Differentiated Tasks by ChatGPT for Secondary Computer Science Education: Useful or not?

Tobias Bahr
bahr@ife.uni-stuttgart.de
University of Stuttgart
Germany

Mario Manzocco
mario.manzocco@stud.ph-ludwigsburg.de
Ludwigsburg University of Education
Germany

Dennis Schuster
dennis.schuster@stud.ph-ludwigsburg.de
Ludwigsburg University of Education
Germany

ABSTRACT

In recent years, there has been a growing interest in exploring the capabilities of AI chatbots, such as *ChatGPT*. Studies have investigated diverse applications, including the response of AI chatbots to undergraduate exam questions and the generation of student exercises for programming. However, the question remains if AI chatbots provide adequate results for K-12 CS in different application scenarios. AI chatbots are increasingly integrated into K-12 education by both students and teachers. In this context, a tool using didactical parameters was created to differentiate tasks with *ChatGPT-4* in an ongoing project. Preliminary findings from this work in progress reveal that teachers see a benefit using the tool. Future directions for using the tool are discussed.

CCS CONCEPTS

• **Social and professional topics → K-12 education**.

KEYWORDS

Computer Science Education; AI chatbots; ChatGPT; Expert rating; K-12

ACM Reference Format:
Tobias Bahr, Mario Manzocco, and Dennis Schuster. 2024. Differentiated Tasks by ChatGPT for Secondary Computer Science Education: Useful or not?. In *The 19th WiPSCE Conference on Primary and Secondary Computing Education Research (WiPSCE '24), September 16–18, 2024, Munich, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3677619.3678092

1 INTRODUCTION AND BACKGROUND

By the time *ChatGPT* was released in November 2022, large language models (LLMs) had become a topic of widespread discussion in the general public as a subset of artificial intelligence (AI) (cf. [11]). Developed by *OpenAI*, *ChatGPT* is a chat interface to the large language model *GPT* (short for Generative Pretrained Transformer), tapping into its extensive database [5]. This database enables the chatbot to generate relatively long texts. Quickly, *ChatGPT* and other AI chatbots were applied by educators to prepare lessons, generate tasks and more [13]. For Computer Science (CS), the use of LLMs to support programmers or educators is nothing new. Potentials (such as more efficient programming) and limits (such as factual false responses) have already been discussed [13]. However, with the frequent use of AI chatbots in educational settings since the release of *ChatGPT*, not much is known how teachers rate the quality and of ChatGPT responses [15]. As a means to reduce teachers' workload in differentiating tasks for different levels, a tool was programmed that uses *ChatGPT-4* to differentiate tasks based on different criteria. In an ongoing project, a first version of that tool was created [3], presented to educators (e.g. teachers) in a workshop, and then revised. Afterwards, an updated version was tested again in a one-day workshop with teachers from different backgrounds (e.g., involved in the creation of the CS curriculum). This led to the current version that will be presented in this work in progress study. When using the tool, the user can choose their subject (here CS) and then select the grade (for the context of this study grade 8 to 10 in lower secondary schools) and the core topic area of the CS curriculum e.g., algorithms [9]. A prompt is sent to search for the fitting facet in the CS curriculum. The user can then either verify that the found facet is correct or regenerate it. In the next step, the user can copy and paste the task that will be differentiated. The different criteria for differentiation can then be selected with the sliders. Overall, the following didactical parameters are encompassed in the tool: (1) CS curriculum from grade 7 to 12 for secondary schools in Baden-Württemberg, Germany (modifiable) [9], (2) Bloom's revised taxonomy (6-point rating scale from remembering to creating) [1], (3) level of complexity (5-point rating scale from one fact to one overarching concept) [7], (4) level of language complexity (5-point rating scale from very simple to very complex) [6, 8], (5) convergent or divergent solution [8, 12], and (6) real-world context (provided or not provided) [14].

In the initial phase of this study, all 75 tasks developed by CS experts for the CS curriculum of Baden-Württemberg for grade 8 to 10 [16] (considered as ground truth) were analyzed by the tool and categorized into one or multiple facets of the CS curriculum. Subsequently, the second and third author of this article reviewed the matching. According to the authors' assessment, 97.5% of the tasks created for the CS curriculum were accurately matched to facets of the CS curriculum. Disagreements among the authors arose only in cases where the association of clean code and code documentation with a task was assumed, even if not explicitly mentioned in the tasks. Therefore, an argument can be made that *ChatGPT-4*, when provided with prompts related to the CS curriculum, can accurately match tasks to the curriculum.

*This project was funded by the Professional School of Education Stuttgart-Ludwigsburg (PSE) and the Federal Ministry of Education and Research (BMBF) No: 01JA23M03C.

WiPSCE '24, September 16–18, 2024, Munich, Germany

ACM ISBN 979-8-4007-1005-6/24/09
https://doi.org/10.1145/3677619.3678092

2 METHOD

A workshop was conducted with 17 secondary teachers with different backgrounds. First the tool was demonstrated and afterwards each participant could use the tool to differentiate tasks. The last 20 minutes were used to get feedback from the participants to further develop the tool. Afterwards three CS teachers rated 38 differentiated tasks of existing 19 CS tasks from educational material for CS secondary education in Germany [16], whether the task described a meaningful problem (*sensibility*) and if it matched a possible solution approach (*readiness*), analogous to Sarsa et al. [13]. Participants could respond with yes, no, or maybe. Next, teachers were asked to rate the practical relevance of the task (on a 6-point likert scale from low to high) [2]. Finally, an open-ended question solicited comments or criticisms regarding the task.

3 RESULTS

The descriptive statistics show that 89.7% of the tasks were rated sensible by two out of the three teachers (one disagreed) and 48.7% of the tasks were sensible (all teachers agreed). 78.4% of the tasks could be directly used and 87.2% of the tasks have a medium to high practical relevance (4 or higher). Correlations between task complexity and practical relevance, as well as sensibility, were all found to be non-significant ($p > .05$). Hence, no indication of whether the tool performs better for easier or more complex tasks could be discerned. Qualitatively, it was observed that *ChatGPT* addressed students formally in seven tasks, which was deemed inappropriate for students aged 14 to 16. Four tasks provided no differentiation, maintaining the same complexity level as before, while seven tasks exhibited overall poor quality, characterized by excessive length or irrelevant text.

The tool underwent optimization after a workshop with educators from various institutions (e.g., schools, universities). The current version includes an English version, the complete CS curriculum of Baden-Württemberg (from grade 7 until 12, A-level), as well as an option to revise the task. The tool can be accessed online[1].

4 DISCUSSION AND FUTURE WORK

The potential of using *ChatGPT*, such as decreasing teaching workload, has been identified in a *SWOT* analysis [10]. In this context, a tool employing didactical parameters was developed to differentiate tasks using *ChatGPT-4* in an ongoing project. Feedback from the initial workshop with educators from diverse institutions proved valuable in enhancing the tool. The findings align with other research on the usage of *ChatGPT* in K-12 educational settings [4], indicating that while many results are sensible, critical thinking is required as some results are factually incorrect or not suitable for the target group (even when selecting the grade and complexity level). Future work entails further refining the tool based on feedback from various stakeholders in the CSE community to enhance its sensibility and practical relevance. Another avenue could involve reorienting the tool's objective to provide the same task but differentiated through tips and hints for students at different competence levels. Alternatively, training a large language model solely with the definitions of didactical parameters and educational materials of certain quality could potentially reduce the generation of tasks with low quality. One feedback from the workshop suggested that the tool performs better for generating exam questions with a fixed set of criteria than tasks intended for classroom teaching. Exploring this direction further could be worthwhile.

[1] https://aufgabendifferenzierungbeta.azurewebsites.net/computer-science

REFERENCES

[1] Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition.* Addison Wesley Longman, Inc.

[2] Tobias Bahr. 2023. Developing a Computer Science Content Knowledge Test for 10th Grade Students. *The 18th WiPSCE Conference on Primary and Secondary Computing Education Research (WiPSCE '23)* (2023), 2 pages. https://doi.org/10.1145/3605468.3609778

[3] Tobias Bahr, Mario Manzocco, Dennis Schuster, and Albrecht Wacker. 2024. Textgenerierende KI zur Erleichterung der Arbeit von Lehrpersonen? Ein Vorschlag am Beispiel der Aufgabendifferenzierung mithilfe von ChatGPT. *Ludwigsburger Beiträge Zur Medienpädagogik* 24 (2024).

[4] Grant Cooper. 2023. Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *Journal of Science Education and Technology* 32, 3 (2023), 444–452. https://doi.org/10.1007/s10956-023-10039-y

[5] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[6] Alexander Jordan, Nathalie Ross, Stefan Krauss, Jürgen Baumert, Werner Blum, Michael Neubrand, Katrin Löwen, Martin Brunner, and Mareike Kunter. 2006. *Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt.* Max-Planck-Institut für Bildungsforschung.

[7] Alexander Kauertz, Hans Ernst Fischer, Jürgen Mayer, Elke Sumfleth, and Maik Walpuski. 2010. Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe. *ZfDN* 16 (2010), 135–153.

[8] Uwe Maier, Marc Kleinknecht, Kerstin Metz, and Thorsten Bohl. 2010. Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerbildung* 28, 1 (2010), 84–96. https://doi.org/10.25656/01:13734

[9] Ministerium für Kultus, Jugend und Sport Baden Württemberg. 2018. Bildungsplan zum Profilfach Informatik, Mathematik, Physik (IMP). (2018). http://bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GYM_IMP.pdf [Accessed on: Sep. 9, 2022].

[10] Omid Noroozi Mohammadreza Farrokhnia, Seyyed Kazem Banihashem and Arjen Wals. 2023. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International* 0, 0 (2023), 1–15. https://doi.org/10.1080/14703297.2023.2195846

[11] Saman Rizvi, Jane Waite, and Sue Sentance. 2023. Artificial Intelligence teaching and learning in K-12 from 2019 to 2022: A systematic literature review. *Computers and Education: Artificial Intelligence* 4 (2023), 100145. https://doi.org/10.1016/j.caeai.2023.100145

[12] Ralf Romeike. 2006. Creative students: what can we learn from them for teaching computer science?. In *Proceedings of the 6th Baltic Sea Conference on Computing Education Research: Koli Calling 2006* (Uppsala, Sweden) *(Baltic Sea '06)*. Association for Computing Machinery, New York, NY, USA, 149–150. https://doi.org/10.1145/1315803.1315835

[13] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1* (Lugano and Virtual Event, Switzerland) *(ICER '22)*. Association for Computing Machinery, New York, NY, USA, 27–43. https://doi.org/10.1145/3501385.3543957

[14] Kirsten Schlüter and Torsten Brinda. 2008. From exercise characteristics to competence dimensions exemplified by theoretical computer science in secondary education. In *Proceedings of the Joint Open and Working IFIP Conference on ICT and Learning for the Net Generation (LYICT 2008), Kuala Lumpur, Malaysia), URL: http://cs. anu. edu. au/iojs/index. php/ifip/article/viewFile/1012/16.*

[15] Ethel Tshukudu, Jane Waite, Saman Rizvi, and Sue Sentance. 2022. Teachers' Motivations to Learn about ML and AI. In *Proceedings of the 27th ACM Conference on Innovation and Technology in Computer Science Education Vol. 2* (Dublin, Ireland) *(ITiCSE '22)*. Association for Computing Machinery, New York, NY, USA, 609. https://doi.org/10.1145/3502717.3532148

[16] ZSL. 2018. Fachportal Informatik, Mathematik, Physik (IMP). (2018). https://lehrerfortbildung-bw.de/u_matnatech/imp/gym/bp2016/ [Accessed on: Sep. 9, 2022].