Politechnika Częstochowska
Wydział Zarządzania

Zeszyty Naukowe Politechniki Częstochowskiej. Zarządzanie
Research Reviews of Czestochowa University of Technology. Management

No. 63 (2026), pp. 117-131, ISSN: 2083-1560, e-ISSN: 3071-9259
DOI: 10.17512/znpcz.2026.3.10, https://znz.pcz.pl/

Received: 22.06.2026; Revised: 04.08.2026;
Accepted: 06.08.2026; Published: 30.09.2026.

EVALUATING LLM-BASED AI SYSTEMS ON THE POLISH MATHEMATICS MATURA EXAMINATION: IMPLICATIONS FOR AI GOVERNANCE AND QUALITY ASSURANCE IN EDUCATIONAL ORGANIZATIONS

Maciej Renz[1], Grzegorz Borkowski[2*], Ewelina Ząb[3]

[1, 2, 3] Wrocław University of Economics and Business, Faculty of Management, Poland

Abstract: LLM-based AI systems are increasingly used in education, including as tools for solving, explaining, and checking mathematical tasks. This exploratory study evaluates the performance of four widely used user-facing AI systems – ChatGPT, Claude, Gemini, and Perplexity – on the Polish mathematics Matura examination at the basic and extended levels. The systems were tested through their chat interfaces using official CKE examination papers, submitted as complete attachments, and a single fixed prompt for each level. The responses were assessed according to the official CKE marking schemes, with points awarded for the final responses and for the correctly completed stages of open-ended tasks. All systems achieved high scores at the basic level (86-96%), whereas performance at the extended level varied considerably (38-100%). At the basic level, most lost points resulted from errors in interpreting charts, figures or other visual information. At the extended level, errors were more often associated with mathematical reasoning or methods. The findings demonstrate the value of supplementing aggregate examination scores with diagnostic error profiles that distinguish visual-data processing from mathematical reasoning. For educational organizations, such benchmarks may provide one source of evidence for human oversight, risk identification and periodic quality assessment. However, the results reflect the performance of complete AI systems under specific interface conditions and should not be interpreted as direct evidence of their general educational effectiveness. Further studies involving multiple examination papers, repeated trials and educational users are needed.

[1] Maciej Renz, Eng, 199317@student.ue.wroc.pl, https://orcid.org/0009-0005-1494-7013
[2] Grzegorz Borkowski, MSc, grzegorz.borkowski@ue.wroc.pl, https://orcid.org/0009-0002-9189-2747
[3] Ewelina Ząb, MSc, ewelina.zab@ue.wroc.pl, https://orcid.org/0009-0008-5147-486X
* Corresponding author: Grzegorz Borkowski, grzegorz.borkowski@ue.wroc.pl

Keywords: AI governance, error analysis, LLM-based AI systems, mathematical reasoning, mathematics education, quality assurance

JEL Classification: I21, M15, O33, C88

Introduction

Large language models (LLMs) have become one of the most widely discussed technologies of recent years (Bommasani et al., 2021; Zhao et al., 2026). Built on the transformer architecture (Vaswani et al., 2017) and trained on very large text corpora, they showed with GPT-3 that increasing model size and training data leads to strong general language abilities (Brown et al., 2020), in line with the scaling laws observed for such models (Kaplan et al., 2020). Since the public release of ChatGPT, tools of this kind, including ChatGPT, Claude, Gemini, and Perplexity, have quickly become everyday aids for students and teachers (Kasneci et al., 2023).

These models are increasingly used not only to generate text, but also to solve tasks that require step-by-step reasoning (Bubeck et al., 2023; OpenAI et al., 2023). Some abilities appeared only beyond a certain scale, a phenomenon called emergent abilities (Wei et al., 2022b). Mathematics has become a central area for testing them, because every step must be correct for the final answer to hold, which makes mistakes easy to detect (Cobbe et al., 2021; Ahn et al., 2024). Dedicated efforts such as Minerva showed that strong quantitative reasoning is possible, but remains difficult (Lewkowycz et al., 2022; Hendrycks et al., 2021).

The Polish Matura is a useful benchmark for such a test. It is standardized, it comes with official marking schemes published by the Central Examination Board (CKE), and it has already been used to evaluate language models in Polish (Jassem et al., 2025). It is divided into a basic and an extended level, allowing a comparison across two clearly different levels of difficulty. Its tasks often include figures, charts, and drawings, so they test more than pure text-based calculation, an aspect on which models are known to be weaker (Pyeon et al., 2025).

The growing presence of these tools also raises a question of management: educational organizations – schools, universities, and the bodies responsible for the quality of teaching and assessment – increasingly have to decide where, and under what safeguards, such models can be relied upon. An accurate picture of what the models actually get right and wrong is a precondition for these decisions.

The aim of this study is therefore twofold: to examine how well four popular language models solve the Polish mathematics Matura at both levels and to describe the types of mistakes they make, and, on this basis, to draw out the implications for managing the use of such tools in educational organizations. Four research questions were formulated: (1) What score, measured as the percentage of points awarded under the CKE marking scheme, do the models reach at the basic and extended levels? (2) Are there clear differences between the models? (3) What kinds of mistakes dominate – are they mathematical (wrong reasoning) or do they come from other causes, such as misreading the data? (4) What do these results imply for educational organizations deciding how far, and with what safeguards, such tools can be relied upon in teaching and learning?

Literature review

Evaluating the mathematical performance of LLM-based systems

Most contemporary large language models are based on the Transformer architecture (Vaswani et al., 2017) and are pretrained on large-scale text corpora. Increasing model scale has been associated with the emergence of broader capabilities, including improved performance on mathematical reasoning tasks (Brown et al., 2020; Zhao et al., 2026). Mathematical performance is commonly evaluated with text-based benchmarks such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021) and MMLU (Hendrycks et al., 2020); recent surveys document both progress and continuing challenges in mathematical reasoning (Ahn et al., 2024).

Step-by-step reasoning

Prompting models to produce intermediate solution steps improves performance on multi-step tasks, whether through chain-of-thought examples (Wei et al., 2022a) or a short zero-shot instruction (Kojima et al., 2022). Training-based approaches point in the same direction: Minerva achieved high results on quantitative benchmarks (Lewkowycz et al., 2022), and reasoning-oriented models such as OpenAI o1 (OpenAI et al., 2024) and DeepSeek-R1 (Guo et al., 2025) have followed.

Known limitations

Despite high benchmark scores, the robustness of LLM reasoning remains debated (Bubeck et al., 2023). Mirzadeh et al. (2024) showed that performance can decline when numerical values are changed or an irrelevant sentence is added, suggesting partial reliance on learned patterns. Data contamination is a further concern: if benchmark questions appeared in training data, high scores may reflect memorization rather than problem-solving (Dong et al., 2024). This issue is relevant here because Matura papers and marking schemes are publicly available online.

LLMs and state examinations

State examinations are useful evaluation material: they are standardized, accompanied by official marking schemes, and contain tasks of varied difficulty. In Poland, the largest related initiative is the LLMzSzŁ benchmark built from archived CKE examination papers (Jassem et al., 2025); other Polish studies evaluated models on medical examinations (Pokrywka et al., 2024; Rosoł et al., 2023). Setälä et al. (2025) used the Finnish extended mathematics Matura, and Pyeon et al. (2025) found on the Korean CSAT that performance was stronger when tasks were provided as text rather than as images, indicating that input format may significantly affect measured performance.

LLMs in mathematics education

A separate strand examines LLM-based systems as tools supporting teaching and learning, including generating explanations, providing hints and checking solutions (Kasneci et al., 2023). These functions require more than a correct final answer: the

explanation must be understandable and based on valid reasoning. Pardos and Bhandari (2024) found that ChatGPT-generated hints produced learning gains comparable to human-authored ones, although this does not imply equivalent outcomes for all systems. Reviews identify opportunities but also limited evidence, with few studies focused specifically on mathematics (Turmuzi et al., 2026; Heung & Chiu, 2025).

Managing AI in educational organizations

From an organizational perspective, the growing use of AI in schools and universities is a challenge of technology adoption, supervision, and quality assurance. Research on technology acceptance and the diffusion of innovations explains under what conditions new tools are adopted (Davis, 1989; Rogers, 2003), change-management research stresses that the benefits of a technology depend on how its introduction is led (Kotter, 1996), knowledge-management research frames the sharing of experience with AI as a knowledge process (Nonaka & Takeuchi, 1995), and service-quality research links the reliability of supporting tools to the perceived quality of the service (Parasuraman et al., 1988). AI-governance research indicates that organizations adopting AI require arrangements for oversight, accountability, and quality control (Floridi et al., 2018; Dwivedi et al., 2021). In this context, examination benchmarks may provide one source of evidence for decision-making. They can help identify tasks requiring additional human verification and demonstrate that different systems should not be treated as interchangeable. However, such benchmarks cannot, by themselves, establish the educational effectiveness of AI use.

Summary of the review

The main points of the review are summarized in Table 1, together with examples of relevant publications. Overall, the literature indicates strong but debated performance on text-based mathematical benchmarks, sensitivity to task format and visual presentation, limited evidence concerning the Polish mathematics Matura, and a management perspective in which benchmark results are one component of a broader organizational assessment rather than a direct measure of educational effectiveness.

Table 1. Summary of the literature review

Area	Key point	Example publications
Mathematical performance of LLM-based systems	Strong results on text-based benchmarks (GSM8K, MATH, MMLU), but robustness and generalizability of reasoning remain debated	Brown et al. (2020); Cobbe et al. (2021); Hendrycks et al. (2021); Ahn et al. (2024)
Step-by-step reasoning	Prompting for intermediate steps and reasoning-oriented training improve multi-step performance	Wei et al. (2022a); Kojima et al. (2022); Lewkowycz et al. (2022); OpenAI et al. (2024); Guo et al. (2025)

Area	Key point	Example publications
Known limitations	Sensitivity to problem perturbations; risk of data contamination when test material is publicly available	Mirzadeh et al. (2024); Bubeck et al. (2023); Dong et al. (2024)
LLMs and state examinations	Standardized papers with official marking schemes enable stage-by-stage evaluation; input format (text vs image) affects performance; Polish evidence on the mathematics Matura remains limited	Jassem et al. (2025); Setälä et al. (2025); Pyeon et al. (2025); Pokrywka et al. (2024); Rosoł et al. (2023)
LLMs in mathematics education	Potential for explanations, hints and feedback; empirical evidence still limited relative to availability of the tools	Kasneci et al. (2023); Pardos & Bhandari (2024); Turmuzi et al. (2026); Heung & Chiu (2025)
Managing AI in educational organizations	Adoption, change management, knowledge sharing and service quality; AI governance requires oversight, accountability and quality control	Davis (1989); Rogers (2003); Kotter (1996); Nonaka & Takeuchi (1995); Parasuraman et al. (1988); Floridi et al. (2018); Dwivedi et al. (2021)

Source: Own elaboration based on the cited literature

Research methodology

Material

The material consisted of two CKE examination papers from the May 2025 Polish mathematics Matura: the basic-level paper with 31 tasks (total 50 points) and the extended-level paper with 12 open tasks (total 50 points). The tasks differ between the levels in format and scoring. At the basic level, half of the points come from closed tasks worth 1 point each, in which only the selected answer is scored; the other half come from open tasks worth 1-4 points, and several tasks are based on visual material such as function graphs, geometric drawings, and a bar chart. At the extended level, all tasks are open, worth 2-6 points and scored stage by stage according to the CKE marking scheme; they cover more advanced material, including proofs, a trigonometric equation and a derivative-based optimization problem, and rely mainly on text and symbols rather than visual input. The two levels therefore provide complementary evidence on visual-data processing and mathematical reasoning.

The models tested were ChatGPT: GPT-5.5 Instant, Claude: Sonnet 4.6, Gemini: 3.5 Flash, Perplexity: 3.5 Turbo, all with the deep research function.

Procedure

Four language models were tested in their chat interfaces: ChatGPT, Claude, Gemini, and Perplexity. Each model received the full paper as an attachment.

To remove any effect of how the task was worded, a single, fixed prompt was used, differing only in the level and the number of tasks. Keeping the prompt constant is important because prompt wording is known to affect answer quality (Wei et al., 2022a; Kojima et al., 2022). For the basic level, it read:

"Hi, attached is the basic-level mathematics Matura paper. Take the role of a student and solve all the tasks (there are 31 of them), giving the answers and explanations."

The extended-level prompt had the same form, referring to the extended paper and its 12 tasks.

Scoring and evaluation

The answers were marked using the official CKE marking schemes, awarding points both for the final answer in closed tasks and for each correct stage in open tasks. Each model's score is given as the percentage of the maximum number of points available at that level. Reporting the score separately for each level, and recording the cause of every lost point, makes it possible to read the results not only as overall accuracy but as a profile of where each tool is and is not reliable.

Results

Basic level scores

At the basic level, all models scored highly, between 86% and 96%. Claude reached the highest score (96%) and Perplexity the lowest (86%). The results are shown in Table 2 and Figure 1.

Table 2. Model scores at the basic level

Model	Score [%]
ChatGPT – 5.5	94
Claude – 4.6	96
Gemini – 3.5	94
Perplexity – 3.5	86

Source: Own study based on research

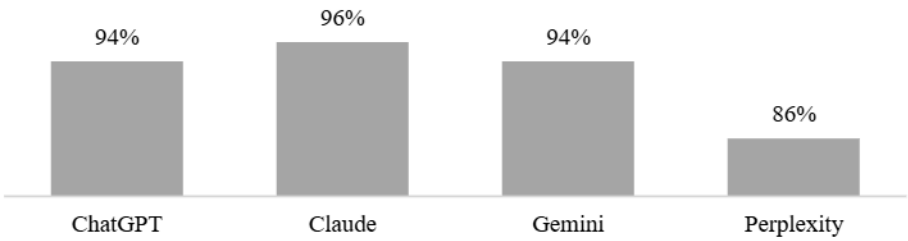


Figure 1. Model performance at the basic level

Source: Own study based on research

Basic level errors

The error analysis showed a clear pattern: most lost points came not from weak mathematics, but from misreading information given as a figure or chart. The errors are summarized in Table 3.

Table 3. Errors at the basic level

Model	Number of errors	Tasks answered incorrectly	Main cause
ChatGPT	2	18.2, 30	Misread figure / chart
Claude	1	30	Misread chart
Gemini	2	6, 30	Misread chart (30); correct work but wrong answer marked (6)
Perplexity	4	6, 11, 18, 30	Misread figure / chart (18, 30); reasoning error (11); answer not marked (6)

Source: Own study based on research

Task 30. The data (the number of students with each grade) were given as a bar chart. None of the four models read the values correctly, and all scored zero. The mathematics was not the problem: they could compute the median and the mode, but they did so on wrongly read data.

Task 18. This task was based on a drawing. ChatGPT misread the triangle's median as the median to the hypotenuse, giving a wrong value for side AC. In task 18.1 (closed), it still got the point by chance, as a second error cancelled the first; but in task 18.2 the wrong sides led to zero points. Perplexity made the same misreading of the median, which caused further errors.

Tasks 6 and 11. In task 6, Gemini and Perplexity solved the mathematics correctly but failed to mark the right answer (Gemini chose an answer that did not match its own work; Perplexity marked none). In task 11, Perplexity read the function correctly but could not give its range or the arguments for positive values, scoring half of the four points.

Extended level scores

At the extended level, the results varied much more. Claude and Gemini solved the paper without any mistakes (100%), while ChatGPT scored 64% and Perplexity only 38%. The gap between the best and the worst model was 62 percentage points, compared with just 10 points at the basic level. The results are shown in Table 4 and Figure 2.

Table 4. Model scores at the extended level

Model	Score [%]
ChatGPT – 5.5	64
Claude – 4.6	100
Gemini – 3.5	100
Perplexity – 3.5	38

Source: Own study based on research

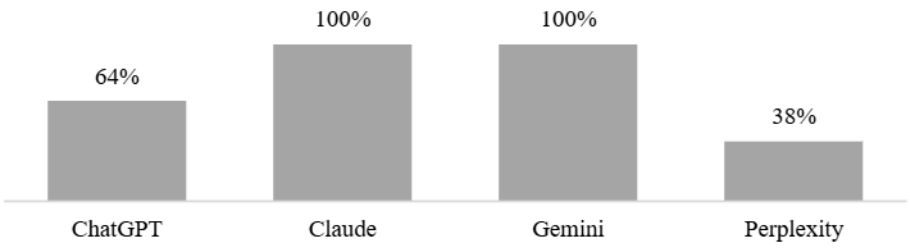


Figure 2. Model performance at the extended level

Source: Own study based on research

Extended level errors

At the extended level, where most tasks are open-ended and need a full solution, the mistakes were mainly mathematical rather than caused by misreading the data. The errors are summarized in Table 5.

Table 5. Errors at the extended level

Model	Number of errors	Tasks answered incorrectly	Main cause
ChatGPT	4	3, 8, 10, 11	Wrong method / did not match scheme
Claude	0	-	-
Gemini	0	-	-
Perplexity	7	3, 6, 7, 8, 9, 10, 11	Wrong reasoning; misread data

Source: Own study based on research

ChatGPT. The model lost points in four tasks. In task 3, it gave a wrong answer using none of the methods in the scheme, earning 1 point for one correct relation between sides. In tasks 8, 10 and 11, the solutions did not match the scheme, so no partial points were given.

Perplexity. The weakest result, losing points in most open tasks. In task 3, it gave a wrong result (1 point for a side relation), in task 6 it misread the data for the sequence, and in task 9 it misread the equation. Elsewhere, it wrote no scored steps and gave no correct answer; in task 11 it listed three conditions but developed none of them, so it scored zero.

Claude and Gemini. Both models solved all twelve tasks correctly and lost no points.

Task-level analysis

Looking at the results task by task, rather than model by model, shows which tasks were objectively difficult. Tables 6 and 7 mark, for each model, whether full marks were obtained (✓) or points were lost (✗); only tasks where at least one model lost points are shown.

Table 6. Task-level results, basic level (✓ = full marks, ✗ = points lost)

Model / Task	6	11	18	30
ChatGPT	✓	✓	✗	✗
Claude	✓	✓	✓	✗
Gemini	✗	✓	✓	✗
Perplexity	✗	✗	✗	✗

Source: Own study based on research

Table 7. Task-level results, extended level (✓ = full marks, ✗ = points lost)

Model / Task	3	6	7	8	9	10	11
ChatGPT	✗	✓	✓	✗	✓	✗	✗
Claude	✓	✓	✓	✓	✓	✓	✓
Gemini	✓	✓	✓	✓	✓	✓	✓
Perplexity	✗	✗	✗	✗	✗	✗	✗

Source: Own study based on research

Two patterns are visible in the task-level data. At the basic level, task 30 forms a full column of failures: all four models lost points on it. At the extended level, the four tasks that ChatGPT failed (3, 8, 10 and 11) are exactly a subset of the seven tasks that Perplexity failed – the tasks that tripped the stronger of the two weaker models also tripped the weaker one, while Claude and Gemini solved everything.

Types of errors

The errors can be grouped into three types: misreading visual or given data (charts, figures, reading the answer off a drawing), mistakes in mathematical reasoning or method, and failures only in reporting the final answer (the mathematics was correct, but no valid answer was selected). Table 8 counts how often each type appeared at each level.

Table 8. Types of errors by level (model-task instances where points were lost)

Error type	Basic level	Extended level
Visual / data reading	6	2
Mathematical reasoning / method	1	9
Answer reporting only	2	0
Total error instances	9	11

Source: Own study based on research

Discussion

Performance and errors across the two levels

Comparing both levels gives two observations. At the basic level, the differences between models were minor, and the main limit was a shared weakness in reading visual information (Pyeon et al., 2025). At the extended level, a clear split appeared: Claude and Gemini stayed at the top, while ChatGPT and especially Perplexity fell behind. As tasks get harder and more open-ended, the real differences in reasoning become much more visible than in closed tasks, in line with reports that harder, less familiar problems expose the limits of these models (Mirzadeh et al., 2024).

The contrast between error types is striking. At the basic level, most errors (6 of 9) came from misreading visual information, whereas at the extended level, most errors (9 of 11) came from reasoning or method. In short, the basic level mainly tested whether a model could read the task; the extended level tested whether it could reason through it, in line with Pyeon et al. (2025) and Mirzadeh et al. (2024).

Claude and Gemini lost points at the basic level only on visual-reading tasks, so for the strongest models the main obstacle was reading visual information, not mathematics. The ranking was only partly stable: Claude stayed at the top, and Perplexity scored last at both levels, while ChatGPT showed the largest decline, from 94% to 64%.

Reading the results more closely

A correct closed answer can hide wrong reasoning. In task 18.1, ChatGPT earned the point even though its method was wrong: two separate errors happened to cancel each other out. Because closed tasks award points only for the final answer, a correct answer does not guarantee correct reasoning, and basic-level scores may slightly overstate true ability. This is consistent with the view that models can reach the right output by pattern matching rather than reasoning (Mirzadeh et al., 2024). Partial credit came from the setup, not the full solution. In extended task 3, both weaker models earned a single point for correctly writing a relation between the sides, but neither could carry the solution through. The models captured the surface structure of the task while failing at the steps that required sustained reasoning.

All four models would pass the basic matura. Every model scored well above the 30% threshold required to pass the basic-level mathematics matura; even the lowest basic score (86%) is a strong result. At the extended level, which has no pass threshold, the scores instead serve to separate the models by ability.

Implications for the management of educational organizations

What is at stake is not whether these tools should solve examinations in place of students, but whether students and teachers can rely on them to study, check a solution or explain a step; the examination is a controlled way to measure how far that reliance is justified. At the basic level, the tools proved dependable enough for such everyday support, whereas the wide gap at the extended level (38-100%) shows that they are not interchangeable, so recommending a particular tool is a decision that needs evidence.

The error analysis shows where relying on these tools is riskiest. The clearest shared weakness was not in mathematics but in reading data presented as a chart or figure: on task 30, every model failed. A student who turns to such a tool to check a statistics task may be confidently misled, because the calculation looks sound and only the reading of the data is wrong. For teachers, and for the staff responsible for the quality of teaching, this marks the points at which the output of these tools should be treated with most caution, and a human check is most worth its cost.

Two features make such errors easy to miss. A solution can reach the right answer through faulty reasoning, as in task 18.1, where two mistakes cancelled each other out; and a wrong solution can still look orderly and plausible. Both are more dangerous for someone who is still learning than an obvious error would be, because nothing signals that anything has gone wrong. Guidance on using AI in teaching therefore cannot rest on final answers alone, and, given how often the underlying models change, it needs to be revisited rather than settled once. In terms of the management literature, relying on these systems is less a single decision than a continuing process of change and governance (Kotter, 1996; Floridi et al., 2018).

Conclusions, limitations and future work

Conclusions

This study evaluated the performance of four user-facing, LLM-based AI systems. They were tested through their chat interfaces using complete examination papers and a fixed prompt. Three principal findings emerge:

1. All four systems achieved high scores at the basic level, ranging from 86% to 96%. At the extended level, performance was much more varied, ranging from 38% to 100%.
2. The dominant error profile differed between the two examination levels. At the basic level, most lost points resulted from incorrect interpretation of visual or given information, particularly charts and figures. At the extended level, errors were more frequently associated with mathematical reasoning, the selected method, or the failure to provide steps recognized by the official marking scheme.
3. The findings are broadly consistent with earlier studies reporting weaker performance on visually presented tasks and greater variation on more demanding mathematical problems (Pyeon et al., 2025; Mirzadeh et al., 2024; Setälä et al., 2025).

For educational organizations, the study offers an exploratory diagnostic perspective. The identified error patterns may help indicate situations in which human verification is particularly important, including tasks based on charts, figures, and extended mathematical reasoning. The results may support periodic quality assessment and the development of guidance for the responsible use of AI systems. They should, however, constitute only one element of a broader assessment of educational usefulness and organizational risk.

Limitations

1. Only one examination paper at each level was analyzed. The sample is therefore too small to support broad generalizations, especially in relation to the 100% scores achieved by two systems at the extended level (Jassem et al., 2025).
2. Each paper was submitted only once to each system. Because generative AI systems are not fully deterministic, repeated runs might produce different solutions and scores. The study does not therefore estimate the variability or statistical significance of the observed differences.
3. The CKE examination papers and marking schemes are publicly available. It cannot be fully excluded that the systems had previously encountered some of this material during training or that research-enabled functions could access related online content. High performance may therefore reflect a combination of problem-solving capability, prior exposure and retrieval rather than reasoning alone (Dong et al., 2024).
4. The systems were tested through user-facing interfaces available at a particular time. Their results may have depended on model versions, system instructions, document-processing mechanisms, visual-recognition components, retrieval functions and other tools.
5. The use of a single fixed prompt improved consistency across the systems but does not show how alternative prompting strategies might affect performance (Wei et al., 2022a; Kojima et al., 2022). Finally, the assessment of open-ended answers involved some judgement and was not independently repeated by a second evaluator.

Directions for future work

Future research should extend the analysis to examination papers from multiple years and include repeated runs for each system, to assess the stability of the findings. Comparing user-facing systems with controlled API access, with retrieval and additional tools documented or disabled, would help distinguish the performance of complete consumer systems from that of the underlying models. A useful diagnostic extension would be to present visually based tasks in both their original and fully textual forms, to isolate errors resulting from visual-data processing (Pyeon et al., 2025), while continuing to assess open tasks stage by stage according to the CKE marking scheme (Jassem et al., 2025). Finally, research involving students and teachers could examine the quality of explanations, effects on independent reasoning and actual learning outcomes, and organizational research could analyze how schools and universities establish rules for AI use, train staff, and define situations requiring human oversight.

References

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). Large language models for mathematical reasoning: Progresses and challenges. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 225-237. https://doi.org/10.18653/v1/2024.eacl-srw.17

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. V., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv. https://arxiv.org/abs/2108.07258

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training verifiers to solve math word problems*. arXiv. https://arxiv.org/abs/2110.14168

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340. https://doi.org/10.2307/249008

Dong, Y., Jiang, X., Liu, H., Jin, Z., & Li, G. (2024). *Generalization or memorization: Data contamination and trustworthy evaluation for large language models*. arXiv. https://arxiv.org/abs/2402.15938

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*, 689-707. https://doi.org/10.1007/s11023-018-9482-5

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., … Zhang. Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*, 633-638. https://doi.org/10.1038/s41586-025-09422-z

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. X., & Steinhardt, J. (2020). *Measuring massive multitask language understanding*. arXiv. https://arxiv.org/abs/2009.03300

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. https://arxiv.org/abs/2103.03874

Heung, Y., & Chiu, T. K. F. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2025.100361

Jassem, K., Ciesiólka, M., Graliński, F., Jabłoński, P., Pokrywka, J., Kubis, M., Jabłońska, M., & Staruch, R. (2025). *LLMzSzŁ: A comprehensive LLM benchmark for Polish*. arXiv. https://arxiv.org/abs/2501.02266

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv. https://arxiv.org/abs/2001.08361

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

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large language models are zero-shot reasoners*. arXiv. https://arxiv.org/abs/2205.11916

Kotter, J. P. (1996). *Leading change*. Harvard Business School Press.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). *Solving quantitative reasoning problems with language models*. arXiv. https://doi.org/10.48550/arXiv.2206.14858

Mirzadeh, I., Alizadeh-Vahid, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models*. arXiv. https://arxiv.org/abs/2410.05229

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press. https://doi.org/10.1093/oso/9780195092691.001.0001

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 technical report*. arXiv. https://arxiv.org/abs/2303.08774

OpenAI, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Allison Tam, Bennett, A., … Li, Z. (2024). *OpenAI o1 system card*. arXiv. https://arxiv.org/abs/2412.16720

Parasuraman, A., Zeithaml, V., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, *64*(1), 12-40.

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS One*, *19*, Article e0304013. https://doi.org/10.1371/journal.pone.0304013

Pokrywka, J., Kaczmarek, J., & Gorzelańczyk, E. (2024). *GPT-4 passes most of the 297 written Polish board certification examinations*. SSRN. https://doi.org/10.2139/ssrn.4861528

Pyeon, G., Heo, I., Jung, J., Hwang, T., Namgoong, H., Seo, H., Han, Y., Kim, E., Kang, H., & Jung, S. (2025). *Evaluating large language models on the 2026 Korean CSAT mathematics exam: Measuring mathematical ability in a zero-data-leakage setting*. arXiv. https://arxiv.org/abs/2511.18649

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., & Młyńczak, M. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Scientific Reports*, *13*, Article 20512. https://doi.org/10.1038/s41598-023-46995-z

Setälä, M., Sikström, P., Heilala, V., & Kärkkäinen, T. (2025). Assessment of evolving large language models in upper secondary mathematics. *Proceedings of ICETC 2025*, 210-214. https://doi.org/10.1109/ICETC66579.2025.11387533

Turmuzi, M., Azmi, S., & Kertiyani, N. M. (2026). ChatGPT in school mathematics education: A systematic review of opportunities, challenges, and pedagogical implications. *Teaching and Teacher Education*, *170*, Article 105286.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.03762

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022b). Emergent abilities of large language models. arXiv. https://arxiv.org/abs/2206.07682

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Xia, F., Le, Q., & Zhou, D. (2022a). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://arxiv.org/abs/2201.11903

Zhao, W., Zhou, K., Junyi, L., Tianyi, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J. R. (2026). A survey of large language models. *Frontiers of Computer Science*, *20*, Article 2012627. https://doi.org/10.1007/s11704-026-60308-3

DOI: 10.17512/znpcz.2026.3.10

Authors' Contribution: Maciej Renz: conducting the study, data collection and analysis, preparation of results. Grzegorz Borkowski: research idea, article outline, initial literature selection. Ewelina Ząb: editing, formatting, finalization of the theoretical section. All authors contributed to the writing and revision of the manuscript and approved the final version.

Acknowledgements and Financial Disclosure: This research received no external funding. The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of Interest: The authors declare no conflict of interest.

Ethics Declaration: This study did not require ethical clearance, as it involved no human or animal subjects, and all materials used were publicly available examination papers.

AI Declaration: During the preparation of this work, the authors used AI-based tools (spelling and grammar checkers) for language polishing and formatting assistance. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

OCENA SYSTEMÓW AI OPARTYCH NA LLM NA EGZAMINIE MATURALNYM Z MATEMATYKI: IMPLIKACJE DLA ZARZĄDZANIA SZTUCZNĄ INTELIGENCJĄ I ZAPEWNIANIA JAKOŚCI W ORGANIZACJACH EDUKACYJNYCH

Streszczenie: Systemy AI oparte na dużych modelach językowych (LLM – Large Language Model) są coraz częściej wykorzystywane w edukacji, między innymi jako narzędzia do rozwiązywania, wyjaśniania i sprawdzania zadań matematycznych. W tym eksploracyjnym badaniu oceniono skuteczność czterech powszechnie używanych, dostępnych dla użytkowników systemów AI (ChatGPT, Claude, Gemini i Perplexity) na maturze z matematyki na poziomie podstawowym i rozszerzonym. Systemy testowano przez ich interfejsy czatu z użyciem oficjalnych arkuszy CKE, przekazywanych jako kompletne załączniki, oraz jednego, stałego promptu dla każdego poziomu. Odpowiedzi oceniano według oficjalnych zasad oceniania CKE, przyznając punkty za odpowiedzi końcowe oraz za poprawnie wykonane etapy zadań otwartych. Wszystkie systemy uzyskały wysokie wyniki na poziomie podstawowym (86-96%), natomiast na poziomie rozszerzonym wyniki były znacznie bardziej zróżnicowane (38-100%). Na poziomie podstawowym większość utraconych punktów wynikała z błędów w interpretacji wykresów, rysunków i innych informacji wizualnych. Na poziomie rozszerzonym błędy częściej wiązały się z rozumowaniem matematycznym lub doborem metody. Wyniki pokazują wartość uzupełniania łącznych wyników egzaminacyjnych o diagnostyczne profile błędów, rozróżniające przetwarzanie danych wizualnych od rozumowania matematycznego. Dla organizacji edukacyjnych takie benchmarki mogą stanowić jedno ze źródeł dowodów na potrzeby nadzoru ludzkiego, identyfikacji ryzyka i okresowej oceny jakości. Wyniki odzwierciedlają jednak działanie kompletnych systemów AI w określonych warunkach interfejsu i nie powinny być interpretowane jako bezpośredni dowód ich ogólnej skuteczności edukacyjnej. Potrzebne są dalsze badania obejmujące większą liczbę arkuszy egzaminacyjnych, powtarzanych prób oraz użytkowników edukacyjnych.

Słowa kluczowe: zarządzanie AI (AI governance), analiza błędów, systemy AI oparte na LLM, rozumowanie matematyczne, edukacja matematyczna, zapewnianie jakości

Articles published in the journal are made available under Creative Commons Attribution 4.0 International.
Certain rights reserved for the Czestochowa University of Technology.

