. **Level 18: Calculating the work done in compressing a spring (requires knowing Hooke's Law)**

    A person compresses a spring a distance of 5cm, which requires a force of 100N. How much work does the person do?

19. **Level 19: Computing the quantity a company should make for its inventory given production cost, demand rate, and other variables (requires understanding Economic Order Quantity)**

    The John Equipment Company estimates its carrying cost at 15% and its ordering cost at $9 per order. The estimated annual requirement is 48,000 units at a price of $4 per unit. What is the most economical number of units to order? How many orders should be placed in a year? How often should an order be placed?

20. **Level 20: Computing the orbital period (requires knowing Kepler's third law of planetary motion)**

    Titan, the largest moon of Saturn, has a mean orbital radius of 1.22x10^9 m. The orbital period of Titan is 15.95 days. Hyperion, another moon of Saturn, orbits at a mean radius of 1.48x10^9 m. Predict the orbital period of Hyperion in days.

Note: The specific prompts used for each problem are detailed in Appendix A (When doing Zero-shot-CoT, the sentence "Let's think step by step" was added at the end of each problem statement).

B. ChatGPT-3.5

In our exploration of zero-shot prompting, we delved into two primary categories of exercises: "Specific Instructions" and "General Instructions". Initially focusing on the former category, the performance of ChatGPT-3.5 exhibited partial success. The evaluation encompassed ten distinct levels, of which the model accurately executed three. Notably, ChatGPT-3.5 successfully navigated Level 2, Level 3, and Level 8 with complete correctness. Conversely, it encountered challenges in providing accurate solutions for Levels 1, 4, 5, 6, 7, 9, and 10.

A closer analysis reveals the varying nature of errors across these levels:

- Levels 1, 4, 6, and 10: ChatGPT-3.5 demonstrated a correct methodological approach. However, computational inaccuracies led to incorrect final answers.
- Level 5: In this case, it delineated the procedural steps but did not carry out all of them.
- Level 7: The initial decision to employ trigonometric identities for transforming the expression was sound. Nonetheless, the incorrect application of these identities resulted in an incorrect outcome.
- Level 9: The error was due to inadequate consideration of the 'Euler function,' crucial for simplifying the expression.

In the evaluation under the "General Instructions" category, ChatGPT-3.5 showed moderately improved performance compared to its results in the "Specific Instructions" category. Out of ten tasks, the model accurately solved half, delivering correct responses for Levels 11, 12, 14, 15, and 20. However, it encountered difficulties with Levels 13, 16, 17, 18, and 19.

A detailed analysis of these levels reveals the following insights:

- Level 13: Successfully identified the correct methodology but faltered in its application, leading to an incorrect solution.
- Level 16: ChatGPT-3.5 accurately found the underlying ciphering technique, but did not apply the shifting process correctly.
- Level 17: Even though it was understood that it was a problem related to Newton's Law of Cooling, the answer does not coincide with the solution.
- Level 18: While understanding the problem, ChatGPT-3.5 failed to account for the spring constant in its calculations, leading to an erroneous outcome.
- Level 19: It knew it was a problem of Economic Order Quantity (EOQ), but the data was used incorrectly.

Further, our evaluation extended to zero-shot-CoT prompting. In the "Specific Instructions" category, ChatGPT-3.5 achieved a moderate level of success. Out of ten levels, five were solved correctly. The model was particularly effective in Levels 2, 3, 7, 8, and 9, suggesting an improved performance in tasks that can benefit from chain-of-thought reasoning. However, the model encountered specific issues in the other levels, as detailed below:

- Level 1: The methodology was correctly identified, but the model failed in executing computations, such as calculating the determinant.
- Level 4: This version of ChatGPT provided the correct methodology but left steps unfinished, offering only explanations on how to proceed without completing them.
- Level 5: Although the methodology was outlined, none of the procedural steps were carried out.
- Level 6: The model's methodology was only partially correct, notably omitting the application of the chain rule.
- Level 10: The methodology was correctly provided, but the final steps necessary to complete the task were not executed.

For the "General Instructions" under zero-shot-CoT prompting, ChatGPT-3.5's performance was limited. The results are stated as follows:

- Level 11: It recognized the need for an integral, but computed it incorrectly.
- Level 12: The solution was incorrect, and likely stemmed from a miscalculation in applying the quadratic formula.
- Level 13: While the methodology was correct, the steps were performed incorrectly.
- Level 16: ChatGPT-3.5 accurately identified the underlying ciphering technique but failed in the execution of the shifting process, reflecting correct methodology but poor application.
- Level 17: It understood the problem but incorrectly assumed missing information, affecting the solution.
- Level 18: The correct methodology was presented, but calculation errors led to an incorrect solution.

These results indicate that only four out of the ten tasks were solved correctly. Successes were noted in Levels 14, 15, 19, and 20, but there were evident struggles in the majority of the levels.

C. Wolfram GPT

In the assessment of zero-shot prompting under specific guidelines, Wolfram exhibited unparalleled performance by solving 9/10 tasks presented to it.

Moving to the general problems category within the same prompting framework, Wolfram continued to display a high degree of precision, successfully addressing nine out of ten tasks. The sole exception occurred at Level 13, in which the shortest path could not be found.

In scenarios utilizing zero-shot-CoT prompting, the GPT's performance remained stellar in the specific statements category, where it achieved a flawless score.

However, when evaluated under general guidelines with zero-shot-CoT prompting, it encountered minor obstacles, particularly at Level 13, where difficulties arose from a mistake in updating distances and selecting nodes.

D. ChatGPT-4

In the experiments conducted with ChatGPT-4 under zero-shot prompting, the model exhibited outstanding performance across both categories of exercises. Remarkably, all twenty tasks were solved correctly.

Significantly, the correct solutions and methodology were successfully received for all ten tasks of Levels 11 to 20 guided only by general instructions.

TABLE II. Performance Comparison of ChatGPT Versions and Wolfram GPT With Zero-Shot-CoT (ZS-CoT) and Zero-Shot Prompting

Level	Problem	ChatGPT-3.5		ChatGPT-4		Wolfram GPT	
		Zero-shot	ZS-CoT	Zero-shot	ZS-CoT	Zero-shot	ZS-CoT
1	Matrix inversion	M	M	S	**M+S**	**M+S**	**M+S**
2	Normal Distribution Generation	**M+S**	**M+S**	S	**M+S**	**M+S**	**M+S**
3	Function Plotting	**M+S**	**M+S**	S	**M+S**	**M+S**	**M+S**
4	Finding Minimum	M	M	**M+S**	**M+S**	**M+S**	**M+S**
5	Function Intersection	M	M	**M+S**	**M+S**	**PI**	**M+S**
6	Function Differentiation	M	**NA**	**M+S**	**M+S**	**M+S**	**M+S**
7	Function Integration	M	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**
8	Series Convergence	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**
9	Fourier Transform	**NA**	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**
10	Data Regression	M	M	**M+S**	**M+S**	**M+S**	**M+S**
11	Area Calculation	**M+S**	M	**M+S**	**M+S**	**M+S**	**M+S**
12	Projectile Range Prediction	**M+S**	M	**M+S**	**M+S**	**M+S**	**M+S**
13	Shortest Path Finding	M	M	**M+S**	**M+S**	**PI**	**PI**
14	Sequence Prediction	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**
15	Compound Interest Calculation	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**
16	Caesar's Cipher Decoding	M	M	**M+S**	**M+S**	**M+S**	**M+S**
17	Cooling Time Calculation	**NA**	M	**M+S**	M	**M+S**	**M+S**
18	Spring Compression Work	M	M	**M+S**	**M+S**	**M+S**	**M+S**
19	Inventory Optimization Analysis	M	**M+S**	**M+S**	**PI**	**M+S**	**M+S**
20	Orbital Period Computation	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**	**M+S**

Abbreviations used in the table: M (Methodology-only), M+S (Methodology+Solution), NA (Nothing), S (Solution-only), PI (Partial/Incorrect Solution).

In the final part of ChatGPT's evaluations, ChatGPT-4, employing zero-shot-CoT prompting, demonstrated commendable performance with the problems described with detailed guidelines, successfully solving all of the problems.

Regarding the results for Levels 11 to 20, although the model showed a high level of proficiency in most tasks, it encountered difficulties in Level 17 and Level 19. Regarding the former, it couldn't obtain the cooling constant; in the latter, even though it answered two questions correctly, it was not able to answer the last one.

E. Google Gemini

When evaluating Gemini's performance across different categories and levels, we observed that there are some areas of improvement in the model. The results will be commented on in further paragraphs.

Under zero-shot prompting, Gemini's computational errors across Levels 4, 5, 6, 7, 12, 16, 19, and 20 demonstrate a consistent challenge in the calculation, irrespective of the instruction type. Similarly, the belief that there is missing information as an obstacle in Levels 10, 17, and 18 reflects a common barrier faced by Gemini in both specific and general contexts.

Similar failing patterns can be observed under zero-shot-CoT prompting: there may be a correct understanding of the process but there are errors in calculations (Levels 3, 7, 11, 12, 16, 19), Gemini believes there is a lack of information in the given instructions (Level 14), or it forgets to consider key steps for solving a problem (Level 18).

F. Meta LLaMA-2-70b

LLaMA, similarly, shows varied performance, but it seems to struggle more consistently than Gemini, especially with zero-shot prompting, where it often provides partial or incorrect solutions.

Let's begin with the first type of prompting. In evaluating this first category, LLaMA correctly solved four exercises, and there were errors along the lines of:

- Levels 1, 5, 6, 11, 12, 13, 15, 16, 18, 20: Correct methodology, errors in calculations.
- Level 4: There wasn't a clear methodology, and the results were incorrect.
- Level 7: The model explained the initial step but then asked the user to complete it independently.
- Levels 9, 19: Incorrect methodology, incorrect solution.
- Level 10: Carried out the problem using R, but considered logarithmic data (unnecessary step).
- Level 14: Incorrectly guessed the sequence.

Finally, regarding zero-shot-CoT prompting, these results show the lowest performance of the evaluations. Let's observe the errors:

- Levels 1, 4, 10, 15, 16: Correct methodology, errors in calculations.
- Levels 5, 6, 7, 9, 11, 12, 19, 20: Incorrect methodology, incorrect solution.
- Level 8: Incorrect identification of the series (it is a telescopic series, not a harmonic series).
- Level 13: Did not use Dijkstra's algorithm.
- Level 14: Incorrectly guessed the sequence.
- Level 17: It did not understand that we were dealing with Newton's Law of Cooling.
- Level 18: Forgot to consider the spring constant.

G. Results

Tables II and III outline the results of all five LLMs used in the experiments: ChatGPT-v3.5, ChatGPT-v4, Wolfram GPT, Google Gemini and Meta LLaMA-2-70b. In each case, the LLMs are evaluated as:

- (NA) stands for Nothing.
- (PI) stands for Partial or Incorrect solution provided.
- (M) stands for (correct) Methodology only, where the LLM explains the theoretical background to solve the problem, but no solution is provided.
- (S) stands for (correct) Solution only, without any explanation about how it was obtained.

TABLE III. Performance Comparison of Gemini and LLaMA With Zero-Shot-CoT (ZS-CoT) and Zero-Shot Prompting

Level	**Problem**	**Gemini**		**LLaMA**	
		Zero-shot	**ZS-CoT**	**Zero-shot**	**ZS-CoT**
1	Matrix inversion	S	**M+S**	**PI**	**PI**
2	Normal Distribution Generation	S	**M+S**	S	M
3	Function Plotting	M	M	S	M
4	Finding Minimum	**PI**	**PI**	**PI**	**PI**
5	Function Intersection	**PI**	**M+S**	**PI**	**PI**
6	Function Differentiation	**PI**	**PI**	**PI**	**PI**
7	Function Integration	**PI**	**M+S**	**NA**	**PI**
8	Series Convergence	**PI**	**M+S**	**M+S**	**PI**
9	Fourier Transform	**PI**	**M+S**	**PI**	**PI**
10	Data Regression	M	**M+S**	**PI**	**PI**
11	Area Calculation	**M+S**	**PI**	**PI**	**PI**
12	Projectile Range Prediction	**PI**	**PI**	**PI**	**PI**
13	Shortest Path Finding	**M+S**	**M+S**	**PI**	**PI**
14	Sequence Prediction	**M+S**	**NA**	**PI**	**PI**
15	Compound Interest Calculation	**M+S**	**M+S**	**PI**	**PI**
16	Caesar's Cipher Decoding	**PI**	**PI**	**PI**	**PI**
17	Cooling Time Calculation	**NA**	**M+S**	**M+S**	**PI**
18	Spring Compression Work	**NA**	**M+S**	**PI**	**PI**
19	Inventory Optimization Analysis	**PI**	**PI**	**PI**	**PI**
20	Orbital Period Computation	**PI**	**M+S**	**PI**	**PI**

Abbreviations used in the table: M (Methodology-only), M+S (Methodology+Solution), NA (Nothing), S (Solution-only), PI (Partial/Incorrect Solution).

- (M+S) stands for both (correct) Methodology and Solution provided by the LLM.

The conclusions from our project, focusing on the performance of ChatGPT-3.5, ChatGPT-4, Wolfram GPT, Gemini, and LLaMA-2-70b can be articulated in the following paragraphs.

On the one hand, **ChatGPT-3.5** displayed a variable level of proficiency, particularly under the zero-shot prompting framework. It demonstrated partial success in specific instructions, handling some tasks with accuracy while struggling with others, particularly in complex mathematical problems. This performance indicated the model's capability to understand basic instructions and methodologies but also highlighted its limitations in executing detailed computational steps. In the zero-shot-CoT (Chain of Thought) prompting, there was a slight improvement, suggesting that the model benefits from a structured approach to problem-solving, particularly in breaking down complex tasks. However, even with this approach, the model faced challenges in fully executing procedures and providing accurate solutions in more complex scenarios.

There might be several underlying reasons that explain these results, and the most obvious one is related to its number of parameters. The model has 175 billion parameters, enabling it to handle a wide range of tasks. However, this number, while substantial, is relatively small compared to current models, which introduce some limitations. Secondly, the transformer architecture of ChatGPT-3.5 enables it to process and generate text effectively, but complex mathematical problem-solving often requires more specialized computational techniques that are not fully developed in this model.

On the other hand, **Wolfram GPT** marked a significant advancement in the model's problem-solving abilities. Under zero-shot prompting, it showcased exceptional performance, successfully solving a wide range of mathematical tasks with specific instructions (except for two problems). This indicated a robust capability in handling computational and analytical problems, likely benefiting from the computational power of the GPT. Under the zero-shot-CoT framework, it continued to display strong capabilities, though it encountered some challenges in updating distances and selecting nodes for, particularly under general instructions. These instances highlighted that while external computational tools significantly enhance AI capabilities, understanding the nuances of complex problem statements remains an area for further development.

Wolfram GPT benefits greatly from Wolfram Mathematica's power, since it boosts the way it deals with complex mathematical tasks, as reflected in its performance. However, the challenges faced suggest limitations in the model's intrinsic understanding of problem contexts without explicit computational guidance. This may be due to the model's reliance on external computational tools, highlighting an area where its internal reasoning capabilities could be further developed.

Remarkably, **ChatGPT-4's performance** also demonstrated outstanding capabilities in handling a diverse array of mathematical problems. This was evident in both specific and general instructions under zero-shot prompting, where the model successfully tackled all tasks. This performance underscores the inherent strength of the model in understanding and addressing complex problems. With zero-shot-CoT prompting, although the model showed high proficiency in most tasks, it encountered difficulties in specific complex scenarios, particularly where advanced problem-solving strategies were required.

Let us take into account that ChatGPT-4, with its enhanced architecture and increased number of parameters to around 1 trillion, demonstrates exceptional performance by using its pre-trained knowledge and advanced processing algorithms. But oddly enough, in zero-shot-CoT prompting, where step-by-step reasoning is required, the model's performance reveals certain limitations. These difficulties often arise in highly complex scenarios that demand advanced problem-solving strategies beyond the model's current capabilities. To address these challenges, enhancing the model's reasoning algorithms to better break down complex tasks is crucial. Additionally, it could also benefit from more extensive training on complex problem-solving techniques.

Drawing conclusions from the performance evaluations of Gemini and LLaMA-2-70b in comparison to ChatGPT versions and Wolfram GPT, the findings reveal distinct capabilities and limitations across these LLMs in handling complex problem-solving tasks under different prompting strategies.

Gemini exhibited strengths in certain areas but demonstrated consistent challenges with computational accuracy and a tendency to perceive missing information in prompts, which hindered its performance. This pattern was evident in both zero-shot and zero-shot-CoT scenarios, indicating a need for improvement in Gemini's computational abilities and its understanding of provided instructions. Despite understanding the process correctly in several cases, Gemini's computational errors and occasional omission of key steps suggest that while its conceptual grasp is on the right track, its execution requires refinement.

LLaMA-2-70b showed a broader range of difficulties, particularly with zero-shot prompting, where it frequently provided partial or incorrect solutions. The issues ranged from calculation errors to incorrect methodologies and misunderstanding of problem statements, being these challenges more pronounced under zero-shot-CoT prompting.

Being the model with the least number of parameters among those evaluated (70 billion), it faced limitations due to its smaller size relative to other models. This smaller parameter count affects its ability to handle complex problems, often resulting in partial or incorrect solutions, as observed in Table 3. Similar to previous models, LLaMA-2-70b could benefit from more extensive and targeted training on complex mathematical and logical reasoning tasks.

Before ending this section, it could be beneficial to express these results numerically. We consider as correct answers those that provide either the solution (S) or the methodology and the solution (M+S). Let's observe the results in the next lines, as well as in the bar graph 1:

- ChatGPT-3.5
  - Zero-shot: 8/20
  - Zero-shot-CoT: 9/20
- ChatGPT-4
  - Zero-shot: 20/20
  - Zero-shot-CoT: 18/20
- Wolfram GPT
  - Zero-shot: 18/20
  - Zero-shot-CoT: 19/20
- Gemini
  - Zero-shot: 6/20
  - – Zero-shot-CoT: 12/20
- LLaMA
  - Zero-shot: 4/20
  - Zero-shot-CoT: 0/20

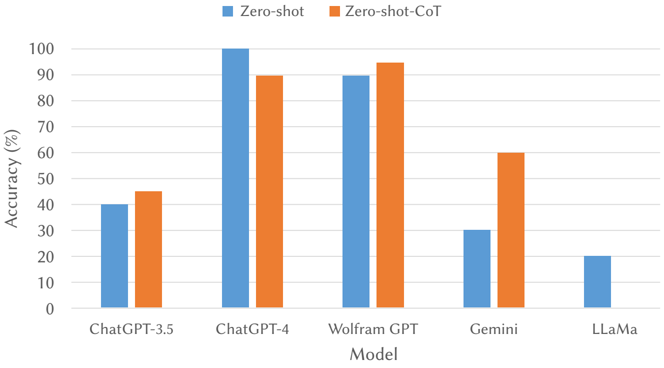

Fig. 1. Accuracy percentages of Methodology+Solution (M+S) and Solution (S) for each model across different promptings.

V. Summary and Conclusions

This article has analysed the capabilities of different Large Language Models (GPT, Gemini and LLaMA-2) at solving different mathematical tasks. We observe that GPT-4 and Wolfram GPT outperforms all others at these tasks.

In particular, GPT-4 using zero-shot prompting performed better than Wolfram GPT in all the categories, achieving perfect scores. When using them with Chain-of-Thoughts prompting, both performed perfectly in the "Specific Instructions" category. However, in the "General Instructions" category, Wolfram GPT performed slightly better, solving 9 out of 10 tasks correctly, just like ChatGPT-4, but provided solutions to all the problems.

Overall, ChatGPT-4 and Wolfram GPT demonstrated almost perfect performance at all the tasks used in the analysis, solving 97.5% and 92.5% of them, respectively. Gemini and LLaMA, on the other hand, had difficulties solving a vast majority of the tasks, with overall success rates of 45.0% and 10.0%, respectively.

In conclusion, we have identified ChatGPT-4 as the best candidate overall. Our comparison of prompting techniques revealed that Zero-shot Chain of Thought generally performs better than Zero-shot for solving mathematical problems. Additionally, while Wolfram GPT did not significantly enhance performance in solving the tested problems, it provided more detailed explanations compared to using ChatGPT-4.

Future work could benefit from exploring additional prompting techniques, such as few-shot learning, which involves providing the model with a few examples to learn from before new tasks. Additionally, we could evaluate a broader range of LLMs with varying architectures and parameter sizes, allowing us to understand the current limitations of conversational AI technologies. Finally, trying out more complex mathematical problems could further challenge the models and identify areas for improvement.

Appendix

A. Zero-Shot Prompting

Here are the detailed prompts used to analyze the responses of the models under various mathematical problems:

- Specific Instructions

1. **Level 1**: Given the matrix A = [[1, -2, 1], [-2, 3, 1], [5, -7, -3]], find the inverse (A^{-1}).
2. **Level 2:** Generate several normally distributed variables with mean = 45.6 and variance = 13.84.
3. **Level 3:** Plotting a function -3*(x-2)^2-5.
4. **Level 4:** Find the minimum of the function f(x)=(x^2- x-2) / (x^2-6x+9).
5. **Level 5:** Find the intersection points of the functions

   $f(x) = |x - 5|$ and $g(x) = \log x$
6. **Level 6:** What is the derivative of f(x)=5/(sqrt(3x-1))?
7. **Level 7:** What is the integral of sin2xcos2x?
8. **Level 8:** Study the convergence of the infinite series starting from n equals 1 to infinity of the sum of the reciprocal of the product of n and n plus 1.
9. **Level 9:** Compute the Fourier transform of f (t) = cos0t.
10. **Level 10:** The following table shows information regarding the sales of daily press in the year 1998, as the number of daily copies sold per thousand inhabitants in 8 autonomous Spanish regions. The sales are assumed to be related to economic activity levels as measured by the Gross Domestic Product (GDP) per inhabitant

in thousands of euros (Source: INE. Anuario Estadistico).

GDP 8.3 9.7 10.7 11.7 12.4 15.4 16.3 17.2
Copies sold 57.4 106.8 104.4 131.9 144.6 146.4 177.4 186.9

Use least squares to estimate a simple regression model that explains the number of copies sold as a function of GDP per capita.

- General Instructions

11. **Level 11**: Find the area bounded by the curve y = x^2 + x + 4, the x-axis and the ordinates x = 1 and x = 3.
12. **Level 12**: A soccer player kicks a ball at an angle of 30 degrees to the horizontal. The initial speed of the ball is 20 meters per second. Assuming no air resistance and that the ball is kicked from ground level, predict how far the ball will travel horizontally before hitting the ground. Use the acceleration due to gravity as 9.8m/s^2.
13. **Level 13**: Imagine I have a graph (with nodes s, s1, s2, s3, s4, s5, t), whose connections (with weights) are represented with the following adjacency matrix:

    ((0,1,3,0,0,0,0), (1,0,0,8,4,0,0), (3,0,0,0,3,0,0), (0,8,0,0,0,1,7), (0,4,3,0,0,1,0), (0,0,0,1,1,0,10),(0,0,0,7,0,10,0)). Obtain the shortest path from s to t.
14. **Level 14**: Consider the following sequence of numbers: 3, 8, 15, 24, 35, 48, ... Your task is to predict the next number in this sequence.
15. **Level 15**: If you invested $5,000 for 10 years at 9% compounded quarterly, how much money would you have? What is the interest earned during the term?
16. **Level 16**: Decode this phrase "SERR CVMMN VA GUR PNSRGREVN" knowing that each letter was shifted by 13.
17. **Level 17**: A freshly baked pie is taken out of the oven and left to cool in a room. The temperature of the oven was 200°C, and the room temperature was a constant 25°C. When the pie is first taken out, its temperature is 180°C. After 20 minutes, the temperature of the pie drops to 100°C. Calculate the time it takes for the pie to cool down to 50°C.
18. **Level 18**: A person compresses a spring a distance of 5cm, which requires a force of 100N. How much work does the person do?
19. **Level 19**: The John Equipment Company estimates its carrying cost at 15% and its ordering cost at $9 per order. The estimated annual requirement is 48,000 units at a price of $4 per unit. What is the most economical number of units to order? How many orders should be placed in a year? How often should an order be placed?
20. **Level 20**: Titan, the largest moon of Saturn, has a mean orbital radius of 1.22x10^9 m. The orbital period of Titan is 15.95 days. Hyperion, another moon of Saturn, orbits at a mean radius of 1.48x10^9 m. Predict the orbital period of Hyperion in days.

B. *Zero-Shot-CoT Prompting*

In the Zero-shot-CoT (Chain of Thought) approach, we utilized the same set of exercises as listed in the previous section. However, to facilitate the generation of step-by-step reasoning, we appended the sentence "Let's think step by step" to the end of each problem statement. This modification aims to prompt the model into providing a more detailed, stepwise breakdown of its thought process in solving the problems. Below are some examples to illustrate how this approach was implemented:

- Specific Instructions

1. **Level 1**: Given the matrix A = [[1, -2, 1], [-2, 3, 1], [5, -7, -3]], find the inverse (Aˆ-1). Let's think step by step.
2. **Level 2:** Generate several normally distributed variables with mean = 45.6 and variance =13.84. Let's think step by step.
3. **Level 3:** Plot the function -3*(x-2)ˆ2-5. Let's think step by step.
4. **Level 4:** Find the minimum of the function f(x)=(xˆ2-x-2)/(xˆ2-6x+9). Let's think step by step.
5. **Level 5:** Find the intersection points of the functions $f(x) = |x - 5|$ and $g(x) = \log x$. Let's think step by step.
6. **Level 6:** What is the derivative of f(x)=5/(sqrt(3x-1))? Let's think step by step.
7. **Level 7:** What is the integral of sin2xcos2x? Let's think step by step.
8. **Level 8:** Study the convergence of the infinite series starting from n equals 1 to infinity of the sum of the reciprocal of the product of n and n plus 1. Let's think step by step.
9. **Level 9:** Compute the Fourier transform of f(t) = cosw0t. Let's think step by step.
10. **Level 10:** The following table shows information regarding the sales of daily press in the year 1998, as the number of daily copies sold per thousand inhabitants in 8 autonomous Spanish regions. The sales are assumed to be related to economic activity levels as measured by the Gross Domestic Product (GDP) per inhabitant in thousands of euros (Source: INE. Anuario Estadistico).

    GDP 8.3 9.7 10.7 11.7 12.4 15.4 16.3 17.2

    Copies sold 57.4 106.8 104.4 131.9 144.6 146.4 177.4 186.9

    Use least squares to estimate a simple regression model that explains the number of copies sold as a function of GDP per capita. Let's think step by step.

- General Instructions

11. **Level 11**: Find the area bounded by the curve y = xˆ2 + x + 4, the x-axis and the ordinates x = 1 and x = 3. Let's think step by step.
12. **Level 12**: A soccer player kicks a ball at an angle of 30 degrees to the horizontal. The initial speed of the ball is 20 meters per second. Assuming no air resistance and that the ball is kicked from ground level, predict how far the ball will travel horizontally before hitting the ground. Use the acceleration due to gravity as 9.8m/sˆ2. Let's think step by step.
13. **Level 13**: Imagine I have a graph (with nodes s, s1, s2, s3, s4, s5, t), whose connections (with weights) are represented with the following adjacency matrix:

    ((0,1,3,0,0,0,0), (1,0,0,8,4,0,0), (3,0,0,0,3,0,0), (0,8,0,0,0,1,7), (0,4,3,0,0,1,0), (0,0,0,1,1,0,10),(0,0,0,7,0,10,0)). Obtain the shortest path from s to t. Let's think step by step.
14. **Level 14**: Consider the following sequence of numbers: 3, 8, 15, 24, 35, 48, ...Your task is to predict the next number in this sequence. Let's think step by step.
15. **Level 15**: If you invested $5,000 for 10 years at 9%compounded quarterly, how much money would you have? What is the interest earned during the term?Let's think step by step.
16. **Level 16**: Decode this phrase "SERR CVMMN VA GUR PNSRGREVN" knowing that each letter was shifted by 13. Let's think step by step.
17. **Level 17**: A freshly baked pie is taken out of the oven and left to cool in a room. The temperature of the oven was 200°C, and the room temperature is a constant 25°C. When the pie is first taken out, its temperature is 180°C. After 20 minutes, the temperature of the pie drops to 100°C. Calculate the time it takes for the pie to cool down to 50°C. Let's think step by step.

18. **Level 18**: A person compresses a spring a distance of 5cm, which requires a force of 100N. How much work does the person do? Let's think step by step.
19. **Level 19**: The John Equipment Company estimates its carrying cost at 15% and its ordering cost at $9 per order. The estimated annual requirement is 48,000 units at a price of $4 per unit. What is the most economical number of units to order? How many orders should be placed in a year? How often should an order be placed?. Let's think step by step.
20. **Level 20**: Titan, the largest moon of Saturn, has a mean orbital radius of 1.22x10ˆ9m. The orbital period of Titan is 15.95 days. Hyperion, another moon of Saturn, orbits at a mean radius of 1.48x10ˆ9m. Predict the orbital period of Hyperion in days. Let's think step by step.

Acknowledgments

The authors would like to acknowledge the support of project 6G-INTEGRATION-3 (grant no. TSI-063000-2021-127), funded by UNICO-5G program (under the Next Generation EU umbrella funds) and ITACA (PDC2022-133888-I00) funded by the Spanish Agencia Estatal de Investigacion (AEI).

References

[1] E. K. Jermakowicz, "The coming transformative impact of large language models and artificial intelligence on global business and education," *Journal of Global Awareness*, vol. 4, no. 2, pp. 1–22, 2023.

[2] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[3] R. C. Staudemeyer, E. R. Morris, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks," 2019. [Online]. Available: https://arxiv.org/abs/1909.09586.

[4] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, "Large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.06196.

[5] K. Jing, J. Xu, "A survey on neural network language models," 2019. [Online]. Available: https://arxiv.org/abs/1906.03591.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762.

[7] OpenAI, "Chatgpt," 2022. [Online]. Available: https://openai.com/blog/chatgpt, Accessed: 23/03/2024.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805.

[9] 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," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[10] J. Zhou, P. Ke, X. Qiu, M. Huang, J. Zhang, "Chatgpt: potential, prospects, and limitations," *Frontiers of Information Technology & Electronic Engineering*, vol. 25, no. 1, pp. 6–11, 2024, doi: 10.1631/FITEE.2300089.

[11] OpenAI, J. Achiam, et al., "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774.

[12] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," 2023. [Online]. Available: https://arxiv.org/abs/2302.04023.

[13] V. Parra, P. Sureda, A. Corica, S. Schiaffino, D. Godoy, "Can generative ai solve geometry problems? strengths and weaknesses of llms for geometric reasoning in spanish," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 65–74, 2024, doi: 10.9781/ijimai.2024.02.009.

[14] N. R. Téllez, P. R. Villela, R. B. Bautista, "Evaluating chatgpt-generated linear algebra formative assessments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 75–82, 2024, doi: 10.9781/ijimai.2024.02.004.

[15] J. A. Hernández, J. Conde, B. Querol, G. Martínez, P. Reviriego, *ChatGPT Tus primeros prompts con 100 ejemplos*. Amazon Kindle Direct Publishing, December 2023. Internet de Nueva Generación.

[16] W. Holmes, M. Bialik, C. Fadel, *Artificial Intelligence in Education. Promise and Implications for Teaching and Learning*. Center for Curriculum Redesign, Mar. 2019.

[17] M. Alier, F.-J. García-Peñalvo, J. D. Camba, "Generative artificial intelligence in education: From deceptive to disruptive," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 5–14, 2024, doi: 10.9781/ijimai.2024.02.011.

[18] J. Izquierdo-Domenech, J. Linares-Pellicer, I. Ferri- Molla, "Virtual reality and language models, a new frontier in learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 46–54, 2024, doi: 10.9781/ijimai.2024.02.007.

[19] OpenAI, "Winding down the chatgpt plugins beta," 2024. [Online]. Available: https://help.openai.com/en/articles/8988022- winding-down-the-chatgpt-plugins-beta, Accessed: 2024-07-24.

[20] OpenAI, "Introducing gpts," 2023. [Online]. Available: https://openai.com/index/introducing- gpts/, Accessed: 2024-07-24.

[21] Wolfram, "Wolfram plugin for chatgpt," 2024. [Online]. Available: https://www.wolfram.com/wolfram-plugin-chatgpt/, Accessed: 2024-07-24.

[22] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, A. Chevalier, J. Berner, "Mathematical capabilities of chatgpt," 01 2023. [Online]. Available: https://arxiv.org/abs/2301.13867.

[23] A. Cherian, K.-C. Peng, S. Lohit, J. Matthiesen, K. Smith, J. B. Tenenbaum, "Evaluating large vision-and-language models on children's mathematical olympiads," 2024. [Online]. Available: https://arxiv.org/abs/2406.15736.

[24] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, W. Yin, "Large language models for mathematical reasoning: Progresses and challenges," 2024. [Online]. Available: https://arxiv.org/abs/2402.00157.

[25] S. Imani, L. Du, H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," 2023. [Online]. Available: https://arxiv.org/abs/2303.05398.

[26] A. Zhou, K. Wang, Z. Lu, W. Shi, S. Luo, Z. Qin, S. Lu, A. Jia, L. Song, M. Zhan, H. Li, "Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification," 2023. [Online]. Available: https://arxiv.org/abs/2308.07921.

[27] E. Davis, S. Aaronson, "Testing gpt-4 with wolfram alpha and code interpreter plug-ins on math and science problems," 2023. [Online]. Available: https://arxiv.org/abs/2308.05713.

[28] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[29] Y. Zhang, H. Fei, D. Li, P. Li, "PromptGen: Automatically generate prompts using generative models," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States, July 2022, pp. 30–37, Association for Computational Linguistics.

[30] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," 2023. [Online]. Available: https://arxiv.org/abs/2302.11382.

[31] Z. Luo, Q. Xie, S. Ananiadou, "Chatgpt as a factual inconsistency evaluator for text summarization," 2023. [Online]. Available: https://arxiv.org/abs/2303.15621.

[32] H. Larochelle, D. Erhan, Y. Bengio, "Zero-data learning of new tasks.," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, 01 2008, pp. 646–651, AAAI Press.

[33] 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. [Online]. Available: https://arxiv.org/abs/2201.11903.

[34] A. Jojic, Z. Wang, N. Jojic, "Gpt is becoming a turing machine: Here are some ways to program it," 2023. [Online]. Available: https://arxiv.org/abs/2303.14310.

[35] G. Team, R. Anil, et al., "Gemini: A family of highly capable multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2312.11805.

[36] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971.

Alejandro L. García-Navarro

Alejandro L. García-Navarro is currently pursuing a Bachelor's degree in Data Science and Engineering at Universidad Carlos III de Madrid, Spain. During the 2023-2024 academic year, he participated in an exchange program at Concordia University in Montréal, Canada, where he deepened his expertise in Data Analytics and Machine Learning. He currently serves as a Research Assistant in the Telematic Engineering Department at Universidad Carlos III de Madrid, contributing to projects and publications involving Machine Learning, Generative AI, and Large Language Models.

Nataliia Koneva

Nataliia Koneva completed her M.Sc. degrees in Bioengineering Systems and Technologies at Moscow Power Engineering Institute in Moscow, Russia, in 2019, and then pursued a specialization in Connected Industry 4.0 at Universidad Carlos III de Madrid, Spain, in 2022. She is currently continuing her Ph.D. studies at the same institution, focusing on the Applications of AI/ML techniques for intelligent network optimization.

José Alberto Hernández

José Alberto Hernández completed the five-year degree in Telecommunications Engineering at Universidad Carlos III de Madrid (Madrid, Spain) in 2002, and the Ph.D. degree in Computer Science at Loughborough University (Leics, United Kingdom) in 2005. Between 2005 and 2009, he was a postdoctoral researcher and teaching assistant at Universidad Autónoma de Madrid, where he participated in several both National and European research projects concerning the modeling and performance evaluation of communication networks, optical WDM networks and energy efficiency in computer communications. At present, José Alberto Hernández is a Senior Lecturer at Universidad Carlos III de Madrid, he has published more than 200 scientific articles in well-known journals and conference proceedings, including IEEE Network, IEEE Communications Magazine, IEEE J. Selected Areas in Communications, IEEE Internet Computing, etc. He has also participated in many National and European research projects related to his main research interests, namely Computer Networks and their performance evaluation along with Generative AI and Machine Learning.

Alfonso Sánchez-Macián

Alfonso Sánchez-Macián received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 2000 and 2007, respectively. He has worked as a Lecturer and a Researcher with several universities, such as the Universidad Politécnica de Madrid; the IT Innovation Centre, , University of Southampton, Southampton, U.K.; the Universidad Antonio de Nebrija, Madrid, and the Universidad Carlos III de Madrid, where he currently works. His current research interests include nonfunctional properties of the systems, including security, fault tolerance, and reliability, as well as Generative AI and its application to different environments.