AI-supported Mathematical Task Design with a GPT Agent Network

Conference Paper · October 2024

CITATIONS
0

READS
30

2 authors:

Franziska Peters
University of Hamburg
18 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Sebastian Schorcht
Technische Universität Dresden
37 PUBLICATIONS **118** CITATIONS
SEE PROFILE

MEDA

Mathematics Education in the Digital Age

Proceedings
of the 17th ERME Topic Conference
MEDA 4

University of Bari Aldo Moro, Italy
3-6 September 2024

Edited by
Eleonora Faggiano
Alison Clark-Wilson
Michal Tabach
Hans-Georg Weigand

erme european society for research in mathematics education

UniBa Università degli Studi di Bari Aldo Moro

Organising Committee

Hans-Georg Weigand (Germany) – Chair
Alison Clark-Wilson (UK)
Eleonora Faggiano (Italy)
Michal Tabach (Israel)

International Program Committee

Hans-Georg Weigand (Germany) – Chair
Bärbel Barzel (Germany)
Rogier Bos (The Netherlands)
Roberto Capone (Italy)
Eirini Geraniou (UK/Greece) – member of the ERME board
Luca Lamanna (Italy) – YR Representative of YERME
Janka Medová (Slovakia)
Ornella Robutti (Italy) – leader of TWG 16 at CERME 13
Helena Rocha (Portugal)
Osama Swidan (Israel) – leader of TWG 15 at CERME 13
Jana Trgalova (Switzerland/France) – member of the ERME board
Melih Turgut (Norway/Turkey)

Local Organisers

Eleonora Faggiano – Chair
Roberto Capone – Co-chair
Maria Lucia Bernardi
Antonio Leserri
Ida Maiellaro
Federica Troilo

Conference webpage: https://www.dm.uniba.it/meda4

Mathematics Education in Digital Age
Proceedings of the 17th ERME Topic Conference MEDA 4

held on 3 – 6 September 2024 in Bari, Italy

Editors: Eleonora Faggiano, Alison Clark-Wilson, Michal Tabach, Hans-Georg Weigand
Publisher: University of Bari Aldo Moro
Place: Bari (Italy)
Year: 2024
ISBN: 978-88-6629-080-3

All contributions were peer-reviewed.

This work was supported by the "National Group for Algebraic and Geometric Structures, and their Applications" (GNSAGA - INDAM).

AI-supported Mathematical Task Design with a GPT Agent Network

Franziska Peters[1] and Sebastian Schorcht[2]

[1] University of Hamburg, Germany; franziska.peters-2@uni-hamburg.de

[2] University of Dresden, Germany; sebastian.schorcht@tu-dresden.de

This study investigates the use of communicative AI agents in designing mathematical tasks. It examines how a network of LLM (large language models) agents can facilitate mathematical task design through collaborative communication in a chat chain. Four specialized AI agents were instructed each focusing on a different perspective: mathematical content, linguistic sensitivity, competence orientation, and differentiation. The AI agents sequentially modified given mathematical tasks, with each contributing a unique focus to the task's evolution. The resulting tasks were evaluated by in-service teachers as human experts. This way, the qualitative study explores the potential of LLM agent networks in educational contexts. First findings suggest that AI agents can support teachers in the development of mathematical tasks for diverse learning needs, but at the same time require adaptation by teachers to the educational situation.

Keywords: Task Design, ChatGPT, AI agent, Large Language Models, problem posing.

Introduction

Large Language Models (LLMs) have gained high attention in the educational research field (e.g. Kasneci et al., 2023; Buchholtz et al. 2023). The disruptive changes and developments in the educational context are still not fully accessible. The aim of the language modeling (LM) approach is to teach machines human language and its characteristics (Hadi et al., 2023) based on statistical calculations (Hiemstra, 2009). While so-called statistical language models make a probability statement regarding the following word within texts, LLMs are based on artificial intelligence (AI) and deep learning techniques (Hadi et al., 2023). A currently much-discussed example of LLMs is the GPT (Generative Pretrained Transformer) architecture, which can imitate the structure of human language and respond appropriately to requests. To achieve this, the LLM is trained to solve specific tasks using large amounts of data. Within this data, the LLM architecture recognizes and analyzes patterns and relationships to generate a coherent, context-dependent output (Floridi & Chiriatti, 2020).

Since the release of access to GPTs by OpenAI (2023), the technology of LLM agents has become accessible to all and can generate customized responses to specified requirements. LLM agents have demonstrated considerable success across a broad spectrum of applications, from reasoning (Yao et al., 2023) to video gaming (Wang et al., 2023) and autopilot systems (Jin et al., 2023). Recent studies have explored the potential of leveraging multiple LLM agents working in collaboration to address a single query, showcasing effectiveness in tackling intricate tasks (Du et al., 2023; Liang et al., 2023; Wu et al., 2023). The integration of text files and their underlying content also allows LLM agents to provide advanced responses. Therefore, research in the field of mathematical task design presents an exciting opportunity for exploring connections with LLM agents.

Theoretical Background

In the past, task design has often been carried out by textbook authors, while new relevant literature emphasizes the significance of involving mathematics teachers as partners in this process (Jones & Pepin, 2016). Even when tasks are not designed from scratch but adapted, this process supports the development of the teacher's mathematical knowledge and mathematics-didactical design capacity (Pepin, 2015).

As tasks play an important role in planning competence-oriented teaching, teachers also need to have a clear understanding of the cognitive demands of the instructional and diagnostic assignments they are planning to use in class (Maier et al., 2014). To plan competence-oriented lessons teachers must know the domain-specific competence development of their learning group and must analyze tasks regarding their suitability for this current competence development (Kleinknecht & Lankes, 2012).

Because of the diversity of learning groups, the ability to design and/or modify mathematical tasks of varying difficulty becomes more important (Maier et al, 2014). Task design can be done by individuals but is also shown to be constructive when realized in collaborative groups. In addition to designing tasks in collaborative groups, Sensevy et al. (2013) created the term ‘cooperative engineering’ between teachers and researchers to design tasks. If those teachers and researchers act out of different professional mandates and responsibilities, they form a ‘multi-professional team’ (summarized in Widmer-Wolf, 2018) which holds the potential to meet the individual requirements of learners. Combining the strengths of a multi-professional team in specialized LLM agents could therefore be promoted as a further step in task design.

Research Interest

To this end, we take up the ‘communicative agents’ approach of Qian et al. (2023), which in the project ChatDev has various AI agents with specific roles and tasks for software development and access to GPT. In our study, we used the LLM GPT-4 from OpenAI and built a so-called GPT agent network consisting of four AI agents with different functions and expertise: a language-sensitive agent, a mathematical content-oriented agent, a competence-oriented agent, and a differentiation-sensitive agent.

The AI agents are assigned to different perspectives and thus set different priorities in the process of task design. This enables interdisciplinary collaboration in the optimizing and adapting of tasks and can be a useful support for teachers and experts in mathematical task design, especially regarding the individual learning requirements of students.

In a qualitative study, the effects of AI agents in mathematical task design are to be investigated aiming at the following research question: To what extent can a GPT agent network working as a multi-professional team support teachers within mathematical task design in four perspectives? We are planning to answer two subordinate questions:

1. Between an original and an AI-modified task, which do teachers prefer to select for their educational context, and what are the reasons for their choice?
2. How do the teachers evaluate the tasks modified by AI agents regarding the four perspectives?

Methods and Materials

To explore those research questions, six tasks with varying levels of difficulty were selected. Four AI agents were created for the GPT agent network, each given access to a PDF document containing guidelines (referred to as the research knowledge base) and instructed to adapt the tasks based on these guidelines. A chat chain mechanism was established to facilitate communication between the AI agents, as proposed by Qian et al. (2023). Following the modifications to the tasks using a consensus-driven, AI-supported design process, the modified tasks were evaluated by in-service teachers who served as human experts in this context.

Mathematical tasks

In choosing the examples to modify, three simple tasks requiring basic competencies at the elementary and secondary levels were considered, as well as three further examples from the domain of problem solving. Problem-solving tasks present a unique challenge for LLMs as they often extend beyond basic calculations and encompass argumentative elements necessary for solving the problem (Schorcht et al., 2024). The first three tasks originate from the areas of arithmetic, patterns, and fraction calculation. For the latter three tasks, problems were selected that require basic arithmetic skills as well as the solving of a system of linear equations. This aims to explore the possibilities of task development with AI agents and demonstrate to what extent the AI agents are capable of enriching both simple and more complex tasks. The tasks presented were always given to the AI agents with a corresponding learning objective and grade level. This ensures the development of the tasks within a defined framework. These tasks were inserted into the chat chain, starting with the mathematical content-oriented AI agent followed by the language-sensitive AI agent, differentiation-sensitive AI agent, and at least the competence-oriented AI agent.

Enhanced AI agents

Documents were developed as foundational materials for each AI agent to serve as a research knowledge base. These documents were provided in PDF format and enabled each AI agent to access relevant information and adapt tasks based on their specific knowledge. Restricting the research knowledge base to a small number of pages is essential due to GPT-4's processing limitation of up to 8192 tokens per prompt. A token might represent an entire word, though fragments of words and punctuation marks are also considered tokens.

Instructions were designed to serve as a functional guide for AI agents in task evaluation using the specified research knowledge base. For example, the role of AI agents was clarified initially, highlighting that tasks should be viewed as objects of study rather than prompts for solutions. Each AI agent approaches the task analysis differently, focusing on areas like mathematical content, linguistic aspects, differentiation, and compliance with content and process standards, as outlined in the research knowledge base.

The mathematical content-oriented AI agent is designed to support teachers in generating a mathematically correct and consistent task. Unlike the others, this agent uses the training data from GPT-4 as its research knowledge base, as it is impossible to limit mathematical knowledge to just a few pages.

The language-sensitive AI agent is designed to support teachers in developing language-sensitive tasks based on knowledge of language barriers in mathematics lessons and design principles of language-sensitive tasks (Abshagen, 2015).

The differentiation-sensitive AI agent is designed to support teachers in natural differentiation and refers to characteristics of naturally differentiated learning opportunities and tasks: openness, complexity, low entry threshold, “ramps” for high achievers, need for discussion and high cognitive activation potential (Krauthausen & Scherer, 2022).

The competence-oriented AI agent is designed to support teachers in encouraging necessary mathematical competencies and refers to the five content-related competencies according to UNESCO (2020) as well as to the NCTM Process Standards (n.d.).

After analysis, a decision is made for potential task modifications, which the AI agent autonomously implements if necessary. Feedback is then provided on the task development, assessing adherence to the guidelines. The process concludes with a summary that reformulates the task's nature as a subject for study, aiming to enhance prompt adherence and minimize hallucinated responses by the LLM, addressing the common issues of prompt following in prompt engineering (Rassin, Ravfogel & Goldberg, 2022; Betker et al., 2023).

After generating instructions for the AI agents, we followed the proposed architecture by Qian et al. (2023) and divided each phase of the task design process into atomic chats. The AI agents sequentially modify the given task, with each AI agent contributing a unique focus to the task design. To enhance communication between AI agents, the output from each agent was used as input for the next prompt. This sequence of prompt, output, prompt, etc., is referred to as a chat chain (Qian et al., 2023). The chat chain thus realizes a communicative process that culminates in a circular, consensus-oriented procedure under the independent influence of certain guidelines, to evaluate tasks and make suggestions for modifications.

Data Analysis

In the second step, the resulting tasks were evaluated by in-service teachers as human experts (each with an average of 11.67 years of teaching mathematics’ experience. The teachers were asked to critically evaluate the adapted tasks to test their practical applicability and relevance in educational settings. This is a process crucial for ensuring the tasks' practical applicability and relevance in classroom settings. Maier et al. (2014) described seven categories and characteristics of interdisciplinary task analysis and showed how teachers can analyze tasks in terms of their suitability for competence-oriented teaching.

Similarly, in this study, four categories of mathematic-specific task analysis were used to guide teachers in evaluating the given tasks. The evaluation process required teachers to consider several criteria, including mathematical content, language sensitivity, differentiation level, and competence requirements. The aim was to assess the AI's effectiveness in customizing educational content to meet diverse learning needs and environments.

Following this we used a frequency analysis of the teacher’s choice of tasks as well as their evaluation based on the four given perspectives. Furthermore, the comments reasoning their decisions are analyzed using Qualitative Content Analysis to inductively develop categories.

Results

In this paper, we want to provide an insight into the results of our study. The extensive analysis of our data can be found in Schorcht et al. (forthcoming; 2024).

RQ1: Overall, 44% of 36 decisions leaned towards the selection of AI-modified tasks with notable differences between problem-based tasks and tasks requiring basic competencies. Specifically, 61% of the decisions concerning problem-based tasks favored the AI-modified task, indicating a recognition of the value added by AI in these contexts. However, the changes made by AI agents to tasks requiring basic competencies were less convincing, with only 28% of experts favoring the AI-modified tasks. This outlines a noticeable variance in expert choices, highlighting the nuanced impact of AI integration depending on the nature of the original task.

The reasons teachers gave for their decisions were analyzed through the Qualitative Content Analysis. The sorting of the paraphrased comments from the experts within the assessments led to categories of positive and negative comments on specific categories for the original tasks and the AI-modified tasks that explain the choices of the experts. Due to the limited space in this conference paper, only a few of the 21 categories are used for each task type to exemplary outline the reasoning of the decisions. A complete description of all categories as well as a table containing their absolute frequencies is presented in Schorcht et al. (forthcoming; 2024).

Regarding the original tasks' positive categories, comments of the category ‘short text’ were used to justify selecting the original task as it was “more concise” and had “less information that overwhelmed”. One expert mentioned students’ reading competence: “I think that many will initially have difficulties in selecting the essential from the texts. Therefore, the simple tasks with less text are more effective in achieving the formulated objectives” (Expert 3). Concerning original tasks' negative categories, the category ‘lack of content’ was evaluated negatively due to the incomplete nature of tasks that require basic competencies.

Positive characteristics of the AI-modified tasks were addressed in eight categories. The following paraphrased comment addresses the categories of ‘language comprehension’, ‘concrete call for action’ and ‘motivating’ and ‘solution approaches given’ in which AI-modified tasks scored very well: “Task 2 [AI-supported task; authors] is formulated in a more motivating and comprehensible way because it requires active action. There is also an approach to a solution” (Expert 1). The range of different strategies offered was also used to explain the selection of an AI-modified task: “Different calculation strategies can be used and described here” (Expert 5). However, we also found considerable criticism of AI-modified tasks, particularly regarding the inclusion of ‘unnecessary information’ and the ‘demanding text level’.

RQ2: All six AI-modified tasks were assessed multidimensionally in terms of mathematical depth, language sensitivity, natural differentiation and competence orientation. The tasks were rated separately but the results don’t seem to be significantly dependent on the different tasks.

Overall, all AI-supported modified task were rated positively in terms of mathematical content, differentiation level, and competence requirements. The only category not evaluated positively was the language sensitivity. Over 40% of the ratings stated, that the tasks were not, not completely or only partial language sensitive. The added comments show that this is due to the fact that most of the modified tasks are longer than the original one as ChatGPT is a language-generating AI and tends to

generate long texts. This should be considered in further development of the language sensitive AI agent.

Discussion

The presented results give useful information about the quality of the modified tasks and their suitability for potential use in classrooms. The evaluation regarding the four perspectives, i.e. the work of the four different agents, also serves as feedback regarding necessary further development of the AI agents. The findings suggest that GPT networks functioning as a multi-professional team can indeed support teachers in developing mathematical tasks and open new perspectives for pedagogical strategies. The AI-modified tasks were especially effective in providing supportive hints, improving language comprehension, and giving clear calls to action. However, it can also be seen that some of the modified tasks are very text-expanding since LLMs are designed for text production. Thus, it can be concluded that the instructions of the language-sensitive AI agent need to be adapted regarding text length. Nevertheless, this tendency to hallucinate also provided an opportunity to enrich tasks requiring basic competencies with context.

In Summary, a GPT network is considered a practical tool for modifying tasks, but teachers are still responsible for adapting tasks to their students.

References

Abshagen, M. (2015). *Praxishandbuch Sprachbildung Mathematik: Sprachsensibel unterrichten – Sprache fördern.* Ernst Klett Sprachen GmbH.

Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Schorcht, S. & Pohl, M. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative KI-Sprachmodelle. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 114*, 19–26.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y. & Ramesh, A. (2023). *Improving Image Generation with Better Captions.* https://cdn.openai.com/papers/dall-e-3.pdf

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate.* https://doi.org/10.48550/arXiv.2305.14325

Floridi, L. & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines, 30*, 681–694.

Hadi, M. U., Al-Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M., Akhtar, N., Wu, J. & Mirjalili, S. (2023). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. 10.36227/techrxiv.23589741.

Hiemstra, D. (2009). Language Models. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems*. Springer. https://doi.org/10.1007/978-0-387-39940-9_923

Jin, Y., Shen, X., Peng, H., Liu, X., Qin, J., Li, J., Xie, J., Gao, P., Zhou, G., & Gong, J. (2023). *SurrealDriver: Designing generative driver agent simulation framework in urban contexts based on large language model*. https://doi.org/10.48550/arXiv.2309.13193

Jones, K., Pepin, B. (2016). Research on mathematics teachers as partners in task design. *Journal of Mathematics Teacher Education, 19*, 105–121. https://doi.org/10.1007/s10857-016-9345-z

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet,

O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. & Kasneci, G. (2023). *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*. https://doi.org/10.35542/osf.io/5er8f

Kleinknecht, M. & Lankes, E.-M. (2012). Kompetenzvermittlung im Unterricht: Eine neue Lern- und Aufgabenkultur an der Schule etablieren. *Schulleitung und Schulentwicklung, 57* (2), 1–16.

Krauthausen, G. & Scherer, P. (2022). *Natürliche Differenzierung im Mathematikunterricht.* Seelze: Kallmeyer.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). *Encouraging divergent thinking in large language models through multi-agent debate*. https://doi.org/10.48550/arXiv.2305.19118

Maier, U., Bohl, T., Drüke-Noe, C., Hoppe, H., Kleinknecht, M. & Metz, K. (2014). Das kognitive Anforderungsniveau von Aufgaben analysieren und modifizieren können: Eine wichtige Fähigkeit von Lehrkräften bei der Planung eines kompetenzorientierten Unterrichts. *Beiträge zur Lehrerinnen- und Lehrerbildung. 32*(3), 340-358.

NCTM. (n.d.) *Principles, Standards, and Expectations.* https://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Principles,-Standards,-and-Expectations/

OpenAI (2023). *GPT-4 Technical Report*. https://arxiv.org/pdf/2303.08774.pdf

Pepin, B. (2015). *Enhancing mathematics/STEM education: A 'resourceful' approach*. Inaugural lecture, 27 November 2015, Technische Universiteit Eindhoven.

Qian, C., Cong, X., Liu, W., Yang, C., Chen, W., Su, Y., Dang, Y., Li, J., Xu, J., Li, D., Liu, Z., & Sun, M. (2023). *Communicative Agents for Software Development.* https://doi.org/10.48550/arXiv.2307.07924

Rassin, R., Ravfogel, S. & Goldberg, Y. (2022). *DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models*. https://doi.org/10.48550/arXiv.2210.10606

Rezat, S. & Sträßer, R. (2012). From the didactical triangle to the socio-didactical tetrahedron: artifacts as fundamental constituents of the didactical situation. *ZDM International Journal on Mathematics Education, 44*, 641–651. https://doi.org/10.1007/s11858-012-0448-4

Sensevy, G., Forest, D., Quilio, S., & Morales, G. (2013). Cooperative engineering as a specific design-based research. *ZDM International Journal on Mathematics Education, 45*(7), 1031–1043.

Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). *Prompt the Problem – Investigating the Mathematics Educational Quality of AI-supported Problem-Solving by facilitating various Prompt Techniques*. Frontiers in Education, 9.

Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F. & Pohl, M. (2023). Ask Smart to Get Smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 115*, 12–24.

Schorcht, S., Peters, F. & Kriegel, J. (forthcoming, 2024). Communicative AI Agents in Mathematical Task Design – A Qualitative Study of GPT Network acting as a Multi-Professional Team. In B. Pepin, N. Buchholtz & U. Salinas- Hérnandez (Hrsg.) *Digital Experiences in Mathematics Education.* Springer.

UNESCO. (2020) *Global Proficiency Framework for Mathematics. Grades 1 to 9.* https://www.edu-links.org/sites/default/files/media/file/GPF-Math-Final.pdf

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). Voyager: *An open-ended embodied agent with large language models*. https://arxiv.org/abs/2305.16291

Widmer-Wolf, P. (2018). Kooperation in multiprofessionellen Teams an inklusiven Schulen. In Sturm, T., & Wagner-Willi, M. (Eds.), *Handbuch schulische Inklusion* (298–313). Verlag Barbara Budrich.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., & Wang, C. (2023). *AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework.* https://doi.org/10.48550/arXiv.2308.08155

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2023). REACT: Synergizing reasoning and acting in language models. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X