Large language models as formative assessment and feedback tools? – A systematic report

Frederik Dilling

University of Siegen, Germany; dilling@mathematik.uni-siegen.de

This paper discusses the use of large language models (LLMs) for formative assessment and feedback in mathematics education. First, a brief introduction to the research on LLMs in mathematics education is given. Subsequently, the LLM ChatGPT 4.0 is systematically evaluated with regard to the aspects 1) Input and localization, 2) Assessment-quality, 3) Content and form of feedback, and 4) Adaptivity and receiver of feedback. It is shown that ChatGPT has the potential to provide meaningful feedback on mathematical work, but that its use is associated with a number of challenges.

Keywords: Artificial intelligence, ChatGPT, Digital assessment and feedback, Formative assessment, Large language models

Introduction

Artificial intelligence (AI) is currently a highly debated topic. At the latest with the free publication of GPT-3 in November 2022, the discussion has arrived in society. The opportunities and challenges for the educational sector were also quickly addressed. For example, a study by Kung et al. (2023), which found that ChatGPT could pass the three-part American medical licence test (USMLE) without further training, received a great deal of media attention. Opportunities and challenges were also investigated in the field of mathematics education. Wardat et al. (2023) conducted interviews with students and teachers and found that ChatGPT is generally perceived as a useful educational tool, but that it does pose some challenges (e.g. development of misconceptions). Other authors, however, analyzed ChatGPT from a theoretical perspective and through extensive testing. For example, Buchholtz et al. (2023) come to rather negative conclusions on this basis and state that the generative AI ChatGPT is not yet suitable for use in mathematics classes.

While AI has been studied intensively in educational research from a technical perspective for around 10 years (see, e.g. "International Journal of Artificial Intelligence in Education"), mathematics education research has only recently begun to address the topic, as can be seen above. In particular, so-called large language models (LLMs) such as ChatGPT are being considered. LLMs are linguistic models that have been trained with a huge amount of text data and are intended to simulate communication. With the help of probability trees, answers to user requests (so-called prompts) are generated. Although the system was trained for linguistic knowledge, it can also contain rational knowledge from the training data (Petroni et al., 2019). However, knowledge databases are not accessed for the answers; the "knowledge" comes solely from the trained linguistic model, which can also result in the output of incorrect information. Kasneci et al. (2023) explain in the context of LLMs in education:

> "Large language models can help teachers to identify areas where students are struggling, which adds to more accurate assessments of student learning development and challenges. Targeted

instruction provided by the models can be used to help students excel and to provide opportunities for further development." (p. 3)

Initial studies on formative assessment with LLMs have already been carried out. Moore et al. (2022) used a fine-tuned GPT-3 model to evaluate student answers in chemistry education and concluded that it is a powerful tool to assist teachers in the quality of their evaluations of students. Zhu and Liu (2020) found that LLMs can support high school students in scientific reasoning in the area of climate activity. Sailer et al. (2023) observed in a teacher education program that the use of LLMs leads to better justifications of diagnoses of students' learning difficulties.

In this article, the opportunities and challenges of ChatGPT as a formative assessment and feedback tool in mathematics education will be discussed. For this purpose, the LLM is tested and analyzed in detail against the background of a previously described framework based on Fahlgren et al. (2021).

Digital formative assessment and feedback – A framework for the analysis

The basis for the analysis framework in this article is the survey report by Fahlgren et al. (2021) on technology-rich assessment in mathematics. In this report, the research or development projects STACK, STEP and SMART are analyzed and compared against the background of selected categories. The categories considered include localization, receiver, content and form, and adaptivity. Localization refers to whether the assessment and feedback takes place on a micro-level (e.g. task level) or a macro-level (e.g. overall performance). The receiver of feedback is the person to whom the feedback is directed (e.g. student, teacher). The content of feedback can also differ (e.g. right/wrong, hints or error information, worked out examples) as well as the form in which it is presented (e.g. language, pictures). Adaptivity refers to the extent to which aspects of the student response appear in the feedback.

For the analysis of ChatGPT as a potential feedback and assessment tool in this article, the above categories were slightly modified. For this purpose, the aspect of localization was expanded to include the possibilities of input by the user. The adaptivity and receiver categories were combined into one analysis aspect. The aspect of assessment quality, which is important to examine in the field of generative AI, has been added. The analysis in this article is therefore based on four aspects:

- Input and localization
- Assessment-quality
- Content and form of feedback
- Adaptivity and receiver of feedback

Evaluation of the large language model ChatGPT

The LLM analyzed in this article is the latest version of ChatGPT 4.0 at the time of the analysis. Extensive testing was carried out with this LLM in January 2024. The Wolfram plugin, which establishes a connection to the Wolfram Alpha computer algebra system (CAS), was used for all user requests. The testing was based on selected tasks in the field of linear algebra and analytical geometry from the publicly accessible mathematics secondary school examinations in North Rhine-Westphalia (Germany) (see https://www.standardsicherung.schulministerium.nrw.de/cms/zentralabitur-gost/faecher/fach.php?fach=2). The tasks include, for example, an inner-mathematical task in which

distances between two points had to be determined and a third point had to be chosen so that a right angle is formed, or an application task with a real-world context in which it had to be justified that the base of a pyramid lies in one of the coordinate planes, it had to be shown that three given corner points of the pyramid approximately form an equilateral triangle with a certain edge length, and the plane in which these three points are located had to be determined. The tasks were translated into English and formulas were converted into LaTex notation. In addition, sample solutions with different types of errors were created, which were then entered into ChatGPT with different prompts. The conversations with ChatGPT formed the data basis for the analysis.

The results of the analysis with regard to the four aspects mentioned above are presented below and explained at selected points with examples from the testing. For readability, the explication is limited to one task in which a linear system of equations is to be solved. It should be emphasized at this point that although the analysis framework used was precise and comparable with other studies, the data collection was rather exploratory and not systematic, which is why this is not a scientific study, but rather an experience report.

Aspect 1: Input and localization

As described above, LLMs like ChatGPT are developed especially for processing text and are intended to simulate communication. Therefore, the input and also the output is mainly symbolic as text. If the Wolfram plugin is switched off, it is also possible to input images or sound recordings, although the analysis options in this case are very limited beside the recognition of text on the images or in the recordings.

Mathematical formulas can be entered in any programming language or as a kind of pseudo-code. In this experiment, formulas were entered as LaTex codes. Various external software is available that automatically converts handwritten formulas or texts into equivalent LaTex codes, which can then be copied into the text input field of ChatGPT.

To enable assessment and feedback by ChatGPT, both the task and the user's own solution must be entered. To complete the prompt, it is also necessary to explain what is to be done in relation to the task and the solution (e.g. "Can you tell me if this is correct?"). Characteristic for an LLM such as ChatGPT is the possibility to ask follow-up questions after the response of the system, which then also changes the feedback (e.g. "Please tell me where exactly the error occurs."). The further conversation automatically includes the previous requests and answers.

However, retaining the context of a conversation is only possible to a certain extent. The number of tokens that can be used (8192 tokens for ChatGPT 4.0 at the time of analysis) limits the number of analysis units included, whereby according to ChatGPT, one token corresponds to approximately four characters. This means that the localization of the assessment and feedback provided by ChatGPT is more likely to be at the micro level. Although several tasks and solutions can be included, a structured modeling of learning paths over a longer period of time is not possible.

Aspect 2: Assessment-quality

In terms of mathematical correctness, the detailed testing revealed that many of the calculated results were correct and that the feedback on the user's solution was also correct on this basis. The Wolfram

plugin establishes the connection between the LLM and the CAS Wolfram Alpha. This means that potentially all operations that can be performed by a CAS (e.g. transformation of equations, calculation of derivatives and integrals) can also be used by ChatGPT. Nevertheless, ChatGPT may misinterprets information when analyzing the task and thus sends an incorrect request to Wolfram. It is also possible that ChatGPT misinterprets the results from Wolfram and thus provides incorrect feedback to the user. However, in the responses from ChatGPT it is made transparent which requests were sent to Wolfram and which responses were given. This means that errors can be quickly identified in most cases. Errors also frequently occur if the Wolfram plugin is not used. This can be problematic, as the reasoning around the incorrect calculations can still be plausible and therefore potentially not recognized as wrong by non-experts (Buchholtz et al., 2023). It is therefore recommended to always explicitly state in a request that the Wolfram plugin should be used.

In general, ChatGPT has proven to be quite reliable in processing math problems, even when real-world contexts occur. However, there are actually no reliable figures on mathematical correctness of ChatGPT responses. Some uncertainties occur in the answers of ChatGPT when conceptual questions are asked (e.g. "What is a probability?"). Problems also arise when mathematical processes outside the capabilities of a CAS are requested. For example, ChatGPT is comparatively unreliable at outputting mathematical proofs or evaluating given proofs. This is mainly due to the fact that LLMs are not mathematically logical systems. Although proofs of classical mathematical theorems can be generated, circular reasoning or incorrect derivations often occur.

Aspect 3: Content and form of feedback

The feedback that ChatGPT provided on the user's solutions in the testing had different contents. If no further information on the desired content is provided (e.g. only asking "Can you tell me, if my solution is correct?"), ChatGPT usually presents the correct solution and uses this as the basis for judging whether the user's solution is correct or incorrect. Figure 1 shows an example in which a system of linear equations consisting of the equations $4x + 2y - 3z = 8$, $2x - 3z = 2$ and $6x + 2y - 5z = 10$ had to be solved. The system of equations was solved by hand on paper, whereby a transformation error was deliberately included in the first step. The entire solution was converted into a LaTex code and entered together with the question "I have calculated like this. Can you tell me if this is correct? Use the Wolfram Plugin for your calculations.". The response from ChatGPT can be seen in Figure 1. The Wolfram plugin was used to calculate the solution, and this was displayed in the response. In addition, feedback was given that the calculated solution differs from the user's input.

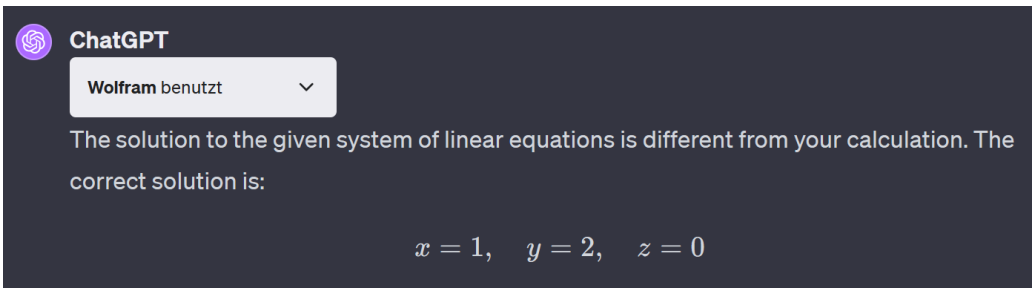

Figure 1: Right-Wrong feedback by ChatGPT

When determining the correct solution, the solution process is often presented directly as a kind of sample solution and provided with some explanations. However, this feedback can also be deliberately generated using suitable prompts, e.g. "Can you solve the task step by step and explain it?". Figure 2 shows such a query in relation to the task above. ChatGPT first writes down the system of linear equations to be solved and explains that this can be solved using the method of Gaussian elimination. The subsequent solution process is divided into four steps. In the first step, the augmented matrix is written down. The second and third steps consist of transformations that convert the matrix into a triangular form. In the fourth step, the concrete values for x, y and z are determined by substitution. The solution is described in relatively great detail and some instructions are given on why the individual steps should be carried out.

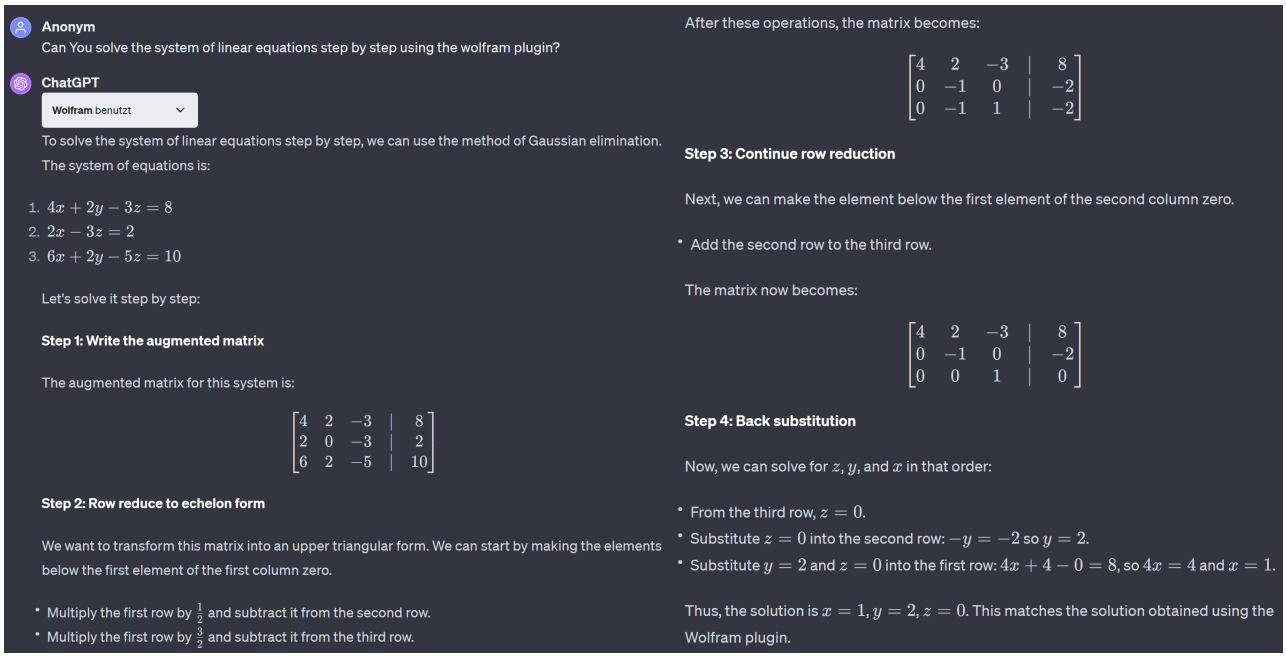

Figure 2: Solving and explaining a task step by step by ChatGPT

In most cases, an explicit prompt is required for a specific feedback content. For example, it is possible to ask at which point the error occurs. ChatGPT can specify this sometimes more and sometimes less precisely and correctly. One of the reasons for this is that the Wolfram plugin does not output the solution path to ChatGPT, but only the final results – against this background, the solution path must therefore be checked by ChatGPT itself.

In Figure 3, for the example above, the user asks where exactly the error is located in the calculation. ChatGPT then checks whether the row operations from the first step are correct and determines that an incorrect equation has been set up in the third line. This is actually the row in which the error was deliberately inserted for the testing. However, when determining the correct equation, ChatGPT makes a mistake, too, and states that the equation $0 + 2y - z = 6$ is created by double subtracting the second line from the third line. This operation actually results in $2x + 2y + 1 = 6$, which does not help to solve the system of equations.

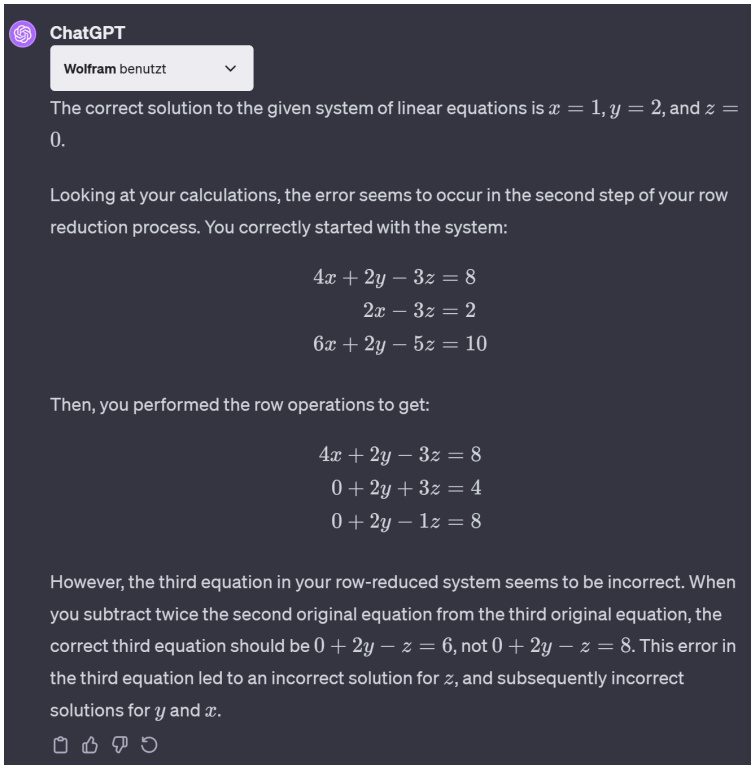

Figure 3: Locating the error by ChatGPT

Many other feedback contents can be generated using suitable prompts. For example, a mathematical process that is related to the task can be explained again. In relation to the task above, for example, you can ask how linear systems of equations can be solved. In a test, ChatGPT has listed different methods (graphical, substitution, elimination, Gaussian matrix elimination, Cramer's rule, software or calculators), briefly explained them and compared them according to the criteria "applicable for" and "best for".

In addition, further exercises including sample solutions can be generated. For this purpose, certain criteria can also be specified for the task, e.g. that only integers should occur in the matrix or that the task is integrated into a real context. A query for the above task resulted in an application, that concerns the relationship between costs and the manufacture of three different products in a company. This is certainly an authentic context in the field of linear equations. However, the task resulted in a system of linear inequalities. From a didactical perspective, it is questionable whether this is a good subsequent task for students struggling with solving systems of linear equations.

Finally, it should be emphasized at this point that it is also possible to deliberately avoid certain feedback content in ChatGPT. This is particularly interesting with regard to the output of the solution. For example, prompts can be formulated in which it is specified that the solution should not be given to the users under any circumstances, but only hints should be provided to help them solve the task or find the errors themselves. A customized version of ChatGPT called "Soctratic Tutor" is already provided specifically for this purpose. If the user asks the tutor how to solve the system of equations mentioned above, the tutor will ask suitable questions in order to help the user actively work on the task step by step (e.g. “Can you identify how many variables and how many equations are presented in this system? And why is this information important for solving a system of linear equations?”).

In addition to the content of the feedback, ChatGPT also allows different forms of presentation. The focus is on feedback in written form. ChatGPT uses in particular mathematical terminology, which is characterized by the corresponding mathematical terms. However, isolated elements of colloquial language can also be identified in the conversation. In relation to the tasks, algebraic expressions can be found at many points. While the input by the user is carried out using LaTex codes or similar codes, which are somewhat confusingly displayed, the formulas in the ChatGPT responses are easy to read (Figures 2-3). In addition to language feedback, ChatGPT can also output iconic representations. For plotting the above linear equation system, for example, it can use the Wolfram plugin and display an appropriate three-dimensional graphic of three intersecting planes.

Aspect 4: Adaptivity and receiver of feedback

The intensive testing has shown that the level of adaptivity of ChatGPT can be very high. For example, the feedback always refers specifically to the task set at the beginning. Instead of giving a general description of the solution procedure, the solution path can be presented step by step according to the task. When checking the user's own solution, ChatGPT can even localize errors in the solution path to a certain extent and is not limited to comparing the final results with the results determined by the Wolfram plugin.

In addition to adaptivity in relation to the task and the solution to be checked, suitable prompts can also be used to make adjustments in relation to the user's characteristics. For example, a studied mathematician should receive different feedback than a student in secondary school. At the beginning of a chat or in the settings of the ChatGPT account, relevant information can be entered. The more detailed the information is, the more accurately ChatGPT can take it into account in the conversation. For example, language difficulties of the user can be pointed out so that responses use simple words and short sentences. It can also be emphasized as positive that users can work with ChatGPT in the language in which they feel most confident regarding mathematics.

Conclusion

The previous analysis has shown that ChatGPT already offers a remarkable amount of potential for formative assessment and feedback in the field of mathematics. This is in line with the results already obtained in studies outside mathematics (Moore et al., 2022; Zhu & Liu, 2020; Sailer et al., 2023). The right prompting has proven to be an important success factor for appropriate feedback. The prompt largely determines the form and content of the feedback, the correctness of the performed calculations and the extent to which the response is adapted to the feedback receiver.

However, a number of challenges remain. Probably the most important one concerns the mathematical correctness of the calculations performed by ChatGPT, which is not guaranteed. This is particularly problematic because learners often do not have the competencies to recognize the errors. Students therefore need well-developed reflection skills, a critical handling of the system and close support from the teacher.

With regard to the analysis categories according to Fahlgren et al. (2021), the strengths of ChatGPT lie in particular in the adaptivity of the responses in relation to the task and the feedback receiver (Aspect 4) as well as the content and form of the feedback (Aspect 3). The possibilities for input are

largely restricted to text and localization is limited to the micro level (Aspect 1). The assessment quality is high for numerical and symbolic calculations as considered in this analysis, but there is generally no certainty for the correctness of the responses (Aspect 2). Thus, LLMs such as ChatGPT cannot replace the assessment and feedback systems developed specifically for learning mathematics – but if used appropriately, they can be suitable additions. The future will show how the development of generative AI will progress and how this will affect the opportunities and challenges identified above.

References

Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Pohl, M., & Schorcht, S. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative KI-Sprachmodelle. *Mitteilungen der GDM, 114*, 19–24.

Fahlgren, M., Brunström, M., Dilling, F., Kristinsdóttir, B., Pinkernell, G. & Weigand, H.-G. (2021). Technology-rich assessment in mathematics. In: A. Clark-Wilson, A. Donevska-Todorova, E. Faggiano, J. Trgalova, & H.-G. Weigand (eds.), *Mathematics Education in the Digital Age* (pp. 69–83). Routledge.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences, 103*, 102274.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE. *PLoS digital health, 2*(2), e0000198.

Moore, S., Nguyen, H.A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. *Proceedings of EC-TEL 2022*, 243–257.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? *arXiv preprint* arXiv:1909.01066.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83*, 101620

Wardat, Y., Tashtoush, M., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(7), em2286.

Zhu, O.M., & Liu, H.L. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668.