Journal of Mathematics Education
2024, Vol. 17, No. 1, 1-15

https://doi.org/10.26711/007577152790171

Challenges in Using ChatGPT for Assessing Conceptual Understanding in Mathematics Education

Corinna Hankeln
TU Dortmund University

Students' learning progressions in mathematics need to cover both procedural skills and conceptual understanding. While the first can more easily be assessed by closed item formats, in-depth insight into students' understanding is often based on self-explanations, reasoning and justifications, thus assessed in open-ended tasks. The accessibility of Generative AI, like ChatGPT, with its ability to understand natural language, makes the automatic real-time evaluation of these item formats seem feasible in near future. However, a Large-Language-Model (LLM) only models human language and the way it "understands" concepts differs from human understanding. This paper explores which consequences this difference has for the assessment of conceptual understanding on the example of multiplication and explains differences by the underlying mechanisms of ChatGPT.

Keywords*:* conceptual understanding, generative AI assessment, ChatGPT mechanisms, multiplication understanding, open-ended task evaluation

Technology enables assessment tools to become more sophisticated in evaluating students' performance and presenting results instantly (Schoenfeld, 2017). However, in mathematics education, this ability has also led to negative effects, as digital assessments tend to focus more on calculations and standard procedures (Hoogland & Tout, 2018), which are easier to construct than assessment tasks of problem solving, reasoning and effective mathematical communication. This is partly because open-ended items, where students show their conceptual understanding through problem-solving and reasoning, cannot (or could not) be scored quickly and deeply by an automatic system (Schoenfeld, 2017). The emergence of Large Language Models (LLM) such as ChatGPT (OpenAI, 2023) or BERT (Devlin et al., 2019) offers new hope to reverse this trend, as this technology can understand and respond to natural human language, allowing the assessment of open-ended responses (Latif & Zhai, 2024). LLMs have rapidly developed the ability to comprehend and interpret human language, deciphering context, meaning, grammar and semantics (Latif & Zhai, 2024). However, they are still models of human language and their learning and skill acquisition differs from human learning. This paper, which is theoretical, but based on empirical case studies, aims to explore how the fundamental differences in understanding between humans and LLMs affect the assessment of conceptual understanding. Using empirical data from a test of conceptual understanding with open-ended items (Hankeln et al., submitted), it illustrates the opportunities and challenges, stemming from the different modes related to processing information processing.

What "Understanding" Means for an LLM

This section briefly explains how an LLM, specifically Chat-GPT, processes natural language, extracts meanings and produces human-like answers, without going into technical details. LLMs use patterns of word co-occurrence in natural language to learn distributed, context-sensitive representations of linguistic meaning at multiple levels, and generate probabilistic predictions of likely next words (Suresh et al., 2023). These patterns are given to an LLM as training data, which, along with the algorithms, determines the quality of the model: the better the data and its exploitation, the better the outcomes. Current LLMs like ChatGPT are trained on huge text data and combine various techniques such as deep learning, unsupervised learning, instruction fine-tuning, multi-task learning, in-context learning and reinforcement learning (Wu et al., 2023). However, they are essentially statistical models that describe the probability distribution of natural language (Zhai, 2009). These language models have been enhanced to pre-trained models that used self-supervised learning over raw large-scale texts (Wu et al., 2023), which improved the semantic description of words from static to dynamic context-aware representation. The proposition of transformers, beginning with the BERT-model (Devlin et al., 2019), made it possible to generate contextualized word vectors (Li et al., 2022), that consider the surrounding of a word in a phrase, not only in a single sentence, but also across several prompts. As the meaning of a word often depends on its position and the accompanying words in the sentence, the use of a transformer model (Vaswani et al., 2017), was a major advance in natural language processing (Li et al., 2022). The transformer model is a type of neural network architecture that is specifically designed for tasks like machine translation or text summarisation (Briganti, 2024). The process between input and output in a pre-trained transformer can be illustrated as follows (example adapted from Briganti (2024)):

When a user enters a prompt, such as "Write down a strategy for the multiplication of two two-digit numbers for grade two students," ChatGPT tokenizes the sentence and maps tokens to nearby points in the vector space. For instance, the word "multiply" could have a high value for "calculus", while the vector for "grade two students" could have a high value for "age group" or "academic level". A recurrent neural network processes these vector sequences and identifies the relationships between the tokens. A decoder processes the encoded input and produces outcome vectors, in this example maybe different ways to explain the multiplication. The decoder uses attention mechanisms to selectively focus on certain parts of the input. For example, an outcome vector that would recommend written multiplication methods ignores the tokens "grade two students," as those students do not usually know this method at this academic stage. Finally, a probability distribution is generated over all outcome vectors, showing the likelihood of each strategy for a grade two student. This also shows the influence of the training data. If ChatGPT lacks the information that grade two students do not know written multiplication, this unsuitable recommendation will have a higher probability. The algorithm was trained with several strategies, including reinforcement, where humans compared different possible outcomes and chose the more appropriate one, allowing the algorithm to learn, so ChatGPT's processing goes beyond pure statistical language models.

This description of the process is of course strongly simplified and varies depending on the language model and its mechanisms (Wu et al., 2023). The exact algorithms how an LLM makes a decision, and the training data are mostly a black box for the user and involve many parameters and layers. This opacity led to the rise of research on explainable AI (xAI) (for an overview see Barredo Arrieta et al., 2020). LLM-Understanding is thus assigning input words to nearby concepts, finding relations between words and weighing different answer options.

What "Understanding" Means for Humans in Mathematics Education

This paper focuses on conceptual understanding, which can be seen as intertwined with procedural knowledge (Pettersson et al., 2019). There are various definitions of conceptual understanding, but because it has been proven to be useful for empirical studies (e.g. Post & Prediger, 2022), this paper uses the definition based on cognitive science models, which views knowledge as mentally organised in networks (Hiebert & Carpenter, 1992):

"A concept is understood if its mental representation is part of a network of representations. The degree of understanding is determined by the number and the strength of the connections. […] [It] is understood thoroughly if it is linked to existing networks with stronger or more numerous connections" (Hiebert & Carpenter, 1992, p. 67)

A deep understanding means being able to explain how an element of a concept is related to another representation (Renkl et al., 2013). Novices need to connect and integrate refined concept elements into a new concept before they can use them effectively (Aebli, 1994), while experts can handle compacted concepts, and unpack and explain their elements and connections. A concept element can have various forms, such as symbolic, visual or verbal, whereby language registers need to be explicitly connected and can counted as separate representation (Post & Prediger, 2022). To assess deep conceptual understanding, activities should capture these abilities: translating a concept between representations, explaining the meaning of concept elements and connections, and connecting concept elements in a wider network (Hiebert & Carpenter, 1992; Swan, 1985). For the acquisition of concepts, students' Grundvorstellungen (GV) (vom Hofe & Blum, 2016) is seen of crucial, because they describe mathematical concepts or procedures and their real-life interpretations. Three aspects are especially important:

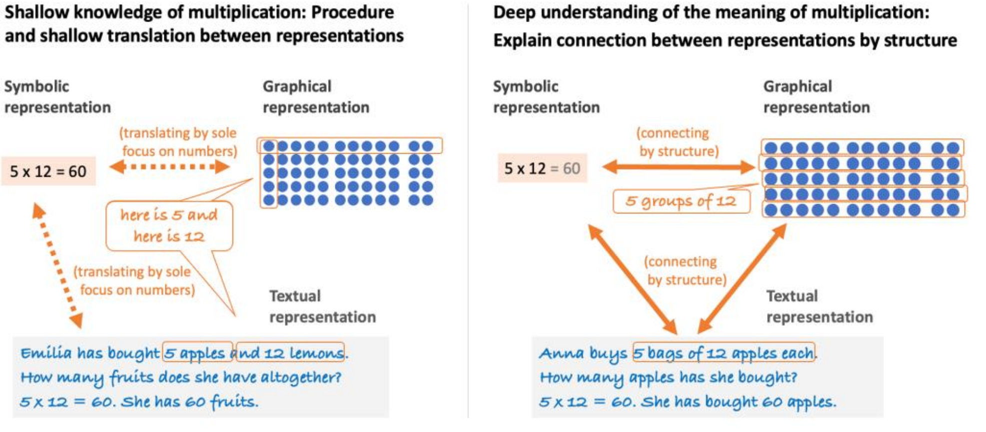
- "The *constitution of meaning* of a mathematical concept by linking it back to a familiar knowledge or experiences, or back to (mentally) represented actions,
- The *generation of a corresponding mental representation* of that concept; that is, an "internalization", which (following Piaget) enables operative action at the level of thought,
- The *ability to apply* a concept *to* real-life situations by recognizing a corresponding structure in subject-related contexts or by modelling a subject-related problem with the aid of mathematical structures." (vom Hofe & Blum, 2016, p. 230)

For example, the meaning of multiplication can be understood as repeated addition or, more relevantly for numeracy competences in each grade (Siemon, 2019), as *counting bundled units* (Götze & Baiker, 2021). While additive perceptions relate to (mental) actions that are repeated a certain number of times, unitizing (Lamon, 1994) refers to grouping object into equally sized groups and counting them. Students with a deep conceptual understanding can for instance explain the symbolic representation of 12 x 5 and a graphical representation like a dot array by articulating the unit structures, e.g., as "12 units of 5" (Götze & Baiker, 2021; Prediger et al., 2019). Students with a superficial conceptual understanding may perform the calculation or transfer the symbolic multiplication into a graphical representation but may not understand how the multiplicative structure is represented (Figure 1). This stresses the relevance of language for the development of conceptual understanding, making it even more pertinent to foster language-production also in assessments (Schoenfeld, 2017).

Questions Guiding the Exploratory Analysis

After explaining how the machine and a human develop understanding, the question is how the differences affect the assessment of conceptual understanding. Assessing responses to open-ended questions can be seen as a text-classification problem. There are various methods to categorize texts (Li et al., 2022). For example, Meyer et al. (2024) showed how they trained a transformer to evaluate students' e-mails and provide instant feedback. Chen et al. (2023) found that GPT can assess text quality, even in reference-free models. However, their study focuses on the evaluation of the coherence of a story line, not on conceptual understanding. Suresh et al. (2023) stated "that the underlying conceptual structure learned by LLMs is brittle" (p.722) and that unlike humans, whose conceptual structure is robust to differences in culture, language and method of estimation, LLMs' structures vary a lot. This paper explores the implications of these findings for conceptual understanding of mathematical concepts.

Figure 1
Shallow Knowledge and Deep Understanding of the Meaning of Multiplication (Hankeln et al., submitted)



Exemplary Reactions of ChatGPT-4 on the Meaning of Multiplication

Procedure of the Exploratory Study

I used a three-step procedure to investigate how ChatGPT-4, the most current and powerful version in February 2024, responds to different queries about the meaning of multiplication. Due to the context-sensitivity of the model, those observances are of course highly dependent on the respective prompts. These explorations are qualitative examples of the potentials and challenges, given ChatGPT's architecture. They aim at raising awareness that assessing conceptual understanding with a LLM is different from assessing text quality, such as coherence or structure.

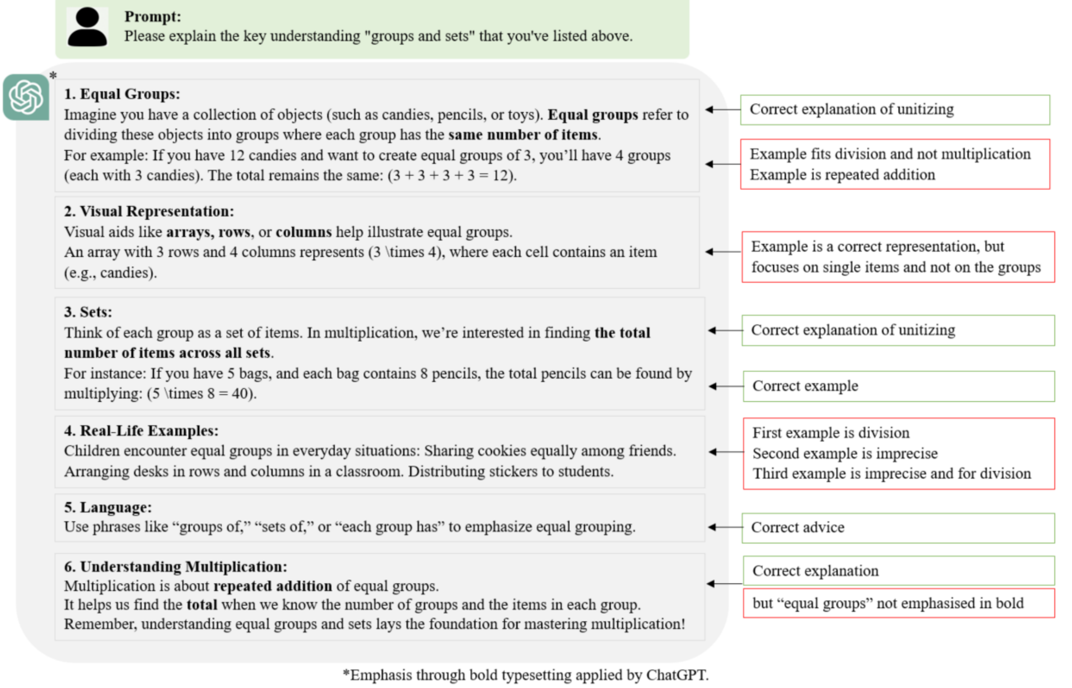
In the first phase, I asked ChatGPT, with different prompts being more or less specific about counting in groups, to explain the meaning of multiplication. As OpenAI does not make it transparent on what data exactly has been trained, it is a priori unclear how ChatGPT is familiar with didactical concepts. In the second phase, I gave ChatGPT 13 tasks on conceptual understanding of multiplication from the Mastering Math – Online Checks (Hankeln et al., submitted). I tested both text and image input for items with graphical representations. In the third phase, I asked ChatGPT to evaluate students' responses to the same tasks it worked on in the second phase. I saved and analysed ChatGPT's answers for problematic aspects that can be explained by its working mechanism. This study focuses on the output of ChatGPT to text responses.

Observations in the Exploratory Study

ChatGPT’s answers showed that it has encountered texts on multiplication as it can give detailed (and mostly accurate) explanations on it. In the exploratory discussions, ChatGPT initially showed a procedural focus on repeated addition. However, with chained prompts, ChatGPT could be guided to explain multiplicative thinking as counting in groups (Figure 2). ChatGPT provides an extensive list with several correct comments. However, its examples often do not match the focus on counting in groups.

Figure 2

ChatGPT’s Explanations on Groups and Sets in Multiplicative Thinking



*Emphasis through bold typesetting applied by ChatGPT.

The purpose of this analysis is not to assess ChatGPT's proficiency in explaining a concept, as fine-tuning or specific prompts can alter its performance. Instead, the goal is to highlight that its output lacks a foundation in solid, coherent knowledge, but relies on patterns learned from training data. This is evident in bullet point four, where ChatGPT associates "sharing" or "arranging in rows and columns" with the input "groups," a general correctness that overlooks the emphasis on multiplicative thinking. To address this word choice, explicit prompting or training would be necessary. In general, and apparent in Figure 2, ChatGPT had difficulties in choosing appropriate examples that not only fit to the context on the surface level, but also illustrate the targeted conceptual understanding.

The impact of transformer and attention mechanisms on ChatGPT's conceptual understanding capabilities is substantial. Among 13 items covering various multiplication representations, ChatGPT successfully solves three, with two relying on superficial strategies. For instance, in a task involving a chocolate bar with three rows of five segments, ChatGPT correctly combines the given numbers 3 and 5 to form a multiplication. The third task, creating a word problem for 5 x 6, aligns with ChatGPT's text generation capability, but does not guarantee deep conceptual understanding.

Regarding translations of non-text representations, ChatGPT exhibits behaviours typical of students with weak conceptual knowledge in ten items. For instance, when prompted several times, each time in a new chat, with "Given the multiplication task 2 times 6 equals 12. Which dice should be shown to match the multiplication?" ChatGPT provides varied answers with multiple errors (see Figure 3). In the first response, ChatGPT identifies a single dot as a unit, a misconception uncommon in students who, by the way, can physically manipulate dice and experience how the dots on one die belong together as one object, which suggests a using a die as a unit. Even when considering one die as a group, ChatGPT selects an incorrect representation, demonstrating a misalignment between its surface-level alignment with group counting and its lack of underlying conceptual understanding. The third answer, which chooses the correct representation, aligns well with the theoretical explanations in section 2. Apparently, the word "dice" was linked during processing with the find concept of "probabilities" where dice games are frequently used. However, the argumentation is again only on the first look convincing but lacks conceptual coherence.

Finally, ChatGPT was tasked with assessing the conceptual understanding of multiplication in students' responses ($n = 124$ students, grade 5 and 6). Iterative prompts designed for ChatGPT-3.5 demonstrated effective detection of correct positive answers (accuracy = 82.2%) (Hankeln, 2024). However, challenges arose when students employed vague expressions or attempted to convey correct ideas without using terms like "groups of," leading to frequent inaccuracies in judgment. Unlike the study by Latif and Zhai (2024), which focuses on a fine-tuned ChatGPT version for rating student answers with high accuracy, this study aims to elucidate challenges stemming from the operational mechanisms of mechanical text comprehension. Thus, the focus is on presenting detailed analyses of two specific responses that highlight the evident consequences of ChatGPT's context-sensitivity.

The example shown in Figure illustrates Grade 5 student Kate's robust conceptual understanding of multiplication, which becomes evident not only in this instance, but also in other tasks she undertook. In this particular item, Kate promptly discerns the discrepancy between the word problem and the symbolic term 2x5. She articulates the notion of grouping, ("get ONE thing twice") but utilizes her own terminology, emphasizing "ONE" to denote equal-sized groups. Kate identifies the surface strategy embedded in the problem, focusing on numbers rather than the operation. A subsequent discussion with Kate validates her reference to "thing" as groups, "one thing" signifying equal-sized groups, and "twice" indicating the repeated operation. Considering ChatGPT's mechanism, its struggle in accurately evaluating this response is unsurprising as Kate does not frequently occurring descriptions for a multiplication.

Figure 3

ChatGPT's Various Responses to Representing a Multiplication with Dice

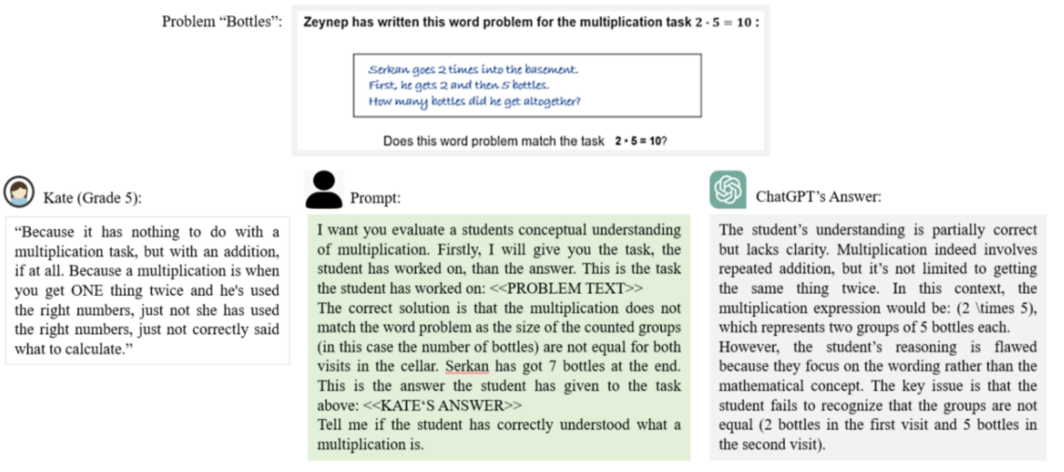
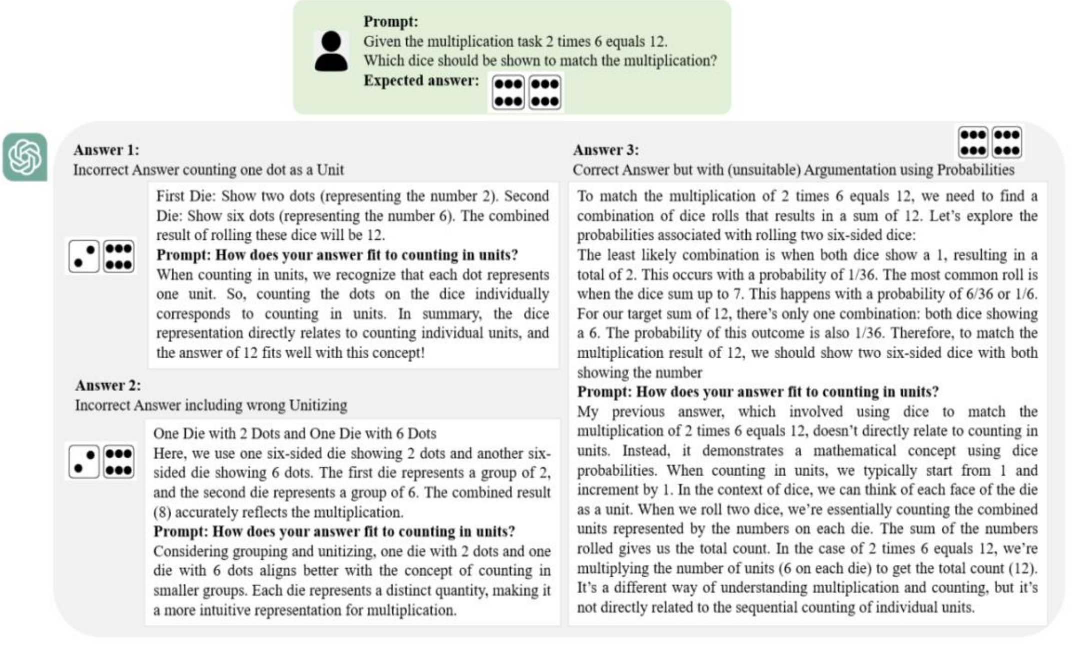


Figure 4

The Item "Bottles" and ChatGPT's Evaluation of a Student's Response



ChatGPT misinterprets the addition Kate mentions. While Kate suggests the word problem aligns better with the addition 2+5, ChatGPT insists on viewing multiplication as repeated addition. The local proximity of “multiplication” and “addition” in Kate's response might have led to this semantic misinterpretation due to ChatGPT’s context-sensitivity. Furthermore, ChatGPT takes "getting the same thing twice" literally, other than Kate intended. ChatGPT's abstract characterization claiming Kate focuses on wording rather than the mathematical concept, is inaccurate. Indeed, the opposite is true. Kate does not focus on the exact wording, but she tries to convey her understanding of the operation.

To gauge the impact of response wording, I replaced "thing" with "group" in Kate’s response and presented the prompt to ChatGPT again (in a new chat). This minor alteration substantially influenced the evaluation, as in this version the wording is closer to the expected answer. The revised evaluation by ChatGPT is more apt for fostering Kate's language proficiency:

> The student argues that multiplication involves getting one group twice, which is partially correct. However, the key aspect of multiplication is that the group size remains the same for each repetition. In this case, the group size (number of bottles) is different for each visit, so it cannot be a multiplication problem.

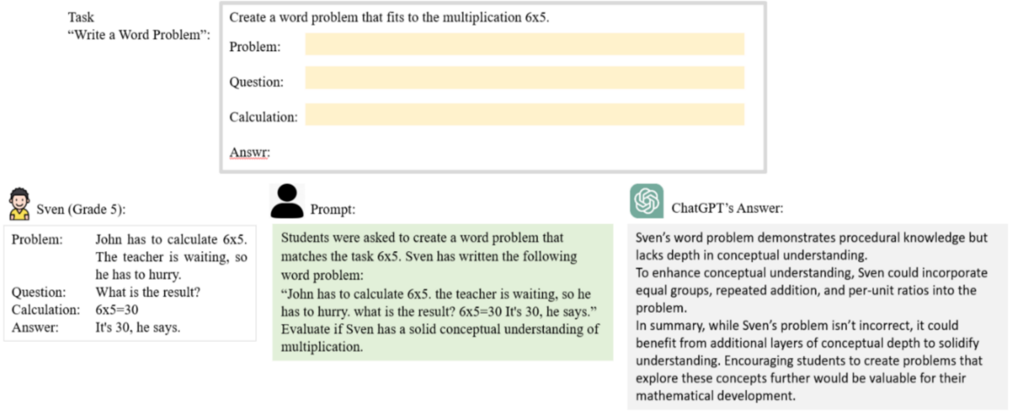
This alteration again illustrates the consequences of ChatGPT's context-sensitivity for the automatic assessment of students’ responses. To effectively evaluate such a rather vague open response, a language model must exhibit flexibility when students employ varied registers and express their thoughts in unexpected ways or/and be perhaps more guided through this process, for example by decision-trees (Hankeln, 2024).

As mentioned earlier, ChatGPT demonstrated the ability to generate a word problem correctly corresponding to the multiplication 6x5. I presented ChatGPT with an authentic response from a Grade 5 student who deviated from the intended approach, opting for a real-life scenario instead of a multiplicative context (Figure 5). In this instance, it becomes apparent that the student does not reveal a firm conceptual understanding but chooses a superficial way-out of the challenge to construct a multiplicative situation. ChatGPT adeptly identifies this structural flaw in the response and provides valuable feedback for improvement.

The examples illustrating the evaluation of students' responses suggest that ChatGPT, particularly when not explicitly pre-trained on students' language and conceptual understandings, adeptly distinguishes whether a response encompasses a description of conceptual understanding or not. However, evaluating the level of conceptual understanding is of higher difficulty and needs a more careful framework to focus the corresponding prompts accordingly.

Figure 5

ChatGPT's Evaluation of a Word Problem Written by a Grade 5 Student



Comparison of Machine and Human Conceptual Understanding

The theoretical discussions in sections 2 and 3 delineated parallels and distinctions between machine and human understanding. In both realms, networks of concepts are pivotal, with understanding manifesting as the ability to correlate verbal elements (words or phrases) with concepts, involving information encoding. The conceptual understanding definition proposed by Hiebert and Carpenter (1992), portraying a robust network of representations, bears resemblance to the foundational idea of neuronal networks underlying LLMs. However, a crucial disparity lies in the current robustness distinction: LLM representations rely on broader linguistic context (Suresh et al., 2023), but lack a cohesive conceptual core, unlike GVs in human understanding, which remain relatively independent of contextual variations or tasks. This distinction is for example evident in ChatGPT's recurrent struggles in this study to represent multiplication with dice.

As elucidated earlier, the emergence of context-sensitivity has been a turning-point for language models (Li et al., 2022), enabling human-like adaptive conversations attuned to subtle wording nuances. This sensitivity arises from the transformer architecture, where each word vector depends on surrounding text or prior prompts. However, this architecture introduces ambiguity regarding whether a word holds meaning outside the given context (Suresh et al., 2023). This contrasts with the definition of conceptual understanding or GV in mathematics education, where representations are linked to actions extendable to new situations. This became evident in the example of the dice, which ChatGPT linked its output to different contexts, including probabilities for no other reason than frequent co-occurrence of dice and probabilities in the text of the training data, lacking content-related reasoning. Humans, in contrast, draw semantic knowledge not only from language but also from visual, tactile, and auditory inputs (Shapiro, 2019).

Efforts to incorporate embodied experiences into language models (McGregor, 2023; Xiang et al., 2023) are currently explored.

Despite the crucial role of context-sensitivity in facilitating human-like conversations, it poses challenges for attaining human-like conceptual representations. This complexity is reflected in the difficulties encountered by LLMs in problem-solving tasks requiring conceptual understanding, as exemplified in the earlier-discussed cases. The machine's tendency to react intensively to context variations can lead to illogical, non-coherent responses or 'hallucinations' (Wu et al., 2023). As the machine reacts intensively to variations in context, formulations that for a human understanding share the same core, appear unrelated for the machine, leading to divergent outcomes. The behaviour of basing responses solely on contextual closeness resembles strategies from students with a superficial conceptual understanding: As they cannot draw upon appropriate representations, they infer what they have to do from signal words in the problem text. Additionally, it remains a black box what exactly is happening when ChatGPT processes a prompt.

The essential differences between human and machine understandings pose challenges in utilizing LLMs like ChatGPT for educational purposes. On the outside, it seems as if the machine's algorithm is employing strategies in problem-solving, we seek to replace in students with more profound conceptual knowledge. Leveraging ChatGPT, especially in a non-specifically fine-tuned version, as a model or instructor for students may thus risk reinforcing non-viable strategies.

Additionally, in assessing students' open-ended responses, challenges emerge: As Post and Prediger (2022) state, different language registers can be seen as different representations and need to explicitly be connected. Additionally, when building up knowledge, students encompass intermediate states of transitions with individual, but typical conceptions (Prediger, 2008). This implies that students' responses typically show variations, imprecision or vagueness containing more or less normatively intended representations. I have shown one example where ChatGPT is not (yet) able to extract the core understanding of multiplication the student Kate had, because Kate was not able to express her ideas unambiguously. During learning processes, these intermediate states, especially concerning the explanations of operations, are quite common, so this problem is not neglectable. It can be assumed that ChatGPT has not been trained on this kind of texts, students produce when communicating mathematically. The explanations have also shown that giving cue words (like the word "group") can change and improve the outcome. Further hope lies in Retrieval Augmented Generative AI (Siriwardhana et al., 2023) where additional information can be given to the GPT as knowledge base for the answers.

In advancing the potential of ChatGPT for mathematics education, explicit training on texts reflecting students' varying stages of conceptual understanding could be pivotal. Such an LLM, with nuanced context-sensitivity could capture students' conceptual understanding, aligning with Schoenfeld's (2017) call to enhance open questions in assessments. The last example, however, gave insight into the potential a LLM could have, as the extraction of textual features and the structure of texts is already better feasible.

Conclusion

This paper has contrasted ChatGPT's understanding with conceptual understanding in mathematics education, highlighting their structural network similarities yet substantial differences. The analysis reveals that many reported study outcomes align with the logic of the underlying transformer mechanism. The examination of ChatGPT's suitability for open-ended task assessment exposes the greater challenge in capturing conceptual understanding compared to structural text elements. Students' language use, particularly imprecision, significantly influences ChatGPT's output. Consequently, further research in this domain is imperative to refine real-time evaluation of open-ended items, particularly focusing on conceptual understanding.

Several limitations have to be discussed for this analysis: Firstly, the available models are progressing rapidly, and it is possible that in near future the lacking robustness in understanding as this study has demonstrated is compensated for by more training. Fine-tuning a machine for specific tasks shows that this is possible, but expensive in time and the need for labelled data. Secondly, this study has focused on the understanding of multiplication. It can be assumed that the quality of the assessments differs strongly between different topics and more specialised topics could be more challenging than more basic topics. And thirdly, students' responses differ depending on the specific task and age group so that also these factors could bring more or less variance for example in the linguistic registers or expression in the responses.

References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Briganti, G. (2024). How ChatGPT works: A mini review. *European Archives of Oto-Rhino-Laryngology*, *281*(3), 1565–1569. https://doi.org/10.1007/s00405-023-08337-7

Chen, Y., Wang, R., Jiang, H., Shi, S., & Xu, R. (2023). *Exploring the use of large language models for reference-free text quality evaluation: An empirical study*. https://doi.org/10.48550/ARXIV.2304.00723

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Götze, D., & Baiker, A. (2021). Language-responsive support for multiplicative thinking as unitizing: Results of an intervention study in the second grade. *ZDM – Mathematics Education*, *53*(2), Article 2. https://doi.org/10.1007/s11858-020-01206-1

Hankeln, C. (2024). *Open-ended items in digital formative assessments: Decision Trees as (AI-compatible) approach to reliably code students' understanding?* FAME, Utrecht.

Hankeln, C., Kroehne, U., Voss, L., Gross, S., & Prediger, S. (submitted). *Developing digital formative assessment for deep conceptual learning goals: Which topic-specific research gaps need to be closed?*

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). Macmillan.

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Lamon, S. J. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 89–122). State University of New York Press.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM transactions on intelligent systems and technology*, *13*(2), 1–41. https://doi.org/10.1145/3495162

McGregor, S. (2023). Is ChatGPT really disembodied? *The 2023 Conference on Artificial Life*. The 2023 Conference on Artificial Life. https://doi.org/10.1162/isal_a_00597

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

OpenAI. (2023). *GPT-4 Technical report*. https://cdn.openai.com/papers/gpt-4.pdf

Post, M., & Prediger, S. (2022). Teaching practices for unfolding information and connecting multiple representations: The case of conditional probability information. *Mathematics Education Research Journal*. https://doi.org/10.1007/s13394-022-00431-z

Prediger, S. (2008). The relevance of didactic categories for analysing obstacles in conceptual change: Revisiting the case of multiplication of fractions. *Learning and Instruction*, *18*(1), Article 1. https://doi.org/10.1016/j.learninstruc.2006.08.001

Prediger, S., Fischer, C., Selter, C., & Schöber, C. (2019). Combining material- and community-based implementation strategies for scaling up: The case of supporting low-achieving middle school students. *Educational Studies in Mathematics*, *102*(3), Article 3. https://doi.org/10.1007/s10649-018-9835-2

Schoenfeld, A. H. (2017). On learning and assessment. *Assessment in Education: Principles, Policy & Practice*, *24*(3), 369–378. https://doi.org/10.1080/0969594X.2017.1336986

Shapiro, L. (2019). *Embodied Cognition* (2nd ed.). Routledge. https://doi.org/10.4324/9781315180380

Siemon, D. (2019). Knowing and building on what students know: The case of multiplicative thinking. In D. Siemon, T. Barkatsas, & R. Seah (Eds.), *Researching and Using Progressions (Trajectories) in Mathematics Education* (pp. 6–31). Brill.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, *11*, 1–17. https://doi.org/10.1162/tacl_a_00530

Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L., & Rogers, T. T. (2023). Conceptual structure coheres in human cognition but not in large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 722–738. https://doi.org/10.48550/ARXIV.2304.02754

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems*. https://arxiv.org/pdf/1706.03762.pdf

vom Hofe, R., & Blum, W. (2016). "Grundvorstellungen" as a category of subject-matter didactics. *Journal Für Mathematikdidaktik*, *37*(S1), Article S1. https://doi.org/10.1007/s13138-016-0107-3

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., & Hu, Z. (2023). *Language models meet world models: Embodied experiences enhance language models*. https://doi.org/10.48550/ARXIV.2305.10626

Zhai, C. (2009). *Statistical language models for information retrieval*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02130-5

Author Note:

The Mastering Math Online-Check is developed and investigated within the projects SchuMaS (funded by the National Ministry of Education and Research, grant No. SchumaS-SMS2101L-01PR2101C to S. Prediger) and MaCo (funded by the Departments of Education, Hamburg and Hessia).
There is no conflict of interests to disclose.

Correspondence concerning this article should be addressed to Corinna Hankeln, Insitut für Entwicklung und Erforschung des Mathematikunterrichts, TU Dortmund University, Vogelpothweg 87, 44227 Dortmund, Germany, E-Mail: Corinna.hankeln@math.tu-dortmund.de