LLMs Take on the Bebras Challenge: How Do Machines Compare to Students?

Germán Capdehourat, María Eugenia Curi and Víctor Koleszar
Ceibal, Uruguay

Keywords: Artificial Intelligence, Computational Thinking, Education, K-12, LLMs.

Abstract: Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains. However, their performance in tasks involving logical reasoning and computational thinking continues to be an active area of research. This study analyzes the behaviour of state-of-the-art LLMs on tasks from Bebras Challenge, a test designed to promote computational thinking skills. We compare the outcomes of LLMs and primary and secondary school students from grades 3rd through 9th in Uruguay, who participated in the Bebras Challenge as part of the country's Computational Thinking and Artificial Intelligence program. The results reveal that LLMs achieve an increasing performance as the model complexity increases, with the most advanced ones outperforming the average younger students' results. Our findings highlight both the promise and the current limitations of LLMs in tackling computational thinking challenges, providing valuable insights for their integration into educational contexts. In particular, the results suggest that LLMs could be used as a complementary tool to analyse the task's difficulty level, which could be very helpful to accelerate the time-consuming exchange and discussion process actually required to categorize the tasks.

1 INTRODUCTION

The rapid development of large language models (LLMs) has revolutionized various domains, showcasing impressive potential capabilities in natural language understanding, reasoning, and task completion. LLMs such as GPT-4, Gemini and Llama have demonstrated proficiency across a wide range of applications, including content generation, coding, problem-solving, and conversational agents (Fan, 2024). Their versatility has extended to educational contexts, where they are increasingly used as tools to support learning, tutoring, and the development of critical skills such as reading comprehension and problem-solving (Li, 2024).

To evaluate and benchmark their performance, researchers commonly use standardized tests such as MMLU, MATH and IFEVAL. These benchmarks help identify strengths and weaknesses, with a consistent finding being that LLMs often excel in tasks requiring factual recall and pattern recognition but face challenges with tasks that demand logical reasoning and multi-step problem-solving. Studies in this area (Wan, 2024) highlight limitations in their ability to perform tasks grounded in logical consistency or requiring abstract reasoning, indicating a gap that merits further exploration.

Computational thinking (CT) is considered a critical competency in today's education landscape, referred to problem-solving based on computer science principles and concepts (Barr & Stephenson, 2011, Shute et al., 2017). CT emphasizes skills such as algorithmic thinking, decomposition, generalization, abstraction, and evaluation (Grover & Pea, 2013). These skills mirror the logical and structured reasoning often required in tasks where LLMs struggle. Different tests used to evaluate CT, such as the Bebras Challenge (Dagiene, 2016), provide a unique lens to analyze the reasoning and problem-solving capabilities of LLMs.

Exploring the performance of LLMs on Bebras tasks is not only interesting from an academic perspective but also practical. Beyond solving tasks themselves, LLMs could contribute to educational settings by assisting educators in identifying the appropriate age range for specific challenges, considering different steps in the resolution process, or automating the thematic classification of tasks. These applications could reduce the time and effort required to design and evaluate activities, enriching the learning process.

Capdehourat, G., Curi, M. E. and Koleszar, V.
LLMs Take on the Bebras Challenge: How Do Machines Compare to Students?.
DOI: 10.5220/0013364100003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 338-346
ISBN: 978-989-758-746-7; ISSN: 2184-5026

In this study, we examine the performance of state-of-the-art (SoTA) LLMs on Bebras tasks, comparing their outcomes with those of students in grades 3rd to 9th in Uruguay. This country has been a pioneer in incorporating technology into the educational system, starting in 2007 with a one-to-one computer program called Ceibal (Ceibal, 2025). In this context, the teaching of computational thinking was incorporated a few years ago through a specific educational program and a curriculum designed for such purposes. The key questions addressed in this study are: *What is the accuracy rate of SoTA LLMs when solving Bebras tasks?* and *How does the performance of those LLMs compare to that of Uruguayan students on the same tasks?*

2 COMPUTATIONAL THINKING EDUCATION IN URUGUAY

Ceibal is the innovative one-to-one educational program initiative that positioned Uruguay as the first country in the world to provide laptops and Internet access to all students and teachers in public K-12 schools. In this context, Ceibal has incorporated computer science education into classrooms through the Computational Thinking and Artificial Intelligence (PCIA, in Spanish) program, a joint initiative with the National Administration of Public Education (ANEP, in Spanish). This program operates on an optional basis but it is integrated into the regular school schedule, with a remote teacher working collaboratively with the classroom teacher (Koleszar et al., 2021a). By 2024, the program coverage reaches approximately 75% of public schools across the country. The main goal is to help students develop foundational computer science concepts starting in primary education, learn different approaches to problem-solving, and express solutions through programming.

Bebras Challenge is an international initiative to promote participation in CT activities. It originated at Vilnius University, with its first edition held in Lithuania in 2004 (Dagienė, 2010). Since then, participation has grown steadily, with nearly 4,000,000 participants from over 70 countries in 2023. Numerous studies highlight the challenging work of producing high-quality tasks. Each year, representatives from all participating countries develop, approve, and validate a shared pool of tasks from which each country selects their challenges. This systematic process involves academics and educators from all member countries and consists of multiple stages. Finally, the revised tasks are then presented to the Bebras community during the annual workshop, where representatives from all countries review, improve, and select a final pool of tasks that can be used in the annual challenge.

Since 2020, Ceibal PCIA has been part of Bebras, organizing the challenge for the schools in Uruguay. After a one-month preparation phase, during which students are provided with resources and taught strategies for problem-solving, the annual challenge is made available on a learning assessment online platform, where students complete it individually. These activities contribute to building and enriching the teaching community while providing valuable data for research and evaluation in computational thinking skills (Stupurienė et al., 2016). For this study, we examined a set of tasks from the 2023 Bebras edition implemented in Uruguay for students in grades 3rd through 9th. This edition featured 26 tasks, distributed across three age categories, designed to evaluate the following computational thinking skills: algorithmic thinking, generalization and evaluation.

3 RELATED WORK

Performance evaluation of LLMs is a rapidly evolving field, with various benchmarks designed to assess different aspects of their capabilities. Among the most widely used are tasks like MMLU (Hendrycks et al., 2021), which tests multi-task language understanding, MATH (Hendrycks et al., 2021), which focuses on mathematical reasoning, and IFEVAL (Zhou et al., 2023), designed for instruction-following abilities. These benchmarks provide a foundation for understanding the strengths and weaknesses of LLMs across different domains. However, a growing interest lies in assessing models using tests that are more specialized. In our case, we focused on computational thinking challenges, which usually include a greater component of logic and reasoning ability, thus introducing greater complexity to the models (Williams et al., 2024). This kind of analysis allows to gauge the LLMs problem-solving and reasoning abilities in scenarios aligned with human cognitive processes.

As previously mentioned we are particularly interested in the computational thinking problems from the Bebras challenge. A previous work that studies LLMs performance on solving these tasks, considers a legacy GPT-3 model (Bellettini et al., 2023). This work investigates the ability of OpenAI's DaVinci model to solve tasks from the Bebras

challenge, posing research questions such as *How often is the model able to answer correctly?* or *Does the model perform better with some specific types of tasks?* Although the study provides valuable insights, its conclusions are limited by the use of an earlier LLM, with significantly lower performance compared to current SoTA models. Moreover, it does not compare the performance to that of real students, leaving an essential gap in understanding how LLMs fare against real-world benchmarks.

Another more recent study (Pădurean et al., 2024), uses a similar approach, but in this case focused on different visual programming and computational thinking tests, such as HoC, ACE, and CT-test. These benchmarks focus on typical block coding programming and computational thinking problems, sometimes involving multimodal inputs, such as textual descriptions and accompanying visuals. The authors examine the performance of advanced LLMs, including multimodal variants, and find that SoTA models like GPT-4o and Llama3 barely match the performance of an average school student. Although these tests differ from Bebras, they provide valuable context for understanding the limitations of LLMs in computational thinking evaluation.

An additional area of interest involves the automated classification of tasks. For example, Lucy et al. (2024) investigates the categorization of mathematical problems. The study highlights the potential utility of automating task tagging to streamline the preparation of educational materials. However, the findings suggest that LLMs often struggle to accurately tag problems according to predefined standards, typically predicting labels that approximate but subtly differ from the ground truth. For Bebras challenges, task categorization involves several dimensions, including difficulty levels, age ranges, and specific computational thinking skills. The framework proposed by Dagienė et al. (2017) introduces a dual-level categorization system for Bebras tasks. The first level relates to computational thinking skills (e.g. abstraction, decomposition, generalization), while the second addresses informatics concepts such as algorithms, data structures and representations, computer processes, and human-computer interaction. To investigate whether LLMs can automate this classification process effectively is a promising avenue for research and could significantly influence how such challenges are organized and utilized.

Our study contributes to this field in several ways. To begin with, up to our knowledge this is the first work to analyze the performance of modern LLMs, including state-of-the-art models, on the Bebras challenge. Secondly, unlike previous works, we directly compare LLM performance with that of students participating in the Bebras challenge in Uruguay. This comparison provides a clearer understanding of how LLMs align with human performance on such tasks. Finally, we explore the potential for LLMs to assist in automating educational processes, such as task categorization and age-range recommendations. The results show that while LLMs demonstrate a strong ability to solve many tasks, their performance varies by task type. While it could be incorporated as an objective measure of the level of difficulty of challenges, its ability to classify challenges according to skills or knowledge domains is still insufficient for practical applications. The study provides valuable insights into the integration of LLMs in educational contexts and contributes to the broader understanding of their strengths and limitations.

4 EXPERIMENTAL ANALYSIS WITH STUDENTS AND LLMS

Our experimental section is divided into two main parts. On the one hand, we present the results of the Bebras challenge 2023 in Uruguay. A set of different tasks were selected in order to analyze the students' performance for different grade levels. On the other hand, we analyze the capacity of various LLMs to solve the Bebras tasks. To do so, we carried out a preprocessing step to get a text-only version for the tasks that include images. Finally, the section concludes with a comparative discussion of the results obtained for students and language models.

Within the Ceibal PCIA program, Bebras is an optional initiative for schools every year since 2020 (Koleszar et al., 2021b; Porto et al., 2024). Different tasks are selected to suit each educational level, distributed in four categories: grades 1st-2nd, 3rd-4th, and 5th-6th of primary education, and 7th-9th of secondary education. For this analysis, the results for students from 1st and 2nd grade were discarded, because the original Bebras tasks were modified in those cases to facilitate reading, making the text simpler and more age appropriate. The final subset of student results analyzed is detailed in Table 1, while the corresponding subset of 22 Bebras tasks was distributed as shown in Figure 1. It should be noted that some of them are repeated between the different categories.

Table 1. Dataset considered from the Bebras 2023 edition in Uruguay.

Grade	Category name	Number of students
3th	Benteveos	3,717
4th	Benteveos	17,972
5th	Cardenales	17,789
6th	Cardenales	19,770
7th	Horneros	1,488
8th	Horneros	1,058
9th	Horneros	741

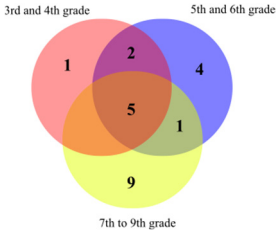

Figure 1. Distribution of the Bebras tasks carried out in Uruguay in 2023.

4.1 Students' Results

Benteveos. The average percentage of correct responses was 45.5% for 3rd-grade students and 53.4% for 4th-grade students. Figure 2 presents the detailed performance results broken down by task. To identify the tasks we simplified the international Bebras code, removing the year and using only the country and the assigned number. For example 2023-LT-01 is represented as LT-01 (as all tasks are from 2023).

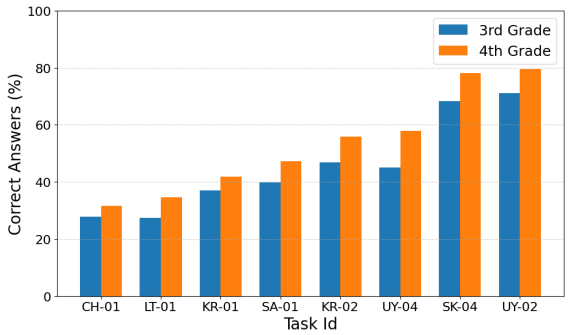

Figure 2. Average percentage of correct responses by task for 3rd- and 4th-grade students.

The first thing that can be observed is the great variability in performance per task, from cases just above 25%, to others that almost reach 80% of correct answers. Additionally, the results show that 4th-grade students systematically perform better than 3rd-grade students, with a significant difference ranging between 4% and 13%, with an average of 8%. This difference is probably explained by the fact that, in addition to the age difference, the Ceibal's PCIA educational program starts at schools during the 4th-grade year. From a more detailed observation of the tasks, LT-01 and CH-01 were particularly complex for all participants, as the average of correct responses was below 35%. On the other hand, SK-04 and UY-02 were more simple as both were answered correctly with an average correct response rate above 65%.

Cardenales. Figure 3 presents the results for the students 5th-6th grade. In this case again the results of the older students are higher, with an average percentage of correct responses of 47.3% for 5th-grade and 50.1% for 6th-grade. However, the differences in this case are smaller than those observed between the 3rd- and 4th-grade, ranging from 0% to 5.6%, with an average of 2.7%. The variability among tasks is similar to the previous case, with a more noticeable break in this case between two groups of tasks, those that are above and below 50% of correct answers. If we consider the tasks analyzed for the previous category, CH-01 and LT-01 achieved a higher average of correct responses compared to the 3rd- and 4th-grade students. Additionally, the task UY-02, which already had a good performance of the students from the Benteveos category, also showed better results for the students in this category.

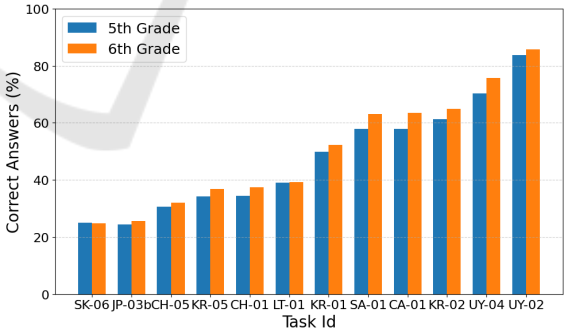

Figure 3. Average percentage of correct responses by task for 5th- and 6th-grade students.

Horneros. The number of tasks used in this category is larger than in the previous ones, as it can be seen in Figure 4. The average percentage of correct responses in this case was 50.0% for the 7th-grade students, 50.6% for 8th-grade students, and 55.3% for 9th-grade students. Thus, an improvement in performance is again observed, as the age of the students increases.

However, the gap between 7th and 8th grade is quite short in this case, with even several tasks in which the results are better for 7th grade students. The differences between 8th and 9th grade are larger, where the results for each task are again systematically better for older students. The performance variability among tasks falls within an even broader range in this case, between 25% and 92%. Upon a deeper analysis of the tasks, CA-01, UY-02, and SA-01, students achieve higher average results compared to the previous category. Tasks BR-04 and PH-03 have the lowest average response rates compared to all other tasks, which could indicate that these are more complex tasks.

Although the set of tasks used for each category is different, it can be seen that the insights found are quite consistent for the three analyzed categories. In all cases, the students' performance increases on average, as the age of the students increases. Furthermore, a great variability in performance is observed for each of the tasks, which shows that the set selected for the challenge covers a fairly wide range of difficulties.

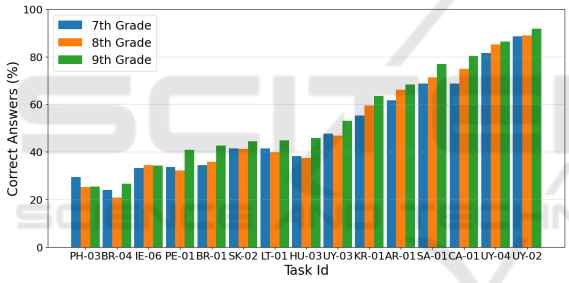

Figure 4. Average percentage of correct responses by task for 7th-, 8th- and 9th-grade students.

4.2 LLMs Performance on Bebras

As previously mentioned, in order to perform the tests with several LLMs, a pre-processing step was applied to prepare the tasks. To do this, a description of the images or graphic elements that provide key information for the formulation and resolution of the problem was generated, with the aim of presenting the task as faithfully as possible to how a student receives it. To generate these text descriptions, the OpenAI GPT-4o model was used and the generated outputs were manually corrected. In this way, the pre-processed tasks used as input for the language models includes both the original task text as well as the corresponding text description of the images.

Various LLMs were selected to run the experiments, some proprietary models and others with open weights. For the proprietary case, OpenAI models were used. In this case, the selected models were GPT-4 and GPT-4o, and the more recent "reasoning" models o1-mini and o1-preview. For the open weights model selection, we took into account the most recent ones available in Ollama and the local hardware infrastructure that we have for running the tests. Thus, the open models selected were of lower capacity than those of Open-AI, such as gemma2:2B and gemma2:9B from Google, llama3.2:3B and llama3.1:8B from Meta, phi3.5:3.8B from Microsoft and qwen2.5:7B from Alibaba. The last number of each open model corresponds to the number of billions of parameters. Thus, with the selected LLMs we manage to cover a wide range of models with different capacity levels.

In addition to the different LLMs, also different prompt variations were analyzed. In this case, two slightly different prompts were considered. The general structure in both cases was the same, including the task description, the question to answer, the available multiple-choice options and a final instruction asking the model to solve the task, indicating the correct answer. The analyzed variation corresponds to a chain-of-thought (CoT) approach, where the phrase "Let's think step by step the problem to reach the correct answer" is also included in the prompt.

In order to analyze the general performance of all the different LLMs and prompts considered, a first experiment was conducted, using all the Bebras tasks used in the 2023 edition in Uruguay. In Figure 5 the results are presented, where the first thing to notice is that the performance is directly related with the model capabilities. The smaller models, which have a few billion parameters, struggle to solve the tasks, with only a few correct responses. Then we have the middle range open models tested, which reach a performance slightly above 30%. The better results in this case correspond to the OpenAI models, with the best results above 50% for the standard models, and

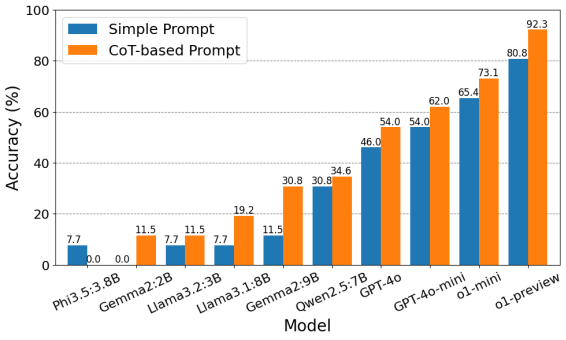

Figure 5. Average accuracy results on the selected Bebras tasks for the different LLMs evaluated.

much better results for the most advanced "reasoning" models. Finally, it is worth highlighting that almost all models present a better performance with the CoT-based prompt.

From now on we concentrate on the OpenAI models, as they were the ones with better results in the previous experiments. The first thing we analyzed is the consistency of the previous results. As we know, the output of an LLM is not always the same, so we repeated the previous experiment 10 times for each task. This way, we computed the number of correct answers for each task on each of the experiment runs. The results are shown in Figure 6, where the histograms indicate that GPT-4o presents more consistent results than GPT-4o-mini. As we can see, the histograms in this case are more concentrated on 1 or 0 values, which indicate that responses for a certain task are always correct or wrong. It is worth noting that little impact of the prompt is observed in this case.

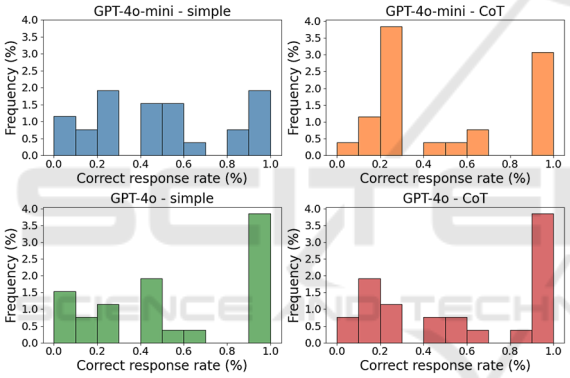

Figure 6. Consistency analysis for the Open AI standard models GPT-4o and GPT-4o-mini.

4.3 Results Comparison and Further Discussion

After looking at the results separately for students and LLMs, we analyze and compare the performance in both cases. Since each test varies according to the students' grade level, a comparison can be made between the LLMs and the grades that share the same tasks in the Bebras challenge. Figures 7, 8 and 9 presents the comparative results for the different categories, analyzing the performance of the LLM models on the tasks corresponding to the challenge for these grades.

Looking at the different graphs, the first thing to notice is that the LLMs results cover the whole range of students' performance. That is to say, that the less capable models have a similar behaviour to the worse students, while the opposite happens with the most advanced models. This result is more than relevant, since it indicates that it would be possible to somehow automate the calibration of the tasks, based on the performance that the different LLMs have when trying to solve it.

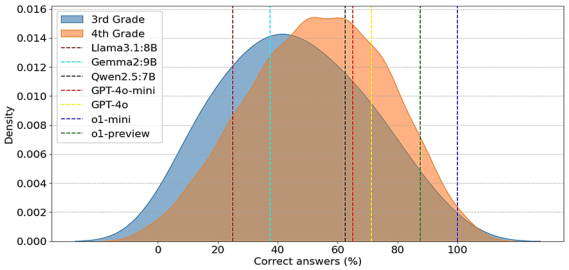

Figure 7. Comparative results between LLMs vs 3rd- and 4th-grade students.

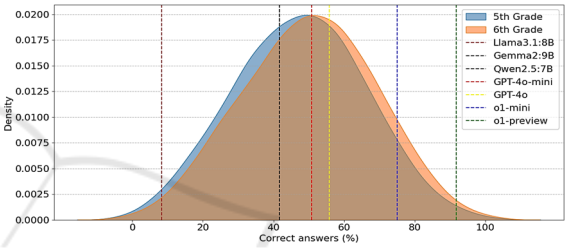

Figure 8. Comparative results between LLMs vs 5th- and 6th-grade students.

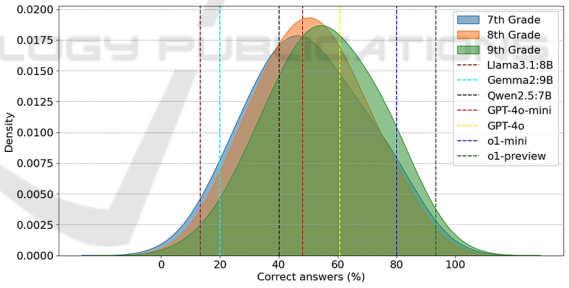

Figure 9. Comparative results between LLMs vs 7th-, 8th- and 9th-grade students.

Furthermore, if we compare the results among the three different graphs, it can be seen that the LLMs performances present a noticeable drift towards worse results (i.e. vertical lines move to the left), as the age level of the categories increases. This makes a lot of sense, since the set of tasks selected for each category is usually associated with the corresponding ages, with the level of difficulty increasing as the ages get older. Thus, the performance degradation of the LLMs is probably explained by the increasing difficulty of the set of tasks selected for each category.

The above results prove that it would be possible to integrate LLMs into the review and categorization processes of the Bebras challenge, as an objective tool to measure the difficulty level of each task. Although this incorporation requires further studies and work on calibration and analysis of the models, the automation of these tasks would be very beneficial and could save significant efforts in the time-consuming exchange and discussion process actually required to categorize the Bebras tasks.

A different automation that could be helpful, concerning LLMs, is the task classification according to the different skills and knowledge required. The aforementioned framework (Dagienė et al., 2017) was used in this case, to test the capabilities of some LLMs concerning the classification task. According to the best results obtained in our preliminary tests, the model struggles to classify correctly in both cases. Concerning the skill associated with each task, the best result for this case was 50% of the tasks classified correctly. The result is not much better for the case of the knowledge domain of each task, where the best model reached 57.7% of correct classifications.

5 CONCLUSIONS AND FUTURE WORK

This study highlights the potential and limitations of state-of-the-art LLMs in solving CT problems such as those presented in the Bebras Challenge. While LLMs have shown strong performance in solving specific types of tasks, their variability across task categories indicates opportunities for improvement. These findings underscore the importance of developing targeted methodologies for integrating LLMs into educational processes.

The results demonstrate the significant impact of prompt design on LLM performance. Incorporating chain-of-thought reasoning into the prompts led to noticeable improvements in accuracy, highlighting the importance of carefully crafting input instructions to align with the cognitive requirements of the tasks. Furthermore, the performance of LLMs consistently improved with increasing model size, a trend that parallels the progression observed in students' results as they advance in age and grade level. This alignment between model size and student performance provides a natural hierarchy of difficulty levels, where tasks can be ranked according to their complexity and solved progressively by students or models with corresponding capabilities.

Advanced reasoning models achieved near-perfect scores on the Bebras tasks, showcasing their ability to handle complex computational thinking challenges. However, even these models exhibited limitations when tasked with categorizing exercises based on the skills required or the knowledge domains involved. This indicates that, while LLMs are effective problem solvers, their meta-cognitive abilities to analyze and classify tasks remain underdeveloped, presenting an avenue for further research and enhancement.

One promising area of future work involves incorporating LLMs into the process of challenge generation and evaluation. By leveraging their ability to solve tasks and analyze patterns of performance, LLMs could provide objective measures of task difficulty. This capability could streamline the current time-intensive process of categorizing challenges by difficulty and assigning appropriate age ranges. Tools based on LLM performance metrics could serve as valuable resources for educators and task designers, enabling a more efficient and data-driven approach to preparing computational thinking activities.

Another avenue for exploration is improving LLMs' performance in automatic task classification. While LLMs have demonstrated some capability in identifying key skills and knowledge domains associated with tasks, their accuracy remains insufficient for practical applications. Enhancing their ability to classify tasks based on computational thinking skills, such as abstraction or algorithmic reasoning, could significantly benefit the design of targeted educational interventions and the organization of challenge databases.

Finally, the potential for modifying the Bebras Challenge format using LLMs represents an exciting opportunity. For example, instead of relying solely on multiple-choice questions, challenges could include intermediate reasoning steps where LLMs assist students in formulating their solutions. Such modifications would still allow for automated grading but would provide richer insights into students' thought processes and problem-solving strategies. LLMs could also play a role in generating adaptive feedback, helping students improve their computational thinking skills in real-time.

REFERENCES

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: What is involved and what is the role

of the computer science education community?. ACM inroads, 2(1), 48-54.

Bellettini, C., Lodi, M., Lonati, V., Monga, M., & Morpurgo, A. (2023, April). DaVinci goes to Bebras: a study on the problem solving ability of GPT-3. In CSEDU 2023-15th International Conference on Computer Supported Education (Vol. 2, pp. 59-69). SCITEPRESS-Science and Technology Publications.

Casal-Otero, L., Catala, A., Fernandez-Morante, C., Taboada, M., Cebreiro, B. and Barro, S. (2023). AI literacy in K-12: A systematic literature review, in International Journal of STEM Education, 10(1), 29. https://doi.org/10.1186/s40594-023-00418-7.

Ceibal (2025). What is Ceibal? https://ceibal.edu.uy/en/what-is-ceibal/

Ceibal (2022). Pensamiento computacional. Propuestas para el aula. https://bibliotecapais.ceibal.edu.uy/info/pensamiento-computacional-propuesta-para-el-aula-00018977

Dagienė, V. (2010). Sustaining informatics education by contests. In Teaching Fundamentals Concepts of Informatics: 4th International Conference on Informatics in Secondary Schools-Evolution and Perspectives, ISSEP 2010, Zurich, Switzerland, January 13-15, 2010. Proceedings 4 (pp. 1-12). Springer Berlin Heidelberg.

Dagiene, V., & Stupuriene, G. (2016). Bebras--A Sustainable Community Building Model for the Concept Based Learning of Informatics and Computational Thinking. Informatics in education, 15(1), 25-44.

Dagienė, V., Sentance, S., & Stupurienė, G. (2017). Developing a two-dimensional categorization system for educational tasks in informatics. Informatica, 28(1), 23-44.

Grover, S., & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. Educational Researcher, 42(1), 38–43. https://doi.org/10.3102/0013189X12463051

Dan Hendrycks and Collin Burns and Steven Basart and Andy Zou and Mantas Mazeika and Dawn Song and Jacob Steinhardt (2021). Measuring Massive Multitask Language Understanding. Proceedings of the International Conference on Learning Representations (ICLR).

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D.X., & Steinhardt, J. (2021). Measuring Mathematical Problem Solving With the MATH Dataset. ArXiv, abs/2103.03874.

Kim, S., Jang, Y., Kim, W., Choi, S., Jung, H., Kim, S. and Kim, H. (2021). Why and What to Teach: AI Curriculum for Elementary School, in Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15569-15576.

Koleszar, V., Pérez Spagnolo, A., & Pereiro, E. (2021a). Pensamiento computacional en educación primaria: El caso de Uruguay. Jornadas Argentinas de Didáctica de las Ciencias de la Computación, Buenos Aires, Argentina.

Koleszar, V., Clavijo, D., Pereiro, E., & Urruticoechea, A. (2021b). Análisis preliminares de los resultados del desafío BEBRAS 2020 en Uruguay. Revista INFAD de Psicología. International Journal of Developmental and Educational Psychology., 1(2), 17-24.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. (2024). A Bibliometric Review of Large Language Models Research from 2017 to 2023. ACM Trans. Intell. Syst. Technol. 15, 5, Article 91 (October 2024), 25 pages. https://doi.org/10.1145/3664930

Long, D. & Magerko, B. (April 2020). What is AI literacy? Competencies and design considerations, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-16.

Lucy, L., August, T., Wang, R. E., Soldaini, L., Allison, C., & Lo, K. (2024). MathFish: Evaluating Language Model Math Reasoning via Grounding in Educational Curricula. arXiv preprint arXiv:2408.04226.

Natali, V., & Nugraheni, C. E. (2023). Indonesian Bebras Challenge 2021 ExploratoryData Analysis. Olympiads in Informatics, 17, 65-85.

Ng, D. T. K., LEUNG, J. K. L., Chu, S. K. W. and Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review, in Computers and Education: Artificial Intelligence, 2, 100041. 10.1016/j.caeai.2021.100041.

Olari, V. and Romeike, R. (October 2021). Addressing AI and Data Literacy in Teacher Education: A Review of Existing Educational Frameworks, in The 16th Workshop in Primary and Secondary Computing Education, pp. 1-2.

Pădurean, V. A., & Singla, A. (2024). Benchmarking Generative Models on Computational Thinking Tests in Elementary Visual Programming. arXiv preprint arXiv:2406.09891.

Porto, C., Pereiro, E., Curi, M. E., Koleszar, V., & Urruticoechea, A. (2024). Gender perspective in the computational thinking program of Uruguay: teachers' perceptions and results of the Bebras tasks. Journal of Research on Technology in Education, 1-15.

Qingyao Li and Lingyue Fu and Weiming Zhang and Xianyu Chen and Jingwei Yu and Wei Xia and Weinan Zhang and Ruiming Tang and Yong Y. (2024). Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges. arXiv preprint arXiv:2401.08664.

Sentance, S. and Waite, J. (2022). Perspectives on AI and data science education. Recovered from https://www.raspberrypi.org/app/uploads/2022/12/Perspectives-on-AI-and-data-science-education-_Sentance-Waite_2022.pdf

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. Educational Research Review, 22, 142–158. doi:10.1016/j.edurev.2017.09.003

Tedre, M., Denning, P. and Toivonen, T. (November 2021). CT 2.0, in Proceedings of the 21st Koli Calling International Conference on Computing Education Research, pp. 1-8.

Touretzky, D., Gardner-McCune, C., Martin, F. and Seehorn, D. (2019). AI for K-12 Envisioning: What Should Every Child Know about AI?, in Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 9795-9799.

UNESCO (2023). Currículos de IA para la enseñanza preescolar, primaria y secundaria: un mapeo de los currículos de IA aprobados por los gobiernos. Recovered from https://unesdoc.unesco.org/ark:/48223/pf0000380602_spa

Williams, S., & Huckle, J. (2024). Easy Problems That LLMs Get Wrong. arXiv preprint arXiv:2405.19616.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2124–2155, Miami, Florida, USA. Association for Computational Linguistics.

Jeffrey Zhou and Tianjian Lu and Swaroop Mishra and Siddhartha Brahma and Sujoy Basu and Yi Luan and Denny Zhou and Le Hou (2023). Instruction-Following Evaluation for Large Language Models. ArXiv, abs/2311.07911.